Research Report
ETS RR-14-32

Examining the *WorkFORCE*™ Assessment for Job Fit and Core Capabilities of the *FACETS*™ Engine

Bobby Naemi

Jacob Seybert

Steven Robbins

Patrick Kyllonen

October 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Examining the *WorkFORCE*™ Assessment for Job Fit and Core Capabilities of the *FACETS*™ Engine

Bobby Naemi, Jacob Seybert, Steven Robbins, and Patrick Kyllonen
Educational Testing Service, Princeton, New Jersey

October 2014

Action Editor: James Carlson

Reviewers: David Klieger and Richard Tannenbaum

Abstract

This report introduces the *WorkFORCE*™ Assessment for Job Fit, a personality assessment utilizing the *FACETS*™ core capability, which is based on innovations in forced-choice assessment and computer adaptive testing. The instrument is derived from the five-factor model (FFM) of personality and encompasses a broad spectrum of personality assessment. This document provides an overview of the assessment, beginning with detailing evidence-based practices for personality measurement and modeling its relationship to workplace outcomes. We address the validity and fairness of this assessment, the creation of composite scores, and the generalizability of the assessment across languages and job types. We conclude with recommendations on the use of this capability for workforce applications and guidelines for future research.

Key words: five-factor model of personality, workplace outcomes, personality assessment

Interest continues to grow in the use of personality and other noncognitive factors and their relation to educational and workplace outcomes. Research and meta-analytic findings suggest that personality constructs in particular are important predictors of educational outcomes (Porchea, Allen, Robbins, & Phelps, 2010; Richardson, Abraham, & Bond, 2012; Robbins, Allen, Casillas, Peterson, & Le, 2006), task performance in the workplace (Barrick & Mount, 1991; Campbell, 1990; Campbell & Knapp, 2001), citizenship behaviors at work (Borman & Motowidlo, 1997), and turnover (Boudreau, Boswell, Judge, & Bretz, 2001). One strong motivator of this interest is the potential for personality variables to reduce the adverse impact typically observed through the use of cognitive ability measures (Sackett, Schmitt, Ellingson, & Kabin, 2001).

The *WorkFORCE*™ Assessment for Job Fit was developed to provide a high-quality measure of personality for use in both developmental and high-stakes assessment in educational and workplace settings. Administered using the *FACETS*™ engine developed by Educational Testing Service (ETS), this assessment measures 13 personality facets using an innovative forced-choice response format designed to reduce response biases and provide normative trait scores. By following a framework based on contemporary views of personality theory and making use of modern scoring techniques, the WorkFORCE Assessment for Job Fit provides a basis for scoring individuals across a range of organizational, educational, and developmental contexts.

An Overview of Personality Measurement

Although a range of theoretical frameworks for describing personality traits has been proposed, over the last 30 years the five-factor model (FFM) of personality (Goldberg, 1990) has emerged as the predominant framework from which to approach personality measurement. The classification of the various potential personality traits into the five broad factors of conscientiousness, agreeableness, extraversion, emotional stability, and openness to experience has been an important advance for synthesizing research findings across a range of measures (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Despite the dominance of these factors for describing personality and theory building, narrow traits that can be categorized within the FFM framework have been found to provide high levels of predictive validity (Ashton, 1998; Paunonen, 1998) and a more detailed set of scores from which to derive composites related to specific and narrow outcomes. Researchers have also argued that narrow lower order personality facets can provide better prediction than higher order factors (see review in Oswald & Hough, 2011). A number of efforts have been made to identify a comprehensive framework of narrow personality traits, including that for the NEO PI-R (Costa, McCrae, & Dye, 1991) and the International Personality Item Pool (IPIP; Goldberg et al., 2006). However, a recent series of analyses summarized by Drasgow et al. (2012) provided a comprehensive taxonomy of 21 lower order personality facets that can be placed under the hierarchical structure of the FFM. These facets have both a strong theoretical foundation and empirical support of their importance in predicting workplace outcomes (see the Psychometric Background of the WorkFORCE Assessment for Job Fit and the Validity Evidence Related to WorkFORCE Assessment for Job Fit Scores sections of this paper).

Despite the research evidence supporting both the FFM and narrow personality facets in predicting academic and workplace success, use of personality assessment in high-stakes contexts has been largely limited due to concerns regarding response distortions that may reduce the usefulness of test scores (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; McGrath, Mitchell, Kim, & Hough, 2010). Of particular concern is the tendency for individuals, rather than responding to personality statements honestly, to identify desirable trait descriptors and purposefully endorse those at higher rate than accurate. This issue of what has been termed *faking* can reduce the validity of observed test scores and thus limit their use. Although, Ones and Viswesvaran (1998) and Hogan, Barrett, and Hogan (2007) have argued that faking is not an issue generally for predictive validity, others (e.g., Mueller-Hanson, Heggestad, & Thornton, 2003) have provided evidence that, in terms of predictive validity, problematic differences still exist for those who have an incentive to fake. In situations where selection ratios are small, in particular, Mueller-Hanson et al. (2003) found that participants with an incentive to fake were more likely to be selected despite having lower mean performance task scores than a control group of participants.

A number of approaches to deal with faking have been explored, including correcting for faking (Ellingson, Sackett, & Hough, 1999), warnings against faking (McFarland, 2003), and obtaining observer ratings (e.g., *ETS® Personal Potential Index* [*ETS® PPI*]; ETS, 2009), alongside more novel idiosyncratic approaches (see Ziegler, MacCann, & Roberts, 2011, for a review). Evidence of the efficacy of these approaches is limited or, in the case of the ETS PPI, requires obtaining data from another source. As an alternative to these strategies, research has focused on attempting to develop tests that are resistant to faking through varied methods such as empirically keyed items, conditional reasoning tests, and rapid responding measures (see Ziegler et al., 2011, for further review). One avenue of particular interest to combat faking, however, is the use of forced-choice item formats, a method with a lengthy history in psychology (e.g., Ghiselli, 1954; Gordon, 1951). Forced-choice methods for this purpose tend to require respondents to select between two equally desirable statements representing different personality facets. For example, a respondent may be presented with the following pair of statements; the first represents a high level of agreeableness, and the second, a high level of conscientiousness:

____ I get along well with others.

____ I always arrive to meetings on time.

For each pairwise preference item, the respondent is instructed to select the statement that is most like the respondent. Although forced-choice methods have typically produced ipsative or quasi-ipsative scores (Hicks, 1970), research has suggested that the approach provides test scores that may be valid predictors of performance, particularly those that are quasi-ipsative (see Salgado & Táuriz, 2012, for a complete review of this evidence). Despite these promising findings, concerns about ipsativity still have limited the application in high-stakes settings. To address this concern, recent research has made advances in the psychometric modeling of these measures and has resulted in the obtaining of test scores that provide normative, rather than ipsative, trait scores.

The past decade has seen the introduction of a number of factor analytic and item response theory (IRT)–based strategies for obtaining normative trait information from forced-choice measures (Maydeu-Olivares & Brown, 2010; Stark, Chernyshenko, & Drasgow, 2005; Stark & Drasgow, 2002). Of particular relevance for the WorkFORCE Assessment for Job Fit, Stark and colleagues developed a strategy for constructing and scoring pairwise preference measures and developed computerized adaptive testing algorithms for the administration of such an approach (Stark, Chernyshenko, Drasgow, & White, 2012). This approach was adopted by the U. S. Army for use in selecting and classifying recruits into job categories (Drasgow et al., 2012; Nye, Drasgow, et al., 2012). The WorkFORCE Assessment for Job Fit is based on the same model and strategies used for the U.S. Army, providing a strong empirical foundation from which to build a comprehensive personality measure for the civilian world.

Purpose of the Current Report

This report is designed to provide an overview of the WorkFORCE Assessment for Job Fit. Specifically, this report first provides an overview of the personality trait taxonomy available for use with the assessment and the psychometric models underlying the forced-choice strategy. Following this, the validity evidence of the assessment across applied and research samples is detailed. Next, an overview of how the individual personality facet scores can be used to form composite scores for predicting educational and organizational outcomes is presented. Finally, evidence supporting the fairness of the WorkFORCE Assessment for Job Fit is described along with the future directions related to its implementation and use.

Psychometric Background of the WorkFORCE Assessment for Job Fit

The WorkFORCE Assessment for Job Fit is rooted in modern personality and psychometric theory. The personality facets measured using the assessment were selected based on both empirical and theoretical arguments that were used to describe a lower order trait taxonomy for the FFM framework. These facets are measured using a 120-item pairwise preference test, administered using the FACETS engine, a computerized adaptive testing environment. This section provides an overview of the development of the personality framework used by the WorkFORCE Assessment for Job Fit and the psychometric models used for the administration and scoring of the pairwise preference items.

The Narrow Trait Taxonomy of FACETS

The FFM conceptualization (Goldberg, 1990) has become the predominant approach to classifying the normal dimensions of personality. The personality factors of conscientiousness, extroversion, openness to experience, agreeableness, and emotional stability have provided a general framework for describing an individual’s behavioral and emotional characteristics. But underlying each FFM dimension are lower order facets that can be used to provide assessment feedback and also make more a precise prediction of criteria of interest. The correlations among these personality facets are relatively low and may demonstrate differential ability to predict behavioral criteria. Studies have indicated the utility of this facet approach (Mershon & Gorsuch, 1988; Oswald & Hough, 2011; Paunonen, 1998), and Drasgow et al. (2012) argued that these lower order facets, when combined, may provide greater levels of validity in predicting broad organizational outcomes as compared to one broad predictor. Across a series of studies (Drasgow et al., 2012; B. Roberts, Chernyshenko, Stark, & Goldberg, 2005; Woo et al., 2014), a lower order trait taxonomy for the FFM dimensions was derived from which the WorkFORCE Assessment for Job Fit's personality dimensions were drawn. These narrow trait domains were determined through the factor analysis of data from a sample of individuals responding to seven major personality inventories over a period of 5 years and an analysis of the lexical structure of those inventories (see Drasgow, et al., 2012, for a complete description of this process). For each higher order personality dimension, an exploratory factor analysis (EFA) was performed, and the resulting factor structure was used to identify logically similar aspects (facets) of each dimension. As a consequence of this process, a framework of 21 facets of personality was derived. The taxonomy described by Drasgow et al. (2012), including definitions and example characteristics for each trait, is detailed below. Table 1 provides a summary of this trait taxonomy available for FACETS.

Table 1

Personality Trait Taxonomy for Existing FACETS Framework

Big Five dimension	Lower order facet name	Brief description
Conscientiousness	Diligence	Feelings and behaviors associated with working toward goals and other positive outcomes
	Rule following	Personal values emphasizing tradition, duty, and traditional moral standards
	Organization	Behaviors and intentions related to the ability to plan and organize tasks and activities
	Dependability	Feelings and actions related to a sense of duty or being answerable for one's behavior
	Self-discipline	Thoughts and behaviors centered around impulsiveness, the ability to focus on tasks without distraction, and the consideration of consequences before taking action
	Character	Beliefs and behaviors associated with adherence to standards of honesty, morality, and good Samaritan behavior
Extroversion	Assertiveness	Behaviors associated with being direct and decisive
	Experience seeking	Intentions and behaviors associated with the exciting components of social interactions and receiving attention
	Friendliness	Interest in engaging in friendly social interactions
Openness to experience	Aesthetic taste	Behaviors and feelings related to emotions, the senses, and art
	Inquisitiveness	Interest and behaviors directed toward understanding how the world around us works
	Introspectiveness	Behaviors and intentions aimed at understanding one's self and/or facilitating self-improvement and self-actualization

Big Five dimension	Lower order facet name	Brief description
	Creativity	Thoughts and behaviors associated with imagination and original thinking
	Intellectual orientation	Interest in and comfort with intellectual and conceptual matters
	Open-mindedness	Thoughts and behaviors related to comfort and interest toward strangers and new stimuli
Agreeableness	Thoughtfulness/ compassion	Feelings and behaviors associated with compassion and sensitivity toward others
	Collaboration	Behaviors and intentions centered on a desire to work or act with others for a common benefit
	Generosity	Behaviors associated with activities such as helping and doing things for others, giving to charity, and volunteering for community improvement
Emotional stability	Stability	Feelings and behaviors associated with various degrees of insecurity and anxiety
	Calmness	Behaviors and feelings associated with mood and strong emotions
	Optimism	Thoughts and behaviors associated with an individual's general emotional tone and world outlook

Conscientiousness. The higher order personality dimension of conscientiousness was found to be best described by six narrow facets: diligence, organization, self-discipline, dependability, rule following, and character (Drasgow et al., 2012).

Diligence. The first facet, diligence, describes behaviors associated with working toward goals and other positive outcomes. Individuals who are high in diligence tend to be described as hard working, ambitious, confident, and resourceful. Individuals who are low in diligence tend to be content with getting by with a minimal amount of work and could be perceived as unmotivated.

Organization. The second facet, organization, describes behaviors related to the ability to plan and organize tasks and activities. Individuals who are high in organization tend to have a strong ability to organize tasks and activities and the desire to maintain neat and clean surroundings. Individuals who are low in organization tend to be more disorganized, have cluttered living space, and do not keep detailed schedules or plans.

Self-discipline. The third facet of conscientiousness, self-discipline, describes behaviors centered on impulsiveness, the ability to focus on tasks without distraction, and the consideration of consequences before taking action. Individuals who are high in self-discipline tend to be cautious, levelheaded, able to delay gratification, and patient. Individuals who are low in self-discipline tend to be impulsive, spontaneous, easily distracted, and careless.

Dependability. The fourth facet, dependability, describes behaviors related to a sense of duty or being answerable for one's behavior. Individuals who are high in dependability like to be of service to others, frequently contribute their time and money to community projects, and tend to be cooperative and dependable. Individuals who are low in dependability tend to be less concerned with the community and not someone considered by friends to be reliable.

Rule following. The fifth facet, rule following, describes personal values emphasizing tradition, duty, and conventional moral standards. Individuals high in rule following tend to comply with current rules, customs, norms, and expectations. They dislike changes and do not challenge authority. Individuals low in rule following tend to be seen as more rebellious and willing to challenge local norms and customs.

Character. Finally, the sixth facet, character, represents a constellation of beliefs and behaviors associated with adherence to standards of honesty, morality, and good Samaritan behavior. Individuals who are high in character have a tendency to act in accordance with accepted standards of good, or moral, behavior and strive to be a moral exemplar. Individuals who are low in character have a tendency to not act within these accepted standards.

Extroversion. The higher order personality dimension of extroversion was found to be best described by three narrow facets: assertiveness, experience seeking, and friendliness (Drasgow et al., 2012).

Assertiveness. The first facet, assertiveness, describes behaviors associated with being direct and decisive. Individuals who are high in assertiveness tend to be domineering, take charge, and are often called natural leaders by their peers. Individuals who are low in assertiveness tend to be more passive and do not have a strong need to exert their own personal influence on other people and events.

Experience seeking. The second facet, experience seeking, describes behaviors associated with the exciting components of social interactions and attention. Individuals who are high in experience seeking tend to engage in behaviors that attract a lot of social attention: They are loud, entertaining, and even boastful. Individuals who are low in experience seeking are more withdrawn in social situations and do not feel the need to be in the spotlight.

Friendliness. The third and final facet, friendliness, describes an individual's level of interest in friendly social interactions. Individuals who are high in friendliness tend to be very interested in socializing with other people. Individuals who are low in friendliness tend to be uninterested in having frequent social interaction.

Openness to experience. The higher order personality dimension of openness to experience was found to be best described by six narrow facets: aesthetic taste, inquisitiveness, introspectiveness, creativity, intellectual orientation, and open-mindedness (Drasgow et al., 2012).

Aesthetic taste. The first facet, aesthetic taste, describes behaviors and feelings related to emotions, the senses, and art. Individuals who are high in aesthetic taste enjoy acquiring, participating, or creating various forms of artistic, musical, or architectural outputs. These high aesthetic taste individuals are not necessarily interested in understanding how or why things they enjoy were created; instead, they are more interested in the experience, while individuals who are low in aesthetic taste have no such interest.

Inquisitiveness. The second facet, inquisitiveness, describes behaviors directed toward understanding how the world works. Individuals who are high in inquisitiveness would be characterized as inquisitive and perceptive; they read popular science/mechanics magazines and are interested in experimenting with objects and substances. Individuals who are low in inquisitiveness tend to be less interested in understanding how things work and do not spend a lot of time learning new things.

Introspectiveness. The third facet of openness to experience, introspectiveness, describes behaviors aimed at understanding one’s self and/or facilitating self-improvement and self-actualization. Individuals who are high in introspectiveness tend to engage in behaviors such as reflection, meditation, introspection, personal growth, and spiritual enlightenment. Individuals who are low in introspectiveness tend to not consider the meaning behind events and feelings and focus on practical matters in life.

Creativity. The fourth facet, creativity, describes behaviors associated with imagination and original thinking. Individuals who are high in creativity tend to be inventive and enjoy making improvements to things. Individuals who are low in creativity are not very creative and tend to rely on other people’s ideas.

Intellectual orientation. The fifth facet, intellectual orientation, involves interest and comfort with intellectual and conceptual matters. Individuals who are high in intellectual orientation are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual. Individuals who are low in intellectual orientation tend to have difficulty understanding new things.

Open-mindedness. The sixth and final facet, open-mindedness, involves behaviors related to comfort and interest toward strangers and new stimuli. Individuals who are high in open-mindedness like to attend cultural events or meet and befriend people with different views. They also tend to better adapt to new situations, unlike people who are low in open-mindedness.

Agreeableness. The higher order personality dimension of agreeableness was found to be best described by three narrow facets: thoughtfulness, collaboration, and generosity (Drasgow et al., 2012).

Thoughtfulness. The first facet, thoughtfulness, describes behaviors and feelings associated with compassion and sensitivity toward others. Individuals scoring high in thoughtfulness are considerate, affectionate, and positive toward others. These high thoughtfulness individuals may be social or nonsocial. However, they tend to be readily able to see to others' emotional needs. Individuals who are low in thoughtfulness tend to not be interested in others' problems and do not connect easily with people.

Collaboration. The second facet, collaboration, describes behaviors centered on a desire to work or act with others for a common benefit. Individuals who are high in collaboration tend to be trusting, cordial, noncritical, and easy to live with. Individuals who are low in collaboration tend to be skeptical, suspicious, and even confrontational.

Generosity. The third and final facet, generosity, describes behaviors associated with activities such as helping and doing things for others, giving to charity, and volunteering for community improvement. Individuals who are high in generosity tend to be willing to share their time and resources. Individuals who are low in generosity are egoistical, greedy, and tend look down on people.

Emotional stability. The higher order personality dimension of emotional stability was found to be best described by three narrow facets: stability, calmness, and optimism (Drasgow et al., 2012).

Stability. The first facet, stability, describes behaviors associated with various degrees of insecurity and anxiety. Individuals who are high in stability tend to be self-assured, relaxed, and confident. Individuals who are low in stability are high-strung, self-conscious, and apprehensive in most contexts.

Calmness. The second facet, calmness, describes behaviors associated with mood and strong emotions. Individuals who are high in calmness tend to be calm and stable, even when threatened. Individuals who are low in calmness tend to experience a range of emotions including irritability, anger, hostility, or even aggression.

Optimism. The third and final facet, optimism, describes behaviors associated with an individual's general emotional tone and world outlook. Individuals who are high in optimism tend to be happy and joyful and have a positive outlook. Individuals who are low in optimism are depressed and hopeless and have a more negative outlook at life.

Statement Pool Development

Given the developed trait taxonomy and corresponding facet level definitions, statements used for WorkFORCE Assessment for Job Fit were developed first by examining available statements and writing new statements intended to assess behaviors, cognition, and affect for each of the 13 lower order facets. Then subject matter experts provided initial judgments of the level of the underlying trait at which an individual is expected to agree with each statement. These judgments were then used to identify gaps in the breath of the pool. Following editing and revision of statements to ensure clarity and quality, data were gathered in a pretesting phase across 12 studies that were conducted with samples of 270 to 588 individuals who were recent recruits into the U.S. Army. For each sample, respondents completed questionnaires presenting questions related to six to 10 facets, each represented by 15 to 30 statements. These respondents were instructed to respond honestly and indicate their level of agreement with each statement on a 1 (*strongly disagree*) to 4 (*strongly agree*) Likert-type scale.

Following collection of these pretesting data, responses to the statement were used to examine the unidimensionality of the scales and statements that did not seem to measure their intended facet were removed from the analyses. Then responses to the remaining statements were dichotomized and parameter estimates were obtained using standard IRT procedures (Koenig & Roberts, 2007; J. S. Roberts, Donoghue, & Laughlin, 2000). Statements were then screened for quality, and those indicating poor discrimination were eliminated from the pool. Finally, social desirability estimates for each statement were obtained using a *fake good* study, where 30 to 40 respondents were instructed to respond to each statement in a way that presented them in the best possible light.

Drasgow et al. (2012) examined the alternative approach of using individuals to directly provide ratings of social desirability. The two sets of desirability ratings correlated .87, indicating that in practice either procedure may be used to obtain social desirability estimates. This effort produced a large statement pool from which FACETS can draw. Psychometric models underpinning statement parameters, pairwise-preference test administration, and scoring are detailed in the next section.

Modeling Statements With the Generalized Graded Unfolding Model (GGUM)

Endorsement probabilities for each statement comprising pairwise preference WorkFORCE Assessment of Job Fit items are calculated with the generalized graded unfolding model (GGUM; J. S. Roberts et al., 2000). The GGUM is an ideal point model, meaning that the probability of endorsing a statement increases as the distance between the person and statement locations on the trait continuum decreases. When applied in the FACETS engine used by the WorkFORCE Assessment for Job Fit, the GGUM is used to compute statement agreement probabilities that underlie *most like* selections and dichotomous parameter estimates. Letting $P(0)$ and $P(1)$ represent the respective probabilities of disagreeing (Z = 0) and agreeing (Z = 1) with a particular statement, given a respondent's latent trait score (θ) on the dimension that statement represents, and three statement parameters (α, δ, τ) reflecting discrimination, statement location (extremity), and threshold, respectively, the GGUM probability equations are given as

$$P(0) = P(Z = 0|\theta) = \frac{1+\exp(\alpha[3(\theta-\delta)])}{\gamma}, \text{ and} \qquad (1)$$

$$P(1) = (Z = 1|\theta) = \frac{\exp(\alpha[(\theta-\delta)-\tau])+\exp(\alpha[2(\theta-\delta)-\tau])}{\gamma}, \qquad (2)$$

where

$$\gamma = 1 + \exp(\alpha[3(\theta-\delta)]) + \exp(\alpha[(\theta-\delta)-\tau]) + \exp(\alpha[2(\theta-\delta)-\tau])$$

is a normalizing factor equal to the sum of the numerators of Equations 1 and 2.

Ideal point models, such as the GGUM, assume that a comparison process governs the decision to agree or disagree with a statement. Specifically, they assume a respondent estimates the distance between his or her location and the location of the statement on the underlying trait continuum. If the distance is small, the respondent agrees with the statement. If the distance is large, the respondent disagrees. Thus, as the perceived distance between the person and the statement increases, the probability of agreeing with the statement decreases. Ideal point models can therefore have *item response functions* (IRFs), which portray the relationship between trait scores and agreement probabilities and are nonmonotonic and possibly bell shaped.

Figure 1 presents an example IRF for the dichotomous GGUM for a statement having discrimination, location, and threshold parameters, $\alpha = 1.75$, $\delta = 0.00$, and $\tau = -1.50$, respectively. The horizontal axis in Figure 1 represents the level of the underlying latent trait, and the vertical axis shows the probability of agreeing with the statement. It can be seen that the probability of agreement is highest when $(\theta - \delta) = 0$, and it decreases in both directions, resulting in a symmetric, unimodal form. The rate of decrease in the probability of agreement depends on the joint relationship between the item discrimination and item threshold parameters, while the location parameter determines where the peak of the IRF occurs. (More details concerning GGUM IRFs can be found in J. S. Roberts et al., 2000; J. S. Roberts & Thompson, 2011; Seybert, Stark, & Chernyshenko, 2014; Stark et al., 2005).

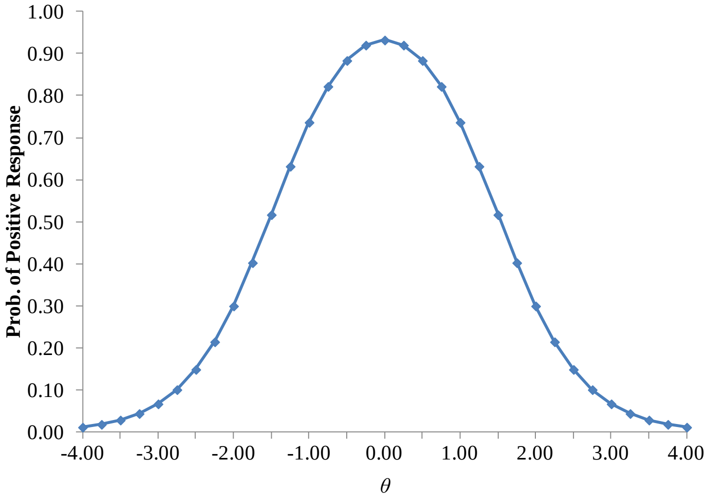

Figure 1**. An item response function (IRF) for a dichotomous generalized graded unfolding model (GGUM) item.**

The Multiunidimensional Pairwise Preference (MUPP) Model and Its Use With WorkFORCE Assessment for Job Fit

Stark et al. (2005) developed the multiunidimensional pairwise preference (MUPP) IRT model for scoring measures such as the WorkFORCE Assessment for Job Fit. This model assumes that when presented with a pair of statements representing the same or different constructs a respondent evaluates each statement and judges whether each will be agreed with or disagreed with. When making the preference choice, those judgments are used to determine the level of preference for selecting one statement while disagreeing with the other. Using this model, each individual statement is described by a unidimensional model, but the test as a whole is composed of multiple dimensions (thus the multiunidimensional description). Formally, the probability of preferring statement *s* to statement *t* in a pairwise preference item is given by

$$P(s > t)_i(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\}+P_{st}\{0,1\}} = \frac{P_s(1)P_t(0)}{P_s(1)P_t(0)+P_s(0)P_t(1)} \ , \qquad (3)$$

where

i = the index for each pairwise preference item, where $i = 1$ to the total number of items (I);

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

d = the dimension associated with a given statement, where $d = 1, \dots, D$;

$\theta_{d_s}, \theta_{d_t}$ = the latent trait values for a respondent on dimensions d_s and d_t, respectively;

$P_{st}\{1,0\}$ = the joint probability of selecting statement *s*, and not selecting statement *t*;

$P_{st}\{0,1\}$ = the joint probability of selecting statement *t*, and not selecting statement *s*;

$P_s(1),\ P_t(1)$ = the probabilities of endorsing statements *s* and *t*, respectively;

$P_s(0),\ P_t(0)$ = the probabilities of not endorsing statements *s* and *t*, respectively; and

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

Stark et al. (2005) described and evaluated a two-stage approach to multidimensional forced-choice (MFC) test construction and scoring where the first step involves obtaining statements and parameter estimates and the second step consists of administering and scoring the forced-choice responses. At the first step, noncognitive statements are written that range in extremity from low to medium to high in the each of the dimensions to be assessed. These statements are then administered to large samples of respondents with instructions to honestly indicate their levels of agreement using an ordered polytomous response format. Statement parameter estimates using the dichotomized polytomous data are obtained using an IRT model for single-statement responses that provides adequate model-data fit. Stark chose the GGUM as the basis for test construction and scoring. After estimating statement parameters using the GGUM, social desirability estimates are obtained for use in MFC item creation by readministering the statements in the context of a *fake good* study (White & Young, 1998) or by collecting ratings from individuals familiar with the constructs being assessed and the population to which they are to be administered.

Next, the multidimensional pairwise preference (MDPP) items are formed by pairing statements similar in social desirability from different dimensions, and the MFC test forms are assembled by combining multidimensional pairs with a small percentage of similarly matched unidimensional pairs to identify the metric of trait scores. In the second step, scoring the MDPP tests is accomplished by multidimensional Bayes modal estimation, via the substitution of observed responses and GGUM statement estimated parameters into Equation 3 for pairwise preference response probabilities. The procedures developed by Stark et al. (2005) for the scoring of MUPP forced-choice response patterns are detailed in Appendix A. The resulting trait score estimates are traditionally reported on a standard normal distribution scale and typically range from -3.00 to 3.00.

Figure 2 presents an IRF for a pair of statements, *s* and *t*, representing different dimensions. Statement *s* has parameters α_s = 1.20, δ_s = 3.00, τ_s = -1.50 and statement *t* has parameters α_s = 1.20, δ_s = 2.50, τ_s = -2.00. The vertical axis represents the probability of preferring statement *s* to *t*, given the latent trait values on the horizontal axes corresponding to the dimensions measured by the statements. The response function for this is a three-dimensional response surface, which is essentially monotonically increasing, with the probability of selecting statement *s* being highest along $\theta_s = \delta_s$ and lowest along $\theta_t = \delta_t$. It should be noted that because the model assumes an ideal-point response process, the shape of the response surface may vary widely, including in a nonmonotonic increasing or decreasing direction, depending on the constituent statement parameters.

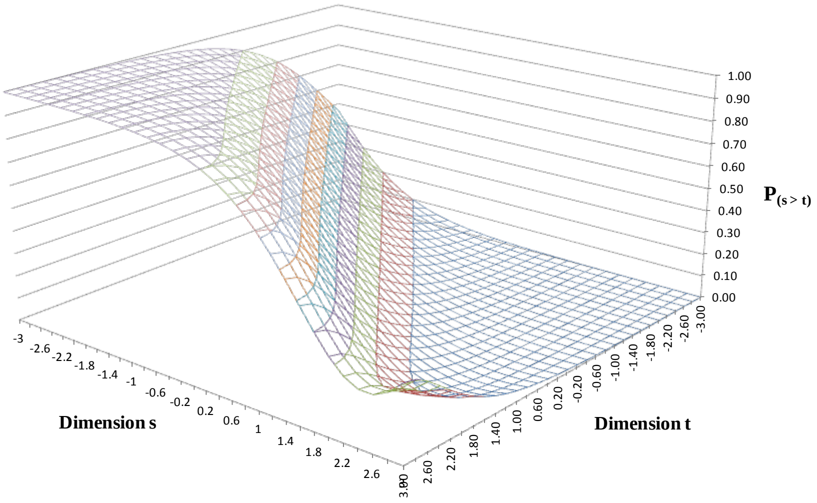

Figure 2. **Example item response surface plot for a multi-unidimensional pairwise preference (MUPP) item for selecting statement *s* (α = 1.20, δ = 3.00, τ = -1.50) over statement *t* (α = 1.20, δ = 2.50, τ = -2.00).**

Implementation of Administration and Scoring for WorkFORCE Assessment for Job Fit

The pairwise preference items used by the WorkFORCE Assessment for Job Fit can be administered either through the use of a fixed form, where all pairs within a test form are constructed prior to administration of the test, or by using a computerized adaptive test (CAT) format, where items are tailored to the respondent’s trait level as the test progresses. One benefit of the use of a CAT-based test is that typically half as many items are required, when compared to fixed-form tests, to obtain the same level of measurement precision (Weiss & Kingsbury, 1984). This reduced number of items also has the benefit of a decreased level of item exposure, which could lead to reduced issues with test security and related problems that occur in high-stakes settings. Consequently, the implementation of WorkFORCE Assessment for Job Fit in a CAT form provides many benefits when measuring a large number of personality dimensions in a short period of time. Below, the CAT algorithm used for administration and scoring of WorkFORCE Assessment for Job Fit is described, followed by a summary of research providing evidence of this approach in accurately recovering normative trait scores. These administration and scoring approaches rely on the FACETS engine for implementation.

Overview of the FACETS computerized adaptive test (CAT) algorithm used by WorkFORCE Assessment for Job Fit. Adaptive testing using an MDPP format must take into consideration the individual statements comprising each constructed item. In particular, resistance to response distortion should be maintained through pairing statements having similar levels of social desirability. Equally important, the dimensional representation and location of the statements must vary in order to measure sufficiently the traits of interest. Drasgow et al. (2012) and Stark et al. (2012) described a *fixed length* adaptive algorithm, used by FACETS, where the total length of a test is predetermined but the individual items are selected to obtain the greatest degree of measurement precision from the test. To accomplish this, the following steps are taken to administer and score a FACETS test:

1. The number of dimensions must be specified by the test user and the total test length determined through selecting the number of items per dimension. For example, a test measuring 10 personality facets may be created so that eight items per dimension are represented, resulting in a total test length of 80 item pairs.

2. Next, the test blueprint, which specifies the dimensional pairing of each of the traits being measured, must be created. Although allowing for all possible pairings of dimensions may be practical for tests when relatively few traits are being measured, this becomes increasing impractical for tests of high dimensionality. Stark et al. (2012) proposed a strategy of circular linking, where dimensional pairings occur through chaining across dimensions. For example, a test measuring five dimensions could link across those dimensions using pairings such as 1-2, 2-3, 3-4, 4-5, and 5-1, so that at least two statements for each dimension are administered and trait scores can be updated as early as possible. The remaining items on the test blueprint are designated as the main subtest, where dimensions are paired with other dimensions based on the specifications of the test user. A small proportion of unidimensional pairs (e.g., 1-1, 2-2) are also specified within the main subtest to facilitate trait score estimation accuracy (see Stark et al., 2012, for a more detailed discussion of this

approach). Tables 2 and 3 present an example of the kind of information used to determine trait recovery accuracy.

3. At the start of the test, the respondent is assumed to have an initial trait score of zero (on a *z*-score scale) on each dimension. Using the test blueprint, pairs of statements are selected so that the items provide high levels of information given the current trait score estimates. These pairs are subject to location and desirability constraints, so that statements should be of similar extremity and social desirability, reducing the likelihood of one statement being viewed as more right than the other.
4. After each item response is obtained, the respondent's trait scores are estimated and the next item pair is selected. This process continues until the total number of test items has been completed. Then final trait score estimates are obtained.
5. Standard error estimates of respondent trait scores are calculated at the end of the test using a process described by Stark et al. (2012). To obtain standard error estimates, the final trait scores are used to generate 30 response patterns and those data are then scored. The standard deviation of each of the respective traits indicates the variation in scores due to error and is used as standard error estimates. (The interested reader is directed to Stark et al., 2012, for more details on this process.)

Use of Normative Information With FACETS Trait Estimates

The IRT-based trait score estimates produced by FACETS are placed on a metric relative to the statement parameter estimates, which are difficult to interpret. Further, because the mean and variance of the traits may vary across each dimension, the combination of the scores through the use of a weighting scheme (see the section titled Forming Composites and Building Overall Selection Indices in this paper) may result in unequal influence of a particular trait because each differs in scale. To address these concerns, the use of normative trait information given an appropriate norm group is recommended. For each IRT-based trait score resulting from FACETS, the relative position of the raw estimate is communicated as a percentile rank or *z*-score (mean 0, variance 1).

To obtain normative scores, trait scores for the normative group are retrieved and a table is created that communicates the percentile rank of each estimate. When trait scores are obtained during subsequent administrations, each estimate is compared to the score table from the norm group and assigned a percentile rank. This percentile rank is then converted into a *z*-score, based on the corresponding cumulative distribution value. For example, a trait estimate of −1.87 may reflect a .01 percentile score in a normative group. Later, a new test taker obtains a score of −1.87 on that trait, which corresponds to a .01 percentile, which in turn is equivalent to a *z*-score of −2.33. Through converting raw IRT-based trait score estimates to reflect normative group scores, interpretation and application of FACETS scores are improved.

Trait Estimation Accuracy Using FACETS

An initial examination of the accuracy of trait recovery using the MUPP model was conducted by Stark et al. (2005). A more detailed study was performed and described by Drasgow et al. (2012) and Stark et al. (2012), comparing the performance of both fixed-form and CAT-based MUPP tests. In these studies, trait recovery accuracy was examined across differing numbers of dimensions, numbers of items per dimension, proportions of unidimensional pairs included in the test blueprint, and the correlations between each of the dimensions. The results of these studies are presented in Tables 2 and 3. Table 2 provides the correlation between the known trait scores used to simulate the response data (i.e., generating trait scores) and estimated trait scores obtained from the simulated response data, averaged across dimensions for each condition. Table 3 provides the average absolute bias for these same conditions. Examining these tables (and as reported by Drasgow et al., 2012, and Stark et al., 2012), it can be seen that the average correlation between generating and estimated trait scores increased as test length increased, along with a corresponding decrease in average absolute bias. In contrast, the proportion of unidimensional pairings had little effect on trait recovery, suggesting that as little as 5% of test items need to be unidimensional for accurate trait recovery. Similarly, increasing the dimensionality of the test forms while keeping the number of items per dimension constant showed no differences in trait recovery. In regard to the correlation between generating trait dimensions, minor decreases in recovery were observed when the generating correlation was increased, due to the assumed independent priors associated with MUPP Bayes modal estimation (see Appendix A and Stark et al., 2005).

Table 2

Correlation Between Generating and Estimated Trait Scores for Nonadaptive and Adaptive Multidimensional Pairwise Preference (MDPP) Tests

			Average Correlation Across Dimensions									
			Nonadaptive					Adaptive				
ρ_{gen}	Percentage Unidimensional	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. From Development of the Tailored Adaptive Personality Assessment (TAPAS) to Support Army Selection and Classification Decisions, by F. Drasgow, S. Stark, O. S. Chernyshenko, C. D. Nye, C. L. Hulin, and L. A. White, 2012, Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences. ρ_{gen} = the correlation between the generating trait scores; 3-*D*, 5-*D*, … 25-*D* = the number of dimensions (*D*) examined in a given condition from 3-dimensions to 25-dimensions.

Table 3

Absolute Bias for Nonadaptive and Adaptive Multidimensional Pairwise Preferences (MDPP) Tests

			Average Absolute Bias Across Dimensions									
			Nonadaptive					Adaptive				
	Percentage	Items per										
ρ_{gen}	Unidimensional	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. From Development of the Tailored Adaptive Personality Assessment (TAPAS) to Support Army Selection and Classification Decisions, by F. Drasgow, S. Stark, O. S. Chernyshenko, C. D. Nye, C. L. Hulin, and L. A. White, 2012, Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences. ρ_{gen} = the correlation between the generating trait scores; 3-*D*, 5-D, … 25-*D* = the number of dimensions (*D*) examined in a given condition from 3-dimensions to 25-dimensions.

Importantly, the results observed in Tables 2 and 3 show a high degree of recovery when a sufficient number of items per dimension are used. With as few as five items per dimension, correlations between generating and estimated thetas were in the low to mid .70s for the fixed-form tests and the low to mid .80s for the CAT conditions. When increased to 20 items per dimension, correlations for both fixed-form and CAT conditions increased to the .90s, with CAT conditions showing slightly higher recovery than fixed-form tests. These results can be used to inform decisions regarding test length that must be made when developing a FACETS-based test. Tradeoffs between test length (and thus test-taking time) and estimation accuracy may be balanced by the test user. Overall these results show that the CAT format for FACETS can provide the highest level of recovery of trait scores and that recovery is markedly high considering the potentially large number of dimensions and relatively few items presented in a short period of time.

Validity Evidence Related to WorkFORCE Assessment for Job Fit Scores

The validity of a selection procedure, such as a personality test, can be thought of as the extent to which theory and data from validation studies support interpretations of its produced scores (Messick, 1989). The process of validation involves developing a sound validity argument: that is, the extent to which existing research supports the intended interpretation of test scores for specific uses, such as personnel selection (American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 2014). The validity argument can draw on multiple types of validity, so long as they are appropriate for a given testing need (Kane, 2006). Validity evidence is generally classified into a number of categories, including that which provides evidence of content coverage, evidence examining the response process, evidence of the internal structure of the test, and evidence relating to external variables (AERA, APA, & NCME, 2014). Some types of validity evidence are gathered for specific selection procedures, while other types of validity evidence involve use of evidence available from earlier reported research. It is widely accepted that construct validity subsumes other types of validity, in that those other types of validity are used to marshal evidence in support of a selection procedure's construct validity (AERA, APA, & NCME, 2014). To support the overall construct validity of scores obtained for the WorkFORCE Assessment for Job Fit, a detailed examination of several kinds of validity evidence follows.

Content-Related Validity Evidence of WorkFORCE Assessment for Job Fit Scores

Content validity is demonstrated by data showing that the content of a selection procedure is representative of important aspects of the domain of interest (e.g., job performance; Sireci, 1998; Society for Industrial and Organizational Psychology [SIOP], 2003). Content validity evidence can include logical or empirical (data-based) analyses of the adequacy with which the selection procedure's content represents the content domain. The content domain, in an employee selection context, is typically identified by a careful job analysis (or, alternatively, through competency modeling; see Sanchez & Levine, 2009, for a comparison of those approaches) that identifies work behaviors and employee attributes important to the job(s) for which the selection procedure is to be used. In the case of the WorkFORCE Assessment for Job Fit, a job analysis will be conducted before operational use for each new job category, using samples of incumbents and supervisors who complete the ETS Job Profiler tool, an online survey of managers based on the framework used by the Occupational Information Network (O*NET), a database developed by the U.S. Department of Labor to group job categories and behaviors/attributes. (For further details on this job analysis process and the samples, please see the forthcoming report *ETS Job Profiler for Validity Studies* [Golubovich & Chatterjee, 2014].)

Personality statements developed for use with the WorkFORCE Assessment for Job Fit were grouped into composites that were mapped onto research-based taxonomies of job performance, as detailed in Table 10. Additionally, during the initial statement development, only statements that conformed to the expected dimensionality for each of the trait dimensions were retained for inclusions in the available statement pool. Consequently, the development procedures, statement analysis, and subject matter expert review provide some indication that WorkFORCE Assessment for Job Fit statements relate to the intended personality facets.

Criterion-Related Validity Evidence of WorkFORCE Assessment for Job Fit Scores

Criterion-related validity is demonstrated by showing a statistical relationship between scores obtained from an assessment procedure and scores on a criterion measure (SIOP, 2003). A criterion is the item with which the selection procedure is expected to predict/correlate: for example, work performance/behavior that can be expected to be predicted by various personality characteristics. This statistical relationship supplements evidence of the relevance of the job performance measure to the overall job performance domain. Evidence of the relevance of the job performance measure to the overall job performance domain is usually based on job analysis (Uniform Guidelines on Employee Selection Procedures, 1978).

Drasgow et al. (2012) performed a meta-analysis of the personality dimensions upon which the trait taxonomy for WorkFORCE Assessment for Job Fit is based. Examining 43 data sources with 1,068 correlation coefficients (primarily in military and government-related jobs), Drasgow et al. showed that important work-related criteria such as turnover, adaptability, contextual performance, and task performance were associated with relatively high correlations with assorted personality facets. For example, contextual performance was best predicted by order ($r = .26$), achievement ($r = .26$), and adjustment ($r = .21$). These validity estimates provide evidence that the WorkFORCE Assessment for Job Fit dimensions may show similar ability to predict these outcome variables across a range of applied situations.

Dimensions associated with WorkFORCE Assessment for Job Fit have demonstrated convergent or criterion-related validity in several samples: a study with 102,000 U.S. military applicants (labeled Army Tech Report in the appendices and tables), a study with 293 ETS incumbents (labeled ETS SWS item tryout), a workforce sample study with 600 U.S. incumbents (labeled workforce incumbents), two workforce sample studies with U.S. and Philippines job applicants (labeled U.S. applicants & PHI applicants), a study with 610 ETS incumbents involving supervisor ratings (labeled U.S. incumbents), and an international study with samples of at least 180 working adults in seven countries (labeled PIAAC). Although some of these studies might face possible generalizability issues if attempting to directly apply results to general U.S. workforce samples (given the international or military nature of certain samples), these studies nevertheless build evidence for validation of the general FACETS capability while also providing a base for extending the use of the assessment to additional populations.

Results for each of these studies are reviewed below, with key correlations also presented in the construct map (see Table B1).

U.S. Army validity findings. The FACETS engine and statement pool are based on initial work performed with the U.S. Army’s Tailored Adaptive Personality Assessment System (TAPAS) for the assessment of personality (Drasgow et al., 2012). The U.S. Army collected criterion-related validity data from roughly 3,500 soldiers who completed a 12-facet version of the TAPAS assessment at the start of basic training. The version of TAPAS used for this data collection, known as TAPAS-95s, was a paper-and-pencil fixed form measuring 12 facets with

95 pairs of statements. Criterion measures collected both alongside and after test administration included technical training exam test scores, self-reported Army Physical Fitness Test scores, self-reported number of disciplinary incidents, self-reported ratings on the Adjustment to Army Life scale (a self-report scale of retention-related attitudes toward the U.S. Army), and attrition rates at the end of the Advanced Individual Training (AIT) or One-Station Unit Training (OSUT) training programs. Further information on these criteria can be found in the Expanded Enlistment Eligibility Metrics (EEEM) research report (Knapp & Heffner, 2010).

Table 4 demonstrates the incremental validity of TAPAS facet scores in predicting eight criteria measures beyond a composite measure of cognitive ability (Armed Forces Qualification Test [AFQT]), measuring arithmetic reasoning, word knowledge, paragraph comprehension, and numerical operations. For the dichotomous criterion variables (e.g., attrition), the logistic regression procedures provide no direct effect size estimate. Instead, Nagelkerke's (1991) *R* was used to obtain a pseudo estimate of *R* for these reported analyses. These results show that a significant additional proportion of the variance of various performance criteria can be accounted for by TAPAS scores, demonstrating the criterion-related validity of the TAPAS assessment in an army sample.

Workforce validity findings. Several sources of validity data for the WorkFORCE Assessment for Job Fit capability also come from ETS efforts to design a recruitment tool for a large global company. Data collection efforts involved incumbent tryout studies in multiple samples, including a sample of ETS incumbents and a sample of incumbents at the large global company's U.S. headquarters, as well as validation studies with job applicants in both America and the Philippines. A summary of these validity data appears below.

***Incumbent tryout studies*.** The goal of the incumbent tryout studies in both organizations was to examine the validity of the assessment before moving forward with implementing WorkFORCE Assessment for Job Fit operationally. In both organizations, the minimum education level for employees is college graduation. The trial demonstrated that pools of WorkFORCE Assessment for Job Fit items are able to reliably recover trait score estimates for each facet and that these facet scores correlate the same way in the college-educated workforce sample as they do in the high-school educated military TAPAS samples for which there was already a considerable validity database ($N > 500{,}000$).

Table 4

Incremental Validity Results for the Tailored Adaptive Personality Assessment System (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
4-month attrition	1,694	.03	.27	.24
6-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. From *Development of the Tailored Adaptive Personality Assessment (TAPAS) to Support Army Selection and Classification Decisions*, by F. Drasgow, S. Stark, O. S. Chernyshenko, C. D. Nye, C. L. Hulin, and L. A. White, 2012, Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences. AFQT = Armed Forces Qualification Test; AIT/OSUT = Advanced Individual Training/One-Station Unit Training; ΔR = Increment in multiple correlation when adding the 12 personality facets to the AFQT score; Nagelkerke's (1991) R was used for the dichotomous criterion variables (comprehensive graduate vs. discharged from training; 4- month attrition, 6-month attrition, AIT/OSUT-graduate vs. discharged).

The ETS Strategic Workforce Solutions (SWS) item tryout study was conducted first with a sample of 293 incumbent employees who completed a 15-facet adaptive version of the assessment drawing from a bank of 600 statements as part of a mock recruitment battery. Participants made choices from 120 pairs of statements. Table 5 demonstrates that interfacet correlations for this incumbent sample are largely similar to those of previous studies. Facets relating to conscientiousness and intellect (openness) predicted self-reported grade point average (GPA) and *SAT*® scores as expected.

Table 5

Interfacet Correlations for Educational Testing Service (ETS) Incumbent Sample (N = 293)

Facet	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Diligence														
2. Stability	-.07													
3. Collaboration	.1	.01												
4. Inquisitiveness	**.17** [a]	**.15** [a]	.02											
5. Assertiveness	**.25** [a]	**.15** [a]	0	**.24** [a]										
6. Calmness	-.05	**.23** [a]	**.26** [a]	.07	.06									
7. Experience seeking	.03	.11	.03	.11	**.17** [a]	.01								
8. Generosity	.07	-.03	**.21** [a]	**.14** [a]	0	**.12** [a]	.06							
9. Creativity	.11	**.12** [a]	-.09	**.35** [a]	**.3** [a]	-.02	.1	**.13** [a]						
10. Intellectual orientation	**.26** [a]	**.16** [a]	-.01	**.29** [a]	**.24** [a]	.09	0	.02	**.31** [a]					
11. Organization	**.19** [a]	-.08	-.03	**-.17** [a]	-.05	.01	-.11	-.09	-.09	-.09				
12. Dependability	**.39** [a]	.05	**.16** [a]	.1	**.13** [a]	**.13** [a]	-.02	**.13** [a]	.08	**.15** [a]	.02			
13. Self -discipline	**.25** [a]	.05	**.14** [a]	.1	-.04	**.15** [a]	**-.23** [a]	.02	**-.1** [a]	0	.08	**.25** [a]		
14. Friendliness	.07	**.17** [a]	**.25** [a]	.02	**.2** [a]	**.15** [a]	**.39** [a]	**.15** [a]	.07	-.07	.01	.11	-.1	
15. Optimism	**.12** [a]	**.36** [a]	**.13** [a]	.02	**.27** [a]	**.18** [a]	**.18** [a]	**.15** [a]	.1	**.13** [a]	0	.11	.08	**.25** [a]

[a] Correlations are significant at $p < .05$ (indicated with boldface).

A workforce incumbent tryout study was also conducted with a sample of 1,128 incumbents across four major job categories (consultants [C]; analysts [A]; customer service representatives [CSR]; and IT solutions staff, [IT]) at a global company's U.S. headquarters. Incumbents completed a 12-facet adaptive version of the FACETS capability that made use of a set of 286 FACETS statements, some of which were customized for the sample alongside a background questionnaire. These customizations largely involved grammatical or wording changes to make statements more work related and less negative in tone. Overall results are as follows:

- The intercorrelations among the 12 facets largely fit an expected pattern, based on prior TAPAS research (Drasgow et al., 2012), and are consistent with the workforce personality literature. For example, openness facets such as inquisitiveness, intellectual orientation, and creativity correlate moderately with each other (around *r* = .30) and agreeableness facets such as collaboration and friendliness were also moderately correlated (*r* = .34; see Table 6).

- Facets relating to conscientiousness and openness correlated with self-reported SAT scores and GPA as expected.
- Composite scores were computed for each of the four job categories (C, A, CSR, IT), to optimize prediction of performance ratings. For each job category, a different set of facet weights was used. These composites had significant positive correlations (*r* values ranging between .30 and .42, uncorrected) with annual performance review ratings for all four job categories. For a detailed overview of these results, please see the *Workforce Sample Research Report* (Naemi & Kyllonen, 2014).

Table 6

Interfacet Correlations for Incumbent Employees (N = 1,128)

Facet	1	2	3	4	5	6	7	8	9	10	11
1. Diligence											
2. Stability	**.12** [a]										
3. Collaboration	**.11** [a]	**.16** [a]									
4. Inquisitiveness	**.24** [a]	**.12** [a]	.00								
5. Assertiveness	**.43** [a]	**.21** [a]	.05	**.21** [a]							
6. Creativity	**.21** [a]	**.11** [a]	.01	**.40** [a]	**.26** [a]						
7. Intellectual orientation	**.28** [a]	**.24** [a]	-.03	**.38** [a]	**.36** [a]	**.35** [a]					
8. Organization	**.18** [a]	.00	.02	**-.07** [a]	**.09** [a]	**-.01** [a]	-.01				
9. Dependability	**.36** [a]	**.13** [a]	**.13** [a]	**.16** [a]	**.20** [a]	**.11** [a]	**.15** [a]	**.15** [a]			
10. Self-discipline	**.24** [a]	**.12** [a]	.03	**.11** [a]	.01	**.06** [a]	**.09** [a]	**.27** [a]	**.31** [a]		
11. Friendliness	**.14** [a]	**.14** [a]	**.34** [a]	.06	**.29** [a]	**.12** [a]	.05	-.02	.05	**-.13** [a]	
12. Optimism	**.21** [a]	**.36** [a]	**.26** [a]	**.09** [a]	**.22** [a]	**.10** [a]	**.13** [a]	**.09** [a]	**.22** [a]	**.17** [a]	**.30** [a]

[a] Correlations are significant at $p < .05$ (indicated with boldface).

United States and Philippines applicant studies. The goal of these validation studies was to examine the validity of FACETS with a sample of job applicants across four broad job categories (C, A, SCR, IT) in the United States and in the Philippines. The overall goals of the studies were to confirm the general validity of the assessment given significant revisions to item content and format as part of a customization process for the client and to examine the convergent and criterion validity of the assessment. The studies were launched in both the United States and the Philippines on April 15, 2013, and data collection was completed on December 31, 2013, resulting in 1,964 valid scores for U.S. applicants and 4,433 valid scores for Philippines applicants.

- Applicants completed a background questionnaire, the assessment battery itself (108 pairs of statements covering 12 facets mapped onto eight job performance factors identified by the company), and a follow-up reactions survey.
- Recruiters were also asked to complete a short survey about the applicant rating overall fit and interview decisions after the human resources interview.
- Finally, data on hiring status and behavioral interviews where available were sent for all applicants who had completed the assessment battery.

Summary Findings

- Results demonstrate statistically significant positive zero-order facet level correlations between facets and outcomes measures, such as recruiter ratings of overall fit, whether or not applicants are moved on to the next step of the hiring process, behavioral interview ratings, and whether or not applicants ultimately receive an offer. Recruiters did not see assessment scores as part of this process to ensure that scores had no influence on hiring decisions. For a detailed overview of these results, please see the *Workforce Sample Research Report* (Naemi & Kyllonen, 2014).
- Behavioral ratings from the interview process, which provided the best indicator of personality ratings among all the outcomes and criteria collected as a part of the study, had the strongest correlation with assessment score composites formed through empirically weighted regression formulas (with *r* values ranging between statistically significant values of .24 and .28), providing evidence that the assessment can be used as an indicator of behavioral fit for recruiters.
- Results from the follow up survey indicate that applicants report overwhelmingly positive user experiences (at least 90% of participants agree that the assessment was valuable, enjoyable, and worth taking as part of the hiring process). These positive responses, however, represent only the subset of job applicants who elected to participate in the study, and all responses were provided before the application process was complete.
- Taken together, these results help demonstrate criterion-related validity for FACETS as a selection tool by linking FACETS scores to hiring outcome measures.

Program for the International Assessment of Adult Competencies (PIAAC) U.S. incumbent validity findings. Another study to obtain validity evidence in support of the FACETS capability was conducted by ETS as part of preparations for a large scale study for the Program for the International Assessment of Adult Competencies (PIAAC) online assessment battery. A sample of 610 U.S. incumbents employed by ETS completed a fixed-length FACETS measure with 104 item pairs that measured 13 of the available personality scales, and supervisors rated these employees' job performance using a performance evaluation tool designed to serve as the criterion measure. The job performance measure supervisors used to evaluate study participants comprised 27 items that described various work-related behaviors. The 6-point response scale supervisors used ranged from 1 (*never*) to 6 (*always*). Performance scale scores were created by averaging responses for each of the six dimensions. The performance dimension scores (criteria for validity analyses) are described in Table 7. Means and standard deviations for the personality scales and performance dimensions are provided in Table 8. It can be seen that the average for each of the performance dimensions is high (or low for counterproductive work behavior), indicating a severe restriction in the distribution of criterion scores.

Table 7

Description of Criterion Measure (Job Performance Dimensions) for the Program for the International Assessment of Adult Competencies (PIAAC) Study

Performance dimension	Dimension description	Items	Internal consistency reliability (α)
General task performance	Items reflect behaviors that are formally recognized as part of an employee's duties. These behaviors contribute to organizational goals.	5	0.93
Safety and rule compliance	Items reflect behaviors that are related to the extent to which employees comply with safety-related rules and regulations.	4	0.91
Counterproductive work behavior	Items reflect voluntary behaviors that violate organizational norms and harm the interests/goals of the organization.	5	0.81
Teamwork	Items reflect behaviors that are related to the extent to which employees work well with others to meet work and organizational goals.	4	0.91
Customer service	Items reflect employee behaviors related to serving and helping customers.	3	0.96
Proactive work behavior	Items reflect employee behaviors that are related to a desire to do a good job and improve work products or processes.	6	0.92

Table 8

Means and Standard Deviations for Criterion and Predictor Measures for the Program for the International Assessment of Adult Competencies (PIAAC) Study

Criterion	Performance dimension	*N*	*M*	*SD*
Cognitive predictor	General task performance	610	5.04	.85
	Safety and rule compliance	256	5.46	.60
	Counterproductive work behavior	610	1.26	.42
	Teamwork	610	4.73	1.01
	Customer service	610	4.87	1.09
	Proactive work behavior	610	4.42	1.00
	Numeracy	313	312.30	79.88
	Literacy	286	362.94	70.01
Noncognitive predictor	Achievement	507	.25	.58
	Adjustment	507	-.39	.69
	Cooperation	507	-.33	.63
	Curiosity	507	.72	.74
	Dominance	507	-.14	.85
	Generosity	507	-.18	.66
	Ingenuity	507	.41	.97
	Intellectual efficiency	507	.33	.76
	Order	507	-.11	.84
	Responsibility	507	-.09	.50
	Self-control	507	.21	.66
	Sociability	507	-.56	.72
	Well being	507	-.28	.72

Table 9 shows correlations of personality facet scores with performance outcomes. This study was intended to be largely exploratory, and the data were expected to be restricted both on the predictor and criterion side. Overall, the relationships were expected to be in alignment with the construct map (see the section titled Forming Composites and Building Overall Selection Indices in this paper). Some observed correlations were in unexpected directions (e.g., lower adjustment was related to higher task performance). Although post hoc explanations could be constructed, the exploratory nature of this study (and not having sufficient samples to explore the moderating effects of job type) suggests that additional research is needed. Overall, however, Table 9 shows that each of the six performance dimensions can be predicted with scores on one or more of the personality facets; that a clear relationship between the facets within each higher order personality trait and the criteria does not exist; and that range restriction likely attenuated these relationships, as indicated by several lowered validity coefficients.

Table 9

Correlations of Facet Scores With Performance Criteria

Performance dimension	General task performance	Safety and rule compliance	Counterproductive work behavior	Teamwork	Customer service	Proactive work behavior
Diligence	**.14**[a]	.04	-.07	**.10**[a]	.08	**.18**[a]
Order	.08	.04	-.02	-.02	0	.01
Dependability	.08	**.19**[a]	-.06	0	.07	.08
Self-discipline	.04	.06	-.03	-.06	.05	.01
Adjustment	**-.10**[a]	0	.06	-.01	-.03	-.06
Optimism	**-.09**[a]	-.08	0	-.02	.01	-.06
Collaboration	-.01	.03	**-.10**[a]	.06	.05	0
Generosity	.05	.03	-.03	**.12**[a]	.04	**.12**[a]
Inquisitiveness	-.04	-.02	.02	**-.10**[a]	-.07	-.02
Ingenuity	-.04	-.03	**.15**[a]	-.06	.06	0
Intellectual orientation	.06	.04	.02	-.03	.05	.08
Friendliness	**-.10**[a]	-.06	.08	.08	-.01	-.04
Assertiveness	.04	.03	.09	.08	**.09**[a]	**.09**[a]

[a] Correlations are significant at *p* < .05 (also indicated with boldface).

Program for the International Assessment of Adult Competencies (PIAAC) international data. FACETS data were also collected as part of the PIAAC online noncognitive battery module, which was designed to capture critical factors associated with work training and workplace success in international samples. Data were collected using online test administration from multiple samples of at least 180 working adults in seven countries: Czech Republic, United States, Canada, Spain, Japan, Italy, and Ireland. All statements were translated and localized by a team of experienced translators (through a process in which multiple translators were separately assigned to translate each statement and resolve any differences in translations after systematic comparisons).

A forced-choice assessment was incorporated as part of the PIAAC personality module and scored via the FACETS engine. This form was composed of a fixed set of 104 pairwise preference items that measured 13 personality facets. To examine the factor structure of the personality facet scores relative to their expected relationship in the FFM framework, an EFA, with an oblique rotation was performed on the data for each country with at least 100 respondents. Overall, the expected pattern of factor loadings onto the intended factors was observed, providing some initial evidence of the construct validity of personality scores across countries and languages.

Overall, the results of these initial analyses demonstrate promising psychometric utility for all assessment modules, with future analyses set to examine measurement invariance across cultures, as well as relations to performance and success outcomes from supervisor ratings to further examine differential prediction issues across cultures.

In sum, a preponderance of evidence from content and criterion-related validity studies conducted in numerous samples help build a validity case for the WorkFORCE Assessment for Job Fit as a tool for selection, a case that is detailed in the next section of this paper.

Forming Composites and Building Overall Selection Indices

We address two major approaches to scoring the FACETS capability in this section. One underlying approach describes building a single overall selection index as part of a specific in/out hiring decision at any stage in the selection process. The second approach involves building and scoring a series of composites. We begin with a discussion of the composite approach and a review of various commonly used methods of weighting scores, followed by methods for building overall selection indices, and finally concluding with an illustration of both approaches for the WorkFORCE Assessment for Job Fit.

Composites for personality, also known as compound traits, combine several overall personality factors (e.g., FFM personality factors such as openness or conscientiousness) or personality facets (e.g., conscientiousness facets such as responsibility and order) to form a composite (Fein & Klein, 2011). Facets are selected to form composites using both theoretical and empirical evidence that support the prediction of specific focal criteria (Hough & Oswald, 2000). Composites are often used to better understand job candidate strengths and challenges and to highlight areas that are commonly identified as important determinants of workforce readiness and success.

Viswesvaran, Deller, and Ones (2007) suggested that composites may serve to improve the validity of selection processes. For instance, customer service orientation (Saxe & Weitz, 1982) has been shown to predict both manager and customer ratings of provided service (Baydoun, Rose, & Emperado, 2001; Hogan & Hogan, 1992). Moreover, substantial predictive validity estimates have been demonstrated for other compound traits that combine several dimensions of FFM personality traits, including constructs such as integrity, violence potential, and management potential (Ones, Viswesvaran, & Dilchert, 2005).

Because broad personality factors encompass several facets, they may contain elements that are both relevant and irrelevant to a given criterion and, consequently, may not afford maximal prediction (Hough & Oswald, 2000; Paunonen & Ashton, 2001). Thus, prediction can be improved by selecting narrow traits or facets into the composite as opposed to broad traits that may contain criterion irrelevant elements.

The use of broad versus narrow measures of personality has been a topic of ongoing debate in the psychological literature (Alge, Gresham, Heneman, Fox, & McMasters, 2002). On the one hand, if the criterion is complex, the most successful predictor will likely be a similarly complex composite. That is, if the bandwidth of the criterion is wide, the chosen predictor should be of relevance and similar bandwidth to the criterion (Hough & Ones, 2001). On the other hand, theory and empirical evidence supporting relations between narrow predictors and criteria can be more illustrative in helping researchers better understand the nature of validity coefficients for complex data, and from a utility perspective, organizations may at times be more interested in narrow criteria of performance on the job. Consequently, although the predictive validities for a composite may be meaningful, the relationships between the specific facets that form the composite and any outcome measure may be obscured (Hough & Oswald, 2008). Hough and Oswald (2008) pointed out that, for example, even if the composite of integrity is equally valid for two different criteria, conscientiousness may be most predictive of attendance, and emotional stability may be most predictive of customer service.

Forming Composites

Previous research (Fein & Klein, 2011) has approached the process of forming composites by first relating trait descriptions to behavioral outcomes using a rational methodology. Behavioral references contained in both scale items and facet-level definitions are examined for clear conceptual links to the behavioral criteria of interest. In this way, researchers can systematically exclude facets or traits that may not be directly related to the desired outcome. This approach reflects the steps that other researchers have followed to form compound trait scores (Alge et al., 2002; Hough & Oswald, 2000).

To this end, composites may be created by averaging scores from targeted facet or broad level traits within a given personality framework. Computing a weighted sum or average using the separate trait or facet scores is the most frequently employed method for forming composite scores (Kane & Case, 2004). Reliability estimates can then be computed for the composite as one would compute the composite of the unidimensional traits, while taking into account the covariance between the scores. Research by Kane and Case (2004) suggested that giving additional weight to the more reliable subscales will improve the overall reliability of the composite, although there are also potential tradeoffs for validity. Validity for the composite score is ultimately a function of its reliability, content, and relationship with criteria of interest (Corcoran, Downing, Tekian, & DaRosa, 2009).

Industrial and organizational (I/O) psychology performance literature competency models review. Composites developed for workforce samples should be consistent with existing broad and general taxonomies of work performance. In this section, we present a group of composites using the FACETS capability designed for WorkFORCE Assessment for Job Fit that is relevant to jobs of moderate complexity (i.e., O*NET Job Zone 3: secretaries, customer service representatives, interviewers); workers in these jobs are anticipated to make most frequent use of the assessment. Job performance dimensions have been identified at a high level of abstraction (e.g., task performance, organizational citizenship, and adaptability), a moderate level of abstraction (e.g., O*NET generalized work activities; Jeanneret, Borman, Kubisiak, & Hanson, 1999), and a specific level of abstraction (e.g., Tett, Guterman, Bleier, & Murphy's [2000] hyperdimensional taxonomy with dimensions such as problem awareness or decision making). See Johnson (2003) for a hierarchical summary of job performance dimensions.

Burrus, Jackson, Xi, and Steinberg (2013) computed descriptive statistics on the importance scale ratings for the O*NET knowledge, skills, abilities, and work style domains for Job Zones 3–5. For Zone 3, the top-ranked competencies were as follows:

1. Dependability
2. Attention to detail
3. Integrity
4. Cooperation
5. Self-control
6. Stress tolerance
7. Initiative
8. Adaptability/flexibility
9. Independence
10. Persistence

The mean importance ratings within Zone 3 jobs for these competencies ranged from 72–87 on a 0–100 scale (Burrus et al., 2013). All of these competencies are classified by O*NET as work styles, or, in other words, personality factors. This work informed the formation of both the composites and the performance dimensions that serve as our criterion variables by identifying and linking together important associated personality factors.

The next step in the process was to link the workforce dimensions to established taxonomic work on work performance dimensions. Johnson's (2003) work performance taxonomy, supplemented with additional work on the relationship between organizational citizenship behavior and counterproductive work behavior, was used. This taxonomy consists of three levels. At the highest level are task, citizenship, and adaptive performance. Task and citizenship performance are well established as separate aspects of the individual job performance domain in the field of industrial and organizational (I/O) psychology (see Borman & Motowidlo, 1997). Because of the increasingly dynamic nature of work environments, adaptive performance has been added, based on the work of Pulakos, Arad, Donovan, and

Plamondon (2000). *Adaptive performance* is the proficiency with which a person alters his or her behavior to meet the demands of the environment, an event, or a new situation (Pulakos et al., 2000). N. Schmitt, Cortina, Ingerick, and Wiechmann (2003) and Johnson (2003) suggested that adaptive performance is distinct from task and citizenship performance.

Sackett, Berry, Wiemann, and Laczo (2006) and Berry, Ones, and Sackett (2007) added to the job performance taxonomic domain by clarifying relationships between positive and negative aspects of nontask performance (i.e., organizational citizenship behaviors and counterproductive work behaviors) that had been treated separately in the literature. *Interpersonal deviance* refers to behaviors that are harmful to individuals in an organization, such as making fun of someone or playing a mean prank on someone. *Organizational deviance* refers to behaviors that are harmful or counterproductive to the organization itself, such as destruction or theft of organizational property, littering, discussing confidential information, or putting little effort into one's work. These two additional dimensions were evaluated empirically with confirmatory factor analysis. Treating them separately from organizational citizenship behavior dimensions resulted in a better fitting solution. As such, they are added to the Johnson (2003) taxonomy as Level 2 dimensions. Based on the previous review, we can now form a taxonomy consisting of four Level 1 dimensions from the literature (task, citizenship, adaptive, and counterproductive work performance) that encompass less broad dimensions of performance also found in the literature, which we describe as Level 2 dimensions (personal support, interpersonal deviance, dealing with uncertain and unpredictable work situations, organizational support, organizational deviance, non-job-specific task proficiency, written and oral communication proficiency, representing the organization to customers and the public, conscientious initiative, non-job-specific task proficiency). Table 10 provides a mapping of the ETS workforce performance dimensions to the first two levels of the literature-based taxonomy described above, as well as to six ETS job dimensions used in supervisor rating studies (these six job dimensions were also based on an integrative review of performance taxonomies from the I/O psychology literature).

Table 10

Educational Testing Service (ETS) Workforce Dimensions, Working Definitions, and Linkages to Integrated Work Performance Taxonomy

Dimension	Working definition	Links to Level 1 performance dimensions	Links to Level 2 performance dimensions	Links to ETS job performance dimensions
Teamwork and citizenship	Working with diverse groups of peers and colleagues; contributing to groups; having a healthy respect of different opinions, customs, and preferences; participating in group decision-making.	Citizenship performance Counterproductive performance	Personal support Interpersonal deviance	Teamwork
Flexibility and resilience	Adjusting well to changing or ambiguous work environments, handling stress, accepting criticism and feedback from others, being positive even when facing setbacks.	Adaptive performance Citizenship performance	Dealing with uncertain and unpredictable work situations Personal support	Proactive work behavior
Responsibility	Conducting oneself with responsibility, accountability, and excellence; adhering to organizational policies; being sensitive to and following safety and other regulatory rules and procedures; demonstrating appropriate workplace behavior and conduct	Citizenship performance Counterproductive performance Task performance	Organizational support Organizational deviance Interpersonal deviance	Safety and rule compliance Counterproductive work behavior
Customer service orientation	Conducting oneself in a courteous, patient, and cooperative manner with external or internal clients or customers; acting to meet client needs and maintain the role as spokesperson when dealing with others; following through with clients to get job done well; managing difficult people and assignments; putting the customer first.	Task performance Citizenship performance	Non-job-specific task proficiency Written and oral communication proficiency Representing the organization to customers and the public	Customer service

Dimension	Working definition	Links to Level 1 performance dimensions	Links to Level 2 performance dimensions	Links to ETS job performance dimensions
Initiative and perseverance	Reflecting behaviors formally recognized as part of job duties and which contribute to assigned work; completing tasks efficiently and accurately; acting as a self-starter; drives to get work accomplished.	Task performance Citizenship performance	Conscientious initiative	General task performance
Problem solving and ingenuity	Using knowledge, facts, and data to solve problems effectively; thinking critically and creatively; using good judgment when making decisions; being a self-directed learner.	Task performance Citizenship performance	Conscientious initiative Non-job-specific task proficiency	Proactive work behavior

As expected, the personality-based criteria were primarily linked to nontask dimensions at Level 1, though task performance is also occasionally represented with respect to certain composites given the nature of task performance as a central component of general evaluations of job performance (Borman & Motowidlo, 1997). It also bears mention that all of the Level 1 dimensions were represented. All five Level 2 nontask performance dimensions were represented. As such, our workforce performance dimensions encompass a substantial portion of the overall criterion space; especially the nontask performance criterion space, which is what personality variables have historically predicted best (e.g., Motowidlo, Borman, & Schmit, 1997).

WorkFORCE Assessment for Job Fit Composites and Overall Indices

Here we outline the approach that is taken for forming composite scores by one use case of the FACETS capability: the WorkFORCE Assessment for Job Fit. We describe how the WorkFORCE Assessment for Job Fit can be used as a tool for selection by making use of six composites, labeled as *behavioral competencies* for client-facing purposes, designed to be linked to job performance ratings.

Several important issues must be addressed when forming composites from facets. The first issue is to decide which facets form each composite. For example, should a teamwork and citizenship composite be composed of two facets, collaboration and generosity, or is friendliness also a facet to include in teamwork and citizenship? In order to address this issue, we formed a construct map (see Table B1) that determined facet groupings by considering meta-analytic evidence from the literature, the FFM theory of personality, and evidence from empirical studies that have made use of the FACETS approach both at ETS and during previous foundational research with TAPAS from the U.S. Army. This construct map is included in Appendix B and contains justifications for composite formation from both theory-based and evidence-based perspectives.

The primary theoretical basis for composite formation seen in the construct map (see Table B1) is the FFM of personality. As discussed in the introduction to this paper, the FFM provides a robust and well-tested approach for grouping facets together. As such, justification for each facet grouping is provided in the column labeled theoretical justification, drawing from the previously outlined TAPAS meta-analysis to link together appropriate facets from a construct validity perspective. In addition, from a criterion-related validity perspective, justification for facet groupings is also displayed by presenting criterion-related validity evidence that links facet scores to appropriate criterion measures, displayed as observed correlation values from the TAPAS meta-analysis. In this sense, if a group of facets all similarly predict a certain criterion (e.g., facets associated with agreeableness predicting customer service behaviors), then there is some further evidence for justification of a particular facet grouping. Finally, from a convergent validity perspective, facet intercorrelations are presented from a series of empirical validity studies occurring in the following data collections:

- Army Tech Report data collection
- ETSSWS item tryout
- Workforce U.S. CSRs
- Workforce U.S. IT
- Workforce U.S. analysts
- Workforce U.S. consultants
- Workforce U.S. applicants overall
- Workforce PHI applicants overall
- Program for the International Assessment of Adult Competencies (PIAAC; fixed form) for 7 countries and languages

Each of these studies was previously described in the Validity Evidence Related to WorkFORCE Assessment for Job Fit Scores section of this paper.

As a result, the following six behavioral competencies were formed with the intention of reflecting a broad criterion space of job performance, as outlined in the taxonomy of job performance presented in Table 10:

- *Initiative and perseverance* (diligence, assertiveness, dependability). Reflecting behaviors formally recognized as part of job duties and that contribute to assigned work; completing tasks efficiently and accurately; acting as a self-starter; and driving to get work accomplished
- *Responsibility* (dependability, self-discipline, organization). Conducting oneself with responsibility, accountability, and excellence; adhering to organizational policies; being sensitive to and following safety and other regulatory rules and procedures; and demonstrating appropriate workplace behavior and conduct

- *Teamwork and citizenship* (collaboration, generosity). Working with diverse groups of peers and colleagues; contributing to groups; having a healthy respect of different opinions, customs, and preferences; participating in group decision making
- *Customer service orientation* (friendliness, collaboration, generosity). Conducting oneself in a courteous, patient, and cooperative manner with external or internal clients or customers; acting to meet client needs and maintain the role as spokesperson when dealing with others; following through with clients to get a job done well; managing difficult people and assignments; putting the customer first
- *Problem solving and ingenuity* (creativity, intellectual orientation, inquisitiveness). Using knowledge, facts, and data to effectively solve problems; thinking critically and creatively; using good judgment when making decisions; being a self-directed learner
- *Flexibility and resilience* (stability, optimism). Adjusting well to changing or ambiguous work environments, handling stress, accepting criticism and feedback from others, being positive even when facing setbacks

Having formed six composites that are tied to relevant job performance dimensions, the next step is to determine an overall index that may be used for selection.

The *overall score index* is constructed as a combination of the underlying personality facets, and in the case of the WorkFORCE Assessment for Job Fit, this refers to the 13 facets measured by the assessment. The best prediction of a criterion may be found using a weighted combination of relevant personality facets. Such an approach provides the ability to specifically tailor each overall score index to a specific job category and requires the collection of appropriate criterion data from which to empirically derive the weights. Thus, to maximize prediction efficacy, in addition to providing six composite scores tied to various job dimensions, the WorkFORCE Assessment for Job Fit also provides a single overall score index. This overall score index is derived from an empirical weighting of the 13 personality facets via linear regression using supervisor ratings of performance. As outlined in Figure 3, the weighting scheme for the overall score index and composites are independent, providing optimal prediction for each set of scores.

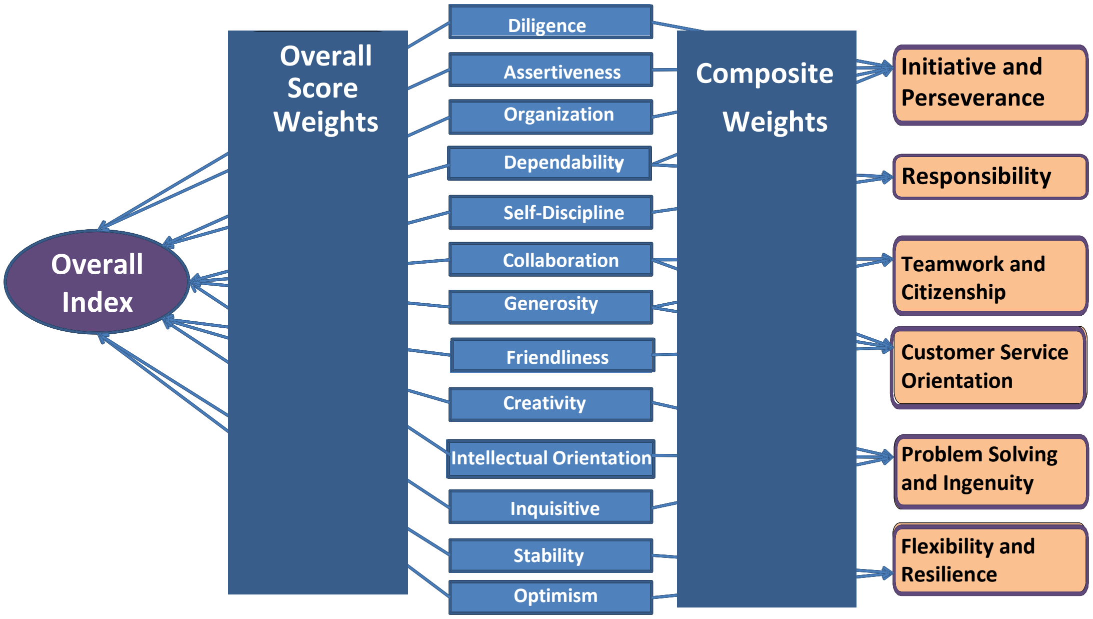

Figure 3. **Outline of the relationship between personality facets, composites, and overall scores when the overall score index is directly linked to facet scores.**

In this way, the WorkFORCE Assessment for Job Fit contains the flexibility to present both composite and overall scores in differing ways to meet varying client needs.

Fairness

An assessment of fairness and consideration of adverse impact is a crucial point to address for all ETS tests. Subgroup differences are a matter of importance not only from a legal defensibility standpoint, but also from a mission-based perspective of providing fair and equitable assessment for all. A vast literature documents significant and sizable subgroup differences in assessments of cognitive ability, but the literature for personality paints a different picture.

Regarding evidence of subgroup differences for assessments of personality, Hough, Oswald, and Ployhart (2001) conducted a meta-analysis in which they investigated subgroup differences by gender, age, and race/ethnicity. They found that for agreeableness the mean for women is about .40 standard deviations higher than for men. Within the domain of conscientiousness, they found facet-level differences. Specifically, women scored higher than men on dependability but about the same as men on achievement. They also found facet-level differences within the extraversion domain. Women scored higher on average than men on affiliation but lower on average on surgency (potency, dominance). Women scored lower on average than men on adjustment. With the exception of rugged individualism (essentially, masculinity), however, all of the subgroup differences could be characterized as *small*.[1] Subgroup differences involving age (40 or higher versus less than 40) and race/ethnicity were also small or nonexistent. An important conclusion of the Hough et al. meta-analysis was that facet-level measurement is beneficial in that broader five factor model (FFM) level measurement can obscure subgroup differences, even if subgroup differences may be *small*. Compared to younger working adults, older working adults (age ≥ 40) have a mean achievement score about .25 standard deviations lower than younger working adults (age < 40), a subgroup difference that can be characterized as *small*.

Costa, Terracciano, and McCrae (2001) found somewhat larger gender-related differences, though based on a smaller sample (1,000 U.S. adults). Women scored higher than men at approximately *medium* effect size levels (*d* >.40) on the anxiety and vulnerability facets of neuroticism. These results were largely replicated across cultures. Costa et al. noted that gender differences vary with culture and are most pronounced in individualistic cultures, like those found in Europe and North America. It is unclear whether these differences are an artifact of the self-report scales or representative of true differences that may be found using alternative strategies such as observer reports.

D. P. Schmitt, Realo, Voracek, and Allik (2008) administered the BFI (Benet-Martínez & John, 1998) to 17,637 individuals in 55 countries, 2,793 of whom were from the United States. Women scored higher on average than men on neuroticism (*d* = .53). The remaining effect sizes for the other four factors were *small*. The average across the 55 countries was comparable (*d* = .40) favoring women on neuroticism, with all other effect sizes close to zero. That said, there was a good deal of variability across the 55 countries on this gender difference (see D. P. Schmitt et al., 2008, Table 1, p. 173).

Soto, John, Gosling, and Potter (2011) examined age differences in FFM domains and facets from late childhood through middle age, using data from a very large cross-sectional sample (*N* = 1,267,218) assessed over the World Wide Web. With regard to neuroticism and its facet anxiety, they found that women scored higher than men on average but that *medium* effect sizes at age 30 became *small* effect sizes at age 50.

Foldes, Duehr, and Ones (2008) conducted a large-scale meta-analysis that focused specifically on race differences in personality. Their study used a single taxonomy to organize scales from 44 personality inventories according to the FFM factors as well as their facets (based on the Hough et al. [2001] results, which argued strongly for meta-analytic estimates at the facet level). While effect sizes were again *small*—at least when the comparisons were between Whites and other races/ethnicities—Foldes et al. noted that residual standard deviations and the size of the lower 90% credibility value suggested that there was room for substantial moderator effects and also noted areas of potential adverse impact for specific comparison groups below specific selection ratios (see Foldes et al., 2008, Table 8, p. 608). In most cases, the selection ratios (or the number of job applicants relative to the number of job positions) were low, but not in all cases. Again, however, the possibility of substantial moderator effects makes these data difficult to interpret. Based on their results, Foldes et al. drew the following conclusions:

> Some concerns for adverse impact may exist for Black [applicants] when Emotional Stability, anxiety, global Extraversion, and sociability scales are used. For Asian [applicants], adverse impact may be a concern when global Emotional Stability (as well as even tempered facet), global Extraversion (as well as dominance and sociability facets), and global Conscientiousness measures are assessed. For Hispanic [applicants], the only scale for which adverse impact concerns exist is the sociability facet of Extraversion. For American Indian [applicants], Emotional Stability and Extraversion measures could potentially result in adverse impact. There may also be concerns for adverse impact against White [applicants] on scales such as global Conscientiousness, and most facets of Conscientiousness (achievement, cautiousness, and order), global Extraversion, and self-esteem facet of Emotional Stability. On the flipside, we are encouraged to find that most effects (over 70%) we meta-analyzed and reported were small or negligible. (p. 611)

Drasgow et al. (2012) investigated gender differences in TAPAS, the instrument on which FACETS is based, in a large sample of military applicants. Other than the physical conditioning scale, which is not currently used in any FACETS use cases, all effect sizes favoring men over women were *small* or close to zero. Women scored .29 standard deviation units higher than men on adjustment, roughly comparable to the Hough et al. (2001) meta-analysis, and somewhat lower than the other gender difference studies cited above. It is also important to note the difference here between mean group differences and adverse impact, as adverse impact depends on cut scores, selection ratios, the distribution of scores, and applicant pool characteristics, such that large group differences can still lead to no adverse impact depending on how the test is used and how cut scores are set (and conversely, that small differences may still result in adverse impact based on selection ratios, cut scores, and the distribution of scores).

In addition, data from all ETS data collections highlighted in the criterion-validity section demonstrate that any FACETS level gender or ethnic differences in workforce samples are either small or not statistically significant.

- For the incumbent validity study in U.S. headquarters, results demonstrated minimal differences in score composite means by gender or ethnicity status. Score composite means in this case consisted of a single regression weighted FACETS score predicting a 3-year average of annual job performance ratings. When combining scores across all job types, women scored significantly higher (*M* = 2.71, *SD* = .24) than men (*M* = 2.67, *SD* = .22), *t* (693) = 1.25, *p* < .05, although the magnitude of this difference was *small*. When examining within job types, no statistically significant gender differences were found. Finally, when examining ethnicity differences across job types, no statistically significant differences were found between White incumbents (*N* = 370) and Asian incumbents (*N* = 131), African-American incumbents (*N* = 71), or Hispanic American incumbents (*N* = 40).
- For the applicant validity studies in both the United States and Philippines, results also demonstrated minimal differences in score composite means: A series of *t*-tests were performed in order to test whether any gender and ethnic differences for an overall regression-weighted FACETS score were statistically significant. In the U.S. overall sample, there is no statistically significant difference between men and women on overall FACETS score.
- In the U.S. overall sample, White applicants scored significantly higher (*M* = 2.64, *SD* = .22) than Black applicants (*M* = 2.61, *SD* = .22) and Asian applicants (*M* = 2.60, *SD* = .20) on overall FACETS score, although these effect sizes were small (Cohen's *d* = .15 and .19 respectively).

- In the Philippines overall sample, there was no significant difference between men and women on overall FACETS score.

Given these results across several data collections, we can conclude that gender and ethnic differences do not play a confounding role in FACETS score differences in several job samples of interest.

Item Fairness

All statements in the FACETS statement bank underwent a formal documented fairness and sensitivity review by trained ETS reviewers, who cleared all statements for desired use as a personality instrument in diverse group samples. These types of reviews are detailed in the *ETS Standards for Quality and Fairness* (2002) and the *ETS Guidelines for Fairness Review of Assessments* (2009). The purpose behind these publications and how they support ETS's fairness and sensitivity reviews are outlined on the ETS *Fairness Guidelines* web page:

> Educational Testing Service is committed to producing tests and other products that acknowledge the multicultural nature of society and treat its diverse populations with respect. Further, ETS is committed to ensuring that test takers and others who make up our increasingly diverse customer base enjoy equal access to our products. To meet these goals, ETS uses sensitivity reviews and differential item functioning (DIF) analysis. (*Fairness Guidelines*, n.d., para. 1).
>
> ETS mandates a process by which all its products—including individual test items, assessments, instructional materials, publications and other products—are evaluated for their sensitivity to and awareness of the contributions of various groups to the society of the United States. This sensitivity review also makes certain that ETS products do not use stereotyping and language, symbols, words or examples that are sexist, racist or otherwise offensive, inappropriate or negative toward any group (*Fairness Guidelines*, n.d., para 3).

Cross-Cultural Review

Research has supported the structural equivalence of the FFM constructs across a variety of countries and cultures to some extent. Whether the structure of the specific facets of the FFM measured by the FACETS capability hold in this instance is unknown. For within-country decisions (e.g., Australians only compared to Australians), there is limited concern about cross-country measurement invariance. If scores are being compared across countries, the need is to establish the measurement invariance of the scales, as well as a lack of differential prediction and possible lack of mean group differences or differential item functioning (DIF).

Forced-choice IRT models are in their infancy, and no methods for examining their invariance across populations have been developed. As a procedure to provide an initial review of statement translation and cross-culture comparability, ETS obtained local ratings of statement properties.

Local rating of statement properties. ETS conducted a series of local rater exercises to determine localized properties of FACETS statements in 17 countries (see Table 11). The procedure for these local rater exercises involved recruiting 10 to 20 respondents (meeting certain criteria) from each country, providing each of them with instructional training, and having those individuals rate 1,000 statements across 22 dimensions.

Table 11

Countries Where Local Rater Studies Were Conducted

Country	*N*
Brazil	13
Chile	17
China	11
Colombia	11
Costa Rica	10
Ecuador	10
France	11
Hong Kong	10
Indonesia	17
Japan	9
Korea	13
Mexico	10
Saudi Arabia	9
Spain	15
Taiwan	16
Thailand	9
Vietnam	16

Ideally, the raters in each country would be a representative sample of the FACETS test-taker target group, but given the variety of potential versions of the FACETS assessment to be used (and thus an undefined target population) and the practical difficulties in such a sampling scheme, it was decided that the raters should meet the following criteria:

- Bilingual in English and the target language
- At least a year of work experience and a college degree or equivalent
- Currently living in the country or have relocated to the United States no more than 3 years prior

An effort was also made to obtain a diversity of age and gender for the rater group in each country.

Once raters were identified, each individual participated in a training/instruction session with an ETS representative, generally lasting about 30 minutes. After reviewing the instructional materials, the training session involved participants completing a series of ratings for a set of example statements. The exercise was discussed and any questions that the participants had about the instructions and ratings were answered. At that point, raters received what we expected to be a sufficient amount of instruction to accurately complete the ratings.

Raters were asked to provide information on the following for each statement:

- Clarity of translation and content validity for jobs in general
- Appropriateness of the statement in the specific international context/culture where the job is situated
- An indicator of statement location
- An indicator of statement social desirability

This statement review activity was conducted on initial sample of 17 countries, and study results were used to identify potentially poorly translated statements or statements that were inappropriate or offensive in that particular context. Similarly, ratings of statement location were used as proxies for the IRT-based location parameters and used to identify items that may not be phrased in the same directional wording (e.g., a negatively worded statement translated as positive) and eliminated from that country’s statement pool.

Analysis of study data resulted in the retention of 36% to 86% of the total statement pool for each country, with those in East Asia showing the lowest rates of retention. The lower retention rates in certain countries were largely based on minor issues of clarity and translation that will be addressed in future localization exercises, allowing for a much greater percentage of statements to be retained. For those statements that were retained for use in each country, the social desirability estimates for each specific country were directly substituted in for the existing U.S.-based values. This analysis provided an empirical check beyond traditional translation activities, the results of which are expected to provide a strong foundation for operational assessment.

Discussion and Conclusions

Based on the information presented in this technical report, we can now consider recommendations for how the WorkFORCE Assessment for Job Fit and other assessments based on FACETS capability may be used in various contexts and use cases. Appendix C presents descriptions of several use cases for selection, describing options for the kinds of assessment scores that can be presented to clients, along with concerns or issues that should be taken to account with each score reporting approach. Thinking carefully about client goals involving the FACETS capability and checking an appropriate user's guide for any given use case is an essential step in taking advantage of the benefits provided by FACETS.

Future Directions

The process of validation of the WorkFORCE Assessment for Job Fit is ongoing, and as more data are collected across cultures and job types, further demonstration of the utility of FACETS as a tool for selection can be obtained. Utility can also be estimated based on validity coefficients (see Hunter & Hunter, 1984; Schmidt & Hunter, 1998). FACETS also represents a rich source for future research. Given the flexibility of the capability and the novelty of the psychometric approach, future research directions will examine questions surrounding scoring methods, algorithm use, cross cultural comparisons, and relations to differing work outcomes.

For example, one line of ongoing research seeks to expand the knowledge related to MFC measures through model development, data simulation, empirical analysis, and scale development. This broad range of activity will address critical issues associated with the expanded use of MFC scales including item analysis, scale development, and the evaluation of conditions under which these measures can both perform adequately and accurately recovery trait estimates.

One topic of particular interest undergoing current development is the use of estimation strategies to obtain statement parameter estimates directly from forced-choice response data rather than using the two-stage estimation approach (see the section titled Psychometric Background of the WorkFORCE Assessment for Job Fit in this paper). Such a capability would allow for the streamlined development of new statements to supplement existing statement pools and the development of methods for assessing measurement invariance and measurement equivalence across countries and languages. This research will extend to strategies for examining model-data fit for these item types and how to extend the methods to longitudinal data.

In addition, given the flexibility of the FACETS capability, future research goals include expanding the construct space of what is measured by the assessment beyond personality. For example, one goal is to develop a forced-choice measure of interests for the use of selection, given the body of literature linking interests to education and work outcomes (Lubinski, 2000; Nye, Su, Rounds, & Drasgow, 2012).

These projects will enhance and inform ongoing efforts using forced-choice measures and support a number of key project areas at ETS including WorkFORCE Assessment for Job Fit, PIAAC, English to Speakers of Other Languages (ESOL), and other work client projects that are in development.

Finally, although this report focuses on the use case of selection, further research and projects surrounding the use of the FACETS capability as a tool for development and training is also planned, helping to expand the space of what FACETS can do and what information can be provided to companies, workers, schools, students, applicants, and other populations.

Summary

To summarize, the purpose of this research report was to provide information on the WorkFORCE Assessment for Job Fit along with details on the FACETS capability and its use for selection. We provided an overview of the development of FACETS through a review of personality assessment and the psychometric models underpinning the capability, described validity evidence for the assessment from numerous sources of data, reviewed scoring options and considerations, described concerns around fairness and cross cultural comparability, and finally concluded with examples of use cases for reporting options. In this way, we have provided a foundation for the use of the FACETS capability for the WorkFORCE Assessment for Job Fit, as well as demonstrated the value of this capability for selection across job types and cultures.

References

Alge, B. J., Gresham, M. T., Heneman, R. L., Fox, J., & McMasters, R. (2002). Measuring customer service orientation using a measure of interpersonal skills: A preliminary test in a public service organization. *Journal of Business and Psychology*, *16*, 467–476.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

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

Baydoun, R., Rose, D., & Emperado, T. (2001). Measuring customer service orientation: An examination of the validity of the customer service profile. *Journal of Business and Psychology, 15*, 605–620.

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait method analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729–750.

Berry, C. M., Ones, D. S., & Sackett, P. R (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*, 410–424.

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

Boudreau, J. W., Boswell, W. R., Judge, T. A., & Bretz, R. D. (2001). Personality and cognitive ability as predictors of job search among employed managers. *Personnel Psychology*, *54*, 25–50.

Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O*NET)* (Research Report No. RR-13-21). Princeton, NJ: Educational Testing Service.

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*. Mahwah, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Corcoran, J., Downing, S. M., Tekian, A., & DaRosa, D. A. (2009). Composite score validity in clerkship grading. *Academic Medicine, 84*, S120–S123.

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.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81,* 322–331.

Drasgow, F., Stark, S, Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (Technical Report No. 1311). Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author.

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.

Fairness guidelines: Goals for fairness review. (n.d.). Retrieved from the Educational Testing Service website: https://www.ets.org/about/fairness/guidelines

Fein, E. C., & Klein, H. J. (2011). Personality predictors of behavioral self-regulation: Linking behavioral self-regulation to five-factor model factors, facets, and a compound trait. *International Journal of Selection and Assessment, 19,* 132–144.

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five racial groups. *Personnel Psychology, 61*, 579–616.

Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology, 7,* 201–208.

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., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Golubovich, J., & Chatterjee, D. (2014). *ETS Job Profiler for validity studies: Overview and planned application.* Manuscript in preparation.

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology, 35*, 407–412. doi:10.1037/h0058853

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

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270–1285.

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

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, 75*, 581–595.

Hough, L. M., & Ones, D. S. (2001). 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. 233–377). London, England: Sage.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future—remembering the past. *Annual Review of Psychology, 51*, 631–664.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 272–290.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Jeanneret, P. R., Borman, W. C., Kubisiak, U. C., & Hanson, M. A. (1999). Generalized work activities. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 105–125). Washington, DC: American Psychological Association.

Johnson, J. W. (2003). Toward a better understanding of the relationship between personality and individual job performance. In M. R. Barrick & A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 83–120). New York, NY: Jossey-Bass.

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Knapp, D. J., & Heffner, T. S. (2010). *Expanded enlistment eligibility metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report No. 1267). Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA523962

Koenig, J. A., & Roberts, J. S. (2007). Linking parameters estimated with the generalized graded unfolding model: A comparison of the accuracy of characteristic curve methods. *Applied Psychological Measurement, 31,* 504–524.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: “Sinking shaft at a few critical points.” *Annual Review of Psychology, 51*, 405–444.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45,* 935–974.

McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment, 11*, 65–276.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450–470.

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.

Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C., III (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*, 348–355.

Naemi, B., & Kyllonen, P. (2014). *Workforce sample research report*. Manuscript in preparation.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Nye, C. D., Drasgow, F., Chernyshenko, O. S., Stark, S, Kubisiak, U. C, White, L. A., & Jose, I. (2012). *Assessing the Tailored Adaptive Personality Assessment System (TAPAS) as an MOS qualification instrument* (Tech. Rep.). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Nye, C. D., Su. R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7*, 384–403.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance, 18*, 389–404.

Oswald, F. L., & Hough, L. M. (2011). Personality and its assessment in organizations: Theoretical and empirical developments. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Vol. 2. Selecting and developing members for the organization* (pp. 153–184). Washington, DC: American Psychological Association.

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

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality & Social Psychology, 81*, 524–539.

Porchea, S. F., Allen, J., Robbins, S., & Phelps, R. P. (2010). Predictors of long-term enrollment and degree outcomes for community college students: Integrating academic, psychosocial, socio-demographic, and situational factors. *The Journal of Higher Education, 81*, 750–778.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85*, 612–624.

Press, W. H. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.) Cambridge, UK: Cambridge University Press.

Ricci et al. v. DeStefano et al., 557 U. S. 557 (2009).

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387.

Robbins, S., Allen, J., Casillas, A., Peterson, C., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98,* 598–616.

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

Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35*, 259–279.

Sackett, P. R., Berry, C. M., Wiemann, S. A., & Laczo, R. M. (2006). Citizenship and counterproductive work behavior: Clarifying relationships between the two domains. *Human Performance, 19,* 441–464.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50,* 707–721.

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

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929–954.

Salgado, J. F., & Táuriz, G. (2012). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23,* 3–30.

Sanchez, J. I., & Levine, E. L. (2009). What is (or should be) the difference between competency modeling and traditional job analysis? *Human Resource Management Review, 19,* 3–63.

Saxe, R., & Weitz, B. (1982). The SOCO scale measure of the customer orientation of salespeople. *Journal of Marketing Research, 19,* 343–351.

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

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can’t a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94,* 168–182.

Schmitt, N., Cortina, J. M., Ingerick, M. J., & Wiechmann, D. (2003). Personnel selection and employee performance. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 77–105). Hoboken, NJ: Wiley.

Seybert, J., Stark, S., & Chernyshenko, O. S. (2014). Detecting DIF with ideal point models: A comparison of area and parameter difference methods. *Applied Psychological Measurement, 38,* 151–165.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299–321.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). College Park, MD: Author.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*, 330–348.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29,* 184–203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15,* 463–487.

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.

Tett, R. P., Guterman, H. A., Bleier, A., & Murphy, P. J. (2000). Development and content validation of a “hyperdimensional” taxonomy of managerial competence. *Human Performance, 13*, 205–251.

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.

Uniform guidelines on employee selection procedures, 43 Fed. Reg. 38290–38315 (1978).

Viswesvaran, C., Deller, J., & Ones, D. S. (2007). Personality measures in personnel selection: Some new contributions. *International Journal of Selection and Assessment, 15*, 354–358.

Weiss, D. J., & Kingsbury, G. G. (1984). Applications of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

White, L. A., & Young, M. C. (1998, April). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the meeting of the American Psychological Association, San Francisco, CA.

Woo, S. E., Chernyshenko, O. S., Longley, A., Zhang, Z. X., Chiu, C. Y., & Stark, S. E. (2014). Openness to experience: Its lower level structure, measurement, and cross-cultural equivalence. *Journal of Personality Assessment, 96*, 29–45.

Ziegler, M., MacCann, C., & Roberts, R. (Eds.). (2011). *New perspectives on faking in personality assessment*. Oxford, England: Oxford University Press.

Notes

[1] In this paper, we refer to effect size differences as *small, medium,* and *large* to note the common nomenclature associated with effect size for Cohen's *d* statistic (where small refers to values closest to .20, medium to values closest to .50, and large to values closest to .80). Although even Cohen (1988) himself noted that these effect size guidelines cannot be applied broadly across fields, we provide these terms as a general guide for the reader, italicized to indicate their reference to technical terms. A general caveat for interpretation is that even if an effect size difference is statistically significant, it may or may not be practically significant in terms of adverse impact. Context matters.

Appendix A

Summary of Scoring Algorithm for FACETS

Formally, the probability of preferring statement *s* to statement *t* in a pairwise preference item is given by

$$P(s > t)_i(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\} + P_{st}\{0,1\}} = \frac{P_s(1)P_t(0)}{P_s(1)P_t(0) + P_s(0)P_t(1)}, \qquad \text{(A1)},$$

where

i = the index for each pairwise preference item, where *i* = 1 to I;

s, *t* = the indices for the first and second statements, respectively, in an item;

d = the dimension associated with a given statement, where *d* = 1, … , *D*;

$\theta_{d_s}, \theta_{d_t}$ = the latent trait values for a respondent on dimensions d_s and d_t, respectively;

$P_{st}\{1,0\}$ = the joint probability of selecting statement *s*, and not selecting statement *t*;

$P_{st}\{0,1\}$ = the joint probability of selecting statement *t*, and not selecting statement *s*;

$P_s(1), P_t(1)$ = the probabilities of endorsing statements *s* and *t*, respectively;

$P_s(0), P_t(0)$ = the probabilities of not endorsing statements *s* and *t*, respectively; and

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

Stark et al. (2005) chose GGUM (J. S. Roberts et al., 2000) as the basis for test construction and scoring. When applied in the FACETS engine, parameter estimates for dichotomous statements are needed. Specifically, the GGUM is used to compute statement agreement probabilities that underlie *most like* selections. If $P(0)$ and $P(1)$ represent the respective probabilities of disagreeing (*Z* = 0) and agreeing (*Z* = 1) with a particular statement, given a respondent's latent trait score (θ) on the dimension that statement represents, and three statement parameters (α, δ, τ) reflecting discrimination, location (extremity), and threshold, respectively, then

$$P(0) = P(Z = 0 | \theta) = \frac{1 + \exp(\alpha[3(\theta - \delta)])}{1 + \exp(\alpha[3(\theta - \delta)]) + \exp(\alpha[(\theta - \delta) - \tau]) + \exp(\alpha[2(\theta - \delta) - \tau])}, \qquad \text{(A2)}$$

and

$$P(1) = (Z = 1|\theta) = \frac{\exp(\alpha[(\theta-\delta)-\tau])+\exp(\alpha[2(\theta-\delta)-\tau])}{1+\exp(\alpha[3(\theta-\delta)])+\exp(\alpha[(\theta-\delta)-\tau])+\exp(\alpha[2(\theta-\delta)-\tau])}. \quad \text{(A3)}$$

Administration of MDPP tests result in a pattern of responses across the pairwise preference items, $\tilde{\mathbf{u}} = (u_1, u_2, \ldots, u_n)$. The scoring of the response pattern is accomplished through a Bayes modal estimation strategy, where the vector of latent trait scores $\tilde{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_D)$, are obtained by maximizing

$$L\left(\tilde{\mathbf{u}},\tilde{\boldsymbol{\theta}}\right) = \left\{\prod_{i=1}^{n}\left[P_{(s>t)_i}\right]^{u_i}\left[1-P_{(s>t)_i}\right]^{1-u_i}\right\} * f\left(\tilde{\boldsymbol{\theta}}\right), \quad \text{(A4)}$$

Where $f(\tilde{\boldsymbol{\theta}})$ is a *D*-dimensional prior density function of the product of independent normals. Because it is more tractable to work with logs, Equation A4 can be rewritten as

$$\ln L\left(\tilde{\mathbf{u}},\tilde{\boldsymbol{\theta}}\right) = \sum_{i=1}^{n}\left[u_i \ln P_{(s>t)_i} + (1-u_i)\ln\left(1-P_{(s>t)_i}\right)\right] + \sum_{d'=1}^{D}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\theta_{d'}^2}{2\sigma^2}\right]. \quad \text{(A5)}$$

This leaves the following set of equations to be solved to obtain a vector of latent trait scores:

$$\frac{\partial \ln L}{\partial \tilde{\boldsymbol{\theta}}} = \begin{bmatrix} \dfrac{\partial \ln L}{\partial \theta_{d'=1}} \\ \dfrac{\partial \ln L}{\partial \theta_{d'=2}} \\ \vdots \\ \dfrac{\partial \ln L}{\partial \theta_{d'=D}} \end{bmatrix} = 0. \quad \text{(A6)}$$

These equations can be solved using a variety of numerical methods. The FACETS algorithm solves this using a quasi-Newtonian approach called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which performs *D*-dimensional minimization. This algorithm is implemented using the numerical recipes DFPMIN routine (Press, 2007).

The solving of Equation A6 requires the first partial derivative elements detailed in Stark et al. (2005), which in turn are composed of the first partial derivatives of the GGUM equation, as detailed in J. S. Roberts et al. (2000).

Appendix B

WorkFORCE Assessment for Job Fit Construct Map and Empirical Support

Table B1

WorkFORCE Assessment for Job Fit Construct Map

Composite	Composite definition	FACETS facet name	FACETS facet definition	Theoretical justification and criterion validity	Empirical support
Initiative and perseverance	Reflecting behaviors formally recognized as part of job duties and that contribute to assigned work; completing tasks efficiently and accurately; acting as a self-starter; drives to get work accomplished.	Diligence	High-scoring individuals are seen as "hard working, ambitious, confident, and resourceful" (Drasgow et al., 2012, p. 17) without direct oversight. They can multitask and produce results. They demonstrate initiative to advance skill levels.	Theoretical justification for this grouping comes from use of the five factor model groupings of conscientiousness, which incorporates the TAPAS defined facets of diligence and dependability. Assertiveness is included in this grouping, despite being an extraversion facet, based on significant positive correlations with conscientiousness facets in workforce samples (see empirical support column).	Please see Table B2 below for the complete summary of facet intercorrelations.
		Assertiveness	"High scoring individuals are domineering, 'take charge' and are often referred to by their peers as 'natural leaders'" (Drasgow et al., 2012, p. 39)	Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to general task performance: "Items reflect behaviors that are formally recognized as part of an employee's duties. These behaviors contribute to organizational goals."	
		Dependability	High scoring individuals have a strong work ethic, "are dependable, reliable and make every effort to keep their promises" (Drasgow et al., 2012, p. 39)	Correlations from TAPAS Drasgow meta-analysis (uncorrected): job/task performance and diligence, *r* = .05 assertiveness, *r* = .03 dependability, *r* = .11 assertiveness and diligence, *r* = .13 assertiveness, *r* = .14 dependability, *r* = .11	

Composite	Composite definition	FACETS facet name	FACETS facet definition	Theoretical justification and criterion validity	Empirical support
Responsibility	Conducting oneself with responsibility, accountability and excellence; adhering to organizational policies; being sensitive to and following safety and other regulatory rules and procedures; demonstrating appropriate workplace behavior and conduct.	Dependability	"High scoring individuals are dependable, reliable and make every effort to keep their promises" (Drasgow et al., 2012, p. 39).	Theoretical justification for this grouping comes from use of the five factor model groupings of conscientiousness, which incorporates the TAPAS defined facets of dependability, self discipline, and organization. Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to counterproductive work behaviors: Items reflect voluntary behaviors that violate organizational norms and harm the interests/goals of the organization. Correlations from TAPAS Drasgow meta-analysis (uncorrected): counterproductive behaviors and dependability, *r* =-.19 self-discipline, *r* =-.18 organization, *r*= -.18	Please see Table B3 for the complete summary of facet intercorrelations.
		Self-Discipline	"High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient" (Drasgow et al., 2012, p. 17).		
		Organization	"High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings" (Drasgow et al., 2012, p. 39).		
Teamwork and citizenship	Working with diverse groups of peers and colleagues; contributing to groups; having a healthy respect of different opinions, customs, and preferences; participating in group decision-making.	Collaboration	"High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with" (Drasgow et al., 2012, p. 32). They are able to bring together cultural and social differences to increase work effectiveness, innovation and quality.	Theoretical justification for this grouping comes from use of the five factor model groupings of agreeableness, which incorporates the TAPAS defined facets of collaboration and generosity. Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to teamwork: Items reflect behaviors that are related to the extent to which employees work well with others to meet work and organizational goals. Correlations from TAPAS Drasgow meta-analysis (uncorrected): contextual performance and collaboration, *r* =.12 generosity, *r* =.13	Please see Table B4 for correlations between collaboration and generosity.
		Generosity	"High scoring individuals are generous with their time and resources" (Drasgow et al., 2012, p. 32).		

Composite	Composite definition	FACETS facet name	FACETS facet definition	Theoretical justification and criterion validity	Empirical support
Customer service orientation	Conducting oneself in a courteous, patient and cooperative manner with external or internal clients or customers; acting to meet client needs and maintain the role as spokesperson when dealing with others; following through with clients to get job done well; managing difficult people, and assignments; putting the customer first.	Collaboration	“High scoring individuals are pleasant, trusting, cordial, non-critical, and easy to get along with” (Drasgow et al., 2012, p. 32). They are able to bring together cultural and social differences to increase work effectiveness, innovation, and quality.	Theoretical justification for this grouping comes from use of the five factor model groupings of agreeableness, which incorporates the TAPAS defined facets of collaboration and generosity. Friendliness is included in this grouping, despite being an extraversion facet, based on significant positive correlations with agreeableness facets in workforce samples (see empirical support column).	Please see Table B4 for the complete summary of facet intercorrelations.
		Generosity	“High scoring individuals are generous with their time and resources” (Drasgow et al., 2012, p. 32)	Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to customer service: Items reflect employee behaviors related to serving and helping customers.	
		Friendliness	“High scoring individuals tend to seek out and initiate social interactions” (Drasgow et al., 2012, p. 32) that can positively affect professional organizations.	Correlations from TAPAS Drasgow meta-analysis (uncorrected): contextual performance and collaboration, *r* = .12 generosity, *r* =.13 friendliness, *r* = .05	
Problem solving and ingenuity	Using knowledge, facts, and data to solve problems effectively; thinking critically and creatively; using good judgment when making decisions; being a self-directed learner.	Creativity	“High scoring individuals are inventive and think ‘outside of the box’” (Drasgow et al., 2012, p. 40).	Theoretical justification for this grouping comes from use of the five factor model groupings of openness, which incorporates the TAPAS-defined facets of creativity, intellectual orientation, and inquisitiveness.	Please see Table B5 for the complete summary of facet intercorrelations.
		Intellectual orientation	“High scoring individuals are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual” (Drasgow et al., 2012, p. 21)	Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to proactive work behavior: Items reflect employee behaviors that are related to a desire to do a good job and improve work products or processes. Correlations from TAPAS Drasgow meta-analysis (uncorrected):	

Composite	Composite definition	FACETS facet name	FACETS facet definition	Theoretical justification and criterion validity	Empirical support
		Inquisitiveness	“High scoring individuals are inquisitive and perceptive” (Drasgow et al., 2012, p. 40). They are interested in going beyond basics to expand their own learning and seek opportunities to gain expertise and knowledge. They believe in learning as a lifelong process.	academic performance or ability and creativity, *r* = .10 intellectual orientation, *r* = .26 inquisitive orientation, *r* = .23 training performance and creativity, *r* = .04 intellectual orientation, *r* = .09 inquisitive orientation, *r* = .09	
Flexibility and resilience	Adjusting well to changing or ambiguous work environments, handling stress, accepting criticism and feedback from others, and being positive even when facing setbacks.	Stability	High scoring individuals are well adjusted, adapt well to various job responsibilities, schedules, and contexts. They work well with changing and evolving work, can function effectively with ambiguity, and handle stress well.	Theoretical justification for this grouping comes from use of the five factor model groupings of emotional stability, which incorporates the TAPAS-defined facets of stability and optimism. Using data derived from the Drasgow TAPAS meta-analysis, facet scores should be linked to adaptability. Correlation from TAPAS Drasgow meta-analysis (uncorrected): adaptability and stability, *r* = .12 optimism, *r* = .08	Please see Table B6 for correlations between stability and optimism.
		Optimism	“High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being” (Drasgow et al., 2012, p. 40). They are able to incorporate and provide feedback, criticism, and praise well. They can cope effectively with setbacks.		

Table B2

Empirical Support: Correlations Between Facets Used to Form Initiative and Perseverance Composite

	Diligence	*r*	Assertiveness	*r*
Assertiveness	Workforce U.S. incumbents overall	.43		
	Workforce U.S. applicants overall	.36		
	Workforce PHI applicants overall	.45		
	Workforce U.S. CSRs	.30		
	Workforce U.S. IT	.42		
	Workforce U.S. analysts	.40		
	Workforce U.S. consultants	.26		
	ETS SWS item tryout	.25		
	PIAAC (fixed form) for seven countries and languages overall [a]	.34		
	Army Tech Report[b]	.32		
	Average correlations across six data collections	.35		
Dependability	Workforce U.S. incumbents overall	.36	Workforce U.S. incumbents overall	.20
	Workforce U.S. applicants overall	.31	Workforce U.S. applicants overall	.25
	Workforce PHI applicants overall	.37	Workforce PHI applicants overall	.26
	Workforce U.S. CSRs	.24	Workforce U.S. CSRs	.20
	Workforce U.S. IT	.36	Workforce U.S. IT	.28
	Workforce U.S. analysts	.30	Workforce U.S. analysts	.30
	Workforce U.S. consultants	.26	Workforce U.S. consultants	.19
	ETS SWS item tryout	.39	ETS SWS item tryout	.13
	PIAAC (fixed form) for seven countries and languages, overall [c]	.17	PIAAC (fixed form) for seven countries and languages, overall [d]	.04
	Army Tech Report, AIM dependability[b]	.17	Army Tech Report, AIM dependability[b]	.01
	Average correlations across six data collections	.29	Average correlations across six data collections	.15

Note. PHI = Philippines; CSRs = customer service representatives; IT = information technology; SWS = Strategic Workforce Solutions; Army Tech Report = Drasgow et al. (2012); AIM = Assessment of Individual Motivation.

[a] With *r* ranging from .19 (Canadian) to .36 (Czech). [b] *N* = 102k military applicants. [c] With *r* ranging from -.05 (Spain) to .22 (U.S.). [d] With *r* ranging from -.10 (Ireland) to.15 (Japan).

Table B3

Empirical Support: Correlations Between Facets Used to Form Responsibility Composite

	Dependability	r	Organization	r
Organization	Workforce U.S. incumbents overall	.15		
	Workforce U.S. applicants overall	.07		
	Workforce PHI applicants overall	.17		
	Workforce U.S. CSRs	.11		
	Workforce U.S. IT	.10		
	Workforce U.S. analysts	.06		
	Workforce U.S. consultants	.02		
	ETS SWS item tryout	.02		
	PIAAC (fixed form) for seven countries and languages overall[a]	.09		
	Army Tech Report, AIM dependability[b]	.06		
	Average correlations across six data collections	.09		
Self-discipline	Workforce U.S. incumbents overall	.31	Workforce U.S. incumbents overall	.27
	Workforce U.S. applicants overall	.21	Workforce U.S. applicants overall	.23
	Workforce PHI applicants overall	.27	Workforce PHI applicants overall	.22
	Workforce U.S. CSRs	.21	Workforce U.S. CSRs	.19
	Workforce U.S. IT	.26	Workforce U.S. IT	.21
	Workforce U.S. analysts	.18	Workforce U.S. analysts	.44
	Workforce U.S. consultants	.16	Workforce U.S. consultants	.39
	ETS SWS item tryout	.25	ETS SWS item tryout	.08
	PIAAC (fixed form) for seven countries and languages overall[c]	.18	PIAAC (fixed form) for seven countries and languages overall[d]	.18
	Army Tech Report, correlation between AIM dependability and self-control [b]	.19	Army Tech Report, correlation between AIM dependability and self-control[b]	.15
	Average correlations across five data collections	.24	Average correlations across six data collections	.39

Note. PHI = Philippines; CSRs = customer service representatives; IT = information technology; SWS = Strategic Workforce Solutions; Army Tech Report = Drasgow et al. (2012); AIM = Assessment of Individual Motivation.

[a] With *r* ranging from .04 (Ireland) to.16 (Italy). [b] *N* = 102k military applicants. [c] With *r* ranging from .03 (Japan) to .23 (Ireland). [d] With *r* ranging from .02 (Japan) to .27 (Czech Republic).

Table B4

Empirical Support: Correlations Between Facets Used to Form Customer Service Orientation Composite

	Collaboration	*r*	Generosity	*r*
Generosity	ETS SWS item tryout	.21		
	PIAAC (fixed form) for seven countries and languages overall[a]	.17		
	Average correlations across six data collections	.19		
Friendliness	Workforce U.S. incumbents overall	.34		
	Workforce U.S. applicants overall	.31		
	Workforce PHI applicants overall	.31		
	Workforce U.S. CSRs	.37		
	Workforce U.S. IT	.26		
	Workforce U.S. analysts	.28		
	Workforce U.S. consultants	.34		
	Army Tech Report, correlation between cooperation and sociability[b]	.18		
	ETS SWS item tryout	.25	ETS SWS item tryout	.15
	PIAAC (fixed form) for seven countries and languages overall[c]	.35	PIAAC (fixed form) for seven countries and languages overall[d]	.10
	Average correlations across five data collections	.27	Average correlations across six data collections	.13

Note. PHI = Philippines; CSRs = customer service representatives; IT = information technology; Army Tech Report = Drasgow et al. (2012); SWS = Strategic Workforce Solutions.

[a] With *r* ranging from .07 (Italy) to .22 (Japan). [b] *N* = 102k military applicants. [c] With *r* ranging from .17 (U.S.) to .35 (Japan). [d] With *r* ranging from -.08 (Spain) to .21 (Ireland).

Table B5

Empirical Support: Correlations Between Facets Used to Form Problem Solving and Ingenuity Composite

	Creativity	*r*	Intellectual orientation	*r*
Intellectual orientation	Workforce U.S. incumbents overall	.35		
	Workforce U.S. applicants overall	.32		
	Workforce PHI applicants overall	.35		
	Workforce U.S. CSRs	.28		
	Workforce U.S. IT	.34		
	Workforce U.S. analysts	.30		
	Workforce U.S. consultants	.34		
	ETS SWS item tryout	.31		
	PIAAC (fixed form) for seven countries and languages overall[a]	.24		
	Average correlations across six data collections	.31		
Inquisitiveness	Workforce U.S. incumbents overall	.40	Workforce U.S. incumbents overall	.38
	Workforce U.S. applicants overall	.31	Workforce U.S. applicants overall	.30
	Workforce PHI applicants overall	.30	Workforce PHI applicants overall	.29
	Workforce U.S. CSRs	.30	Workforce U.S. CSRs	.31
	Workforce U.S. IT	.28	Workforce U.S. IT	.28
	Workforce U.S. analysts	.36	Workforce U.S. analysts	.34
	Workforce U.S. consultants	.33	Workforce U.S. consultants	.29
	ETS SWS item tryout	.35	ETS SWS item tryout	.29
	PIAAC (fixed form) for seven countries and languages overall[b]	.20	PIAAC (fixed form) for seven countries and languages overall[c]	.27
	Average correlations across five data collections	.31	Average correlations across six data collections	.31

Note. PHI = Philippines; CSRs = customer service representatives; IT = information technology; SWS = Strategic Workforce Solutions.

[a] With *r* ranging from .15 (Canadian) to .31 (U.S.). [b] With *r* ranging from .06 (Canadian) to .26 (Czech Republic). [c] With *r* ranging from .20 (Italy) to.37 (Ireland).

Table B6

Empirical Support: Correlations Between Facets Used to Form Stability and Optimism Composite

	Optimism	*r*
Stability	Workforce U.S. incumbents overall	.36
	Workforce U.S. applicants overall	.28
	Workforce PHI applicants overall	.26
	Workforce U.S. CSRs	.29
	Workforce U.S. IT	.31
	Workforce U.S. analysts	.25
	Workforce U.S. consultants	.26
	ETS SWS item tryout	.36
	PIAAC (fixed form) for seven countries and languages overall[a]	.41
	Army Tech Report, correlations between adjustment and optimism	.26
	Average correlations across six data collections	.32

Note. PHI = Philippines; CSRs = customer service representatives; IT = information technology; SWS = Strategic Workforce Solutions; Army Tech Report = Drasgow et al. (2012).

[a] With *r* ranging from .35 (Spain) to .52 (Japan).

Appendix C

Use Cases for Score Report Options

The use cases for score report options shown below display the different ways to present scores to clients.

FACETS Dimension Scores

1. **Scores on FACETS dimensions reported for use for developmental purposes at the facet level, based on norm referenced data**

   This represents a possibility for future reporting options outside of the selection context.

2. **Scores on individual FACETS dimensions not reported for selection purposes**

   Given that prediction of selection and work outcomes is more robust at the composite or overall index level than at the facet level, this is the preferred reporting option for selection.

FACETS Overall Selection Index

3. **Continuous score with cut score recommendation**

   Continuous score (not in terms of score categories). Recommend a cut score but allow a company some freedom to slide score up/down based on company's specific needs (e.g., number of job vacancies, selection ratio, performance and employee diversity goals). In this case, it is important to be mindful of case law, as 1991 Civil Rights Act section 2000e-2[1] prohibits different cutoffs for different groups, and Ricci et al. v. DeStefano et al. (2009) prohibits lowering cutoffs to achieve diversity goals.

   Can arithmetically combine the score with scores on other assessments (see below for use cases related to "Integrate FACETS [Overall Index Score] With Scores on Other Assessments"). It is important in this case to consider the criteria on which a company will slide scores up or down; if these choices are not applied appropriately, then there certainly could be issues from a legal standpoint.

   Although there is a body of research on how to set cut scores, companies also have a right to set a cut score based on projections on volume of hiring (e.g., one might say that a minimum score of 5 out of 10 is needed to perform the job well, but if a company is aware there will be 20,000 applicants for 100 positions and is unable to interview all applicants, a decision can be made to only take those who score a 9 or 10 if past history shows that this cut score will give sufficient numbers for interviewing. So a company can have a business reason for a cut score but will also often look to the test vendor to determine what score relates to minimal competence (especially if there is a job where supply is not much greater than demand).

4. **Continuous score without a cut score recommendation (e.g., top down selection)**

   Continuous score (not in terms of score categories). Do not recommend a cut score and assume that either a company will (a) determine its own cut score or (b) choose top scorers on the test to move forward in the selection process.

Scenario A may be problematic because the company is likely not to use the assessment in a standard/fair way across applicants. Scenario B is not likely to be acceptable to companies that screen and hire a large number of applicants and early on in the selection process just want to identify "bottom of the barrel" applicants to screen out (based on not meeting a cut score). Top-down selection can also be used to identify the lowest performers.

Given that Scenario B is common for selection, questions are raised about how cut scores should be set and reported in Case 1. That is, if a company is using FACETS as a tool to "screen out the bottom," it leads to a very different cut score than a top-down tool, such as a case where a huge applicant pool needs to be cut down substantially. Both are possibilities depending on factors such as how well the company is sourcing, the labor market supply and other factors along those lines.

Both of these options assume that the overall selection index score is used by itself, without any other composite scores.

FACETS Composites

5. **Even when the cut on the index score is met, examine composite scores to understand candidates' relative strengths and weaknesses**

   Gather additional information during later stages of selection process (e.g., ask relevant questions during interview, work sample) to better understand candidates' weaknesses as identified by the FACETS composite scores. Consider candidates' strengths and weaknesses when deciding among candidates to hire.

   Company can choose not to look at certain composite scores.

6. **Use index score (with cut) for selection but reject applicant if scores on the composites are not high enough**

   Company decides which composites are important for success in the particular job and decides what the minimum acceptable score level is on those composites. Applicants who make the cut on the index score but not on the important composites are still rejected.

   Another way to think about this use case is that a candidate has to meet the cuts on all elements of the FACETS assessment (on the index and on each composite—with cut scores on the less critical composites set lower than the cut scores on the more critical composites) to pass. This approach may present issues if a company decides on the composites that are important for success without justification or research evidence. Even if choices are made that take into account ETS recommendations based on justified evidence, the use of composites would still vary across jobs in order to meet a company's potentially arbitrary or unjustified standard requirements.

7. **Use composites and not the index score to make decisions**

   Company disregards index score and examines each composite to decide for each composite whether it is relevant for success in the particular job and what minimum score level (e.g., 1, 2, 3, or 4) on that composite the company is willing to accept. A low score may be something that can be remedied via available training. The company can also look to gather additional information from the candidate during an interview or work sample to probe a low score on a particular dimension and check whether the individual would be motivated to improve.

8. **Use index score (with cut) for selection but examine composites for additional information if cut is not met**

If there are not enough qualified candidates in the pool and a particular individual does not meet cut on the index score (at whichever level that cut score is set at the time, based on ETS recommendation and organizational considerations), examine candidate's composite scores.

Company examines the composites with low scores, decides if these composites are relevant for success in the particular job (ideally based on an a priori job analysis), and decides what minimum score level (1, 2, 3, or 4) on these composites it is willing to accept. Considerations:

> The company can also look to gather additional information from the candidate during an interview or work sample to probe a low score on a particular dimension, check whether the low score is accurate, and if accurate, whether the individual would be motivated to improve.
>
> A low score may be something that can be remedied via available training.

Rather than focusing on the composites with low scores, a company may identify an important composite with a very high score and decide that the high score can trump the fact that the index score cut was not met and that some of the other composite scores are low.

This use case may not be viable considering that index scores should be used in a consistent (standardized) way across candidates rather than making exceptions for particular candidates (by ignoring the index score) whenever a hiring manager desires. It may also be difficult to identify someone who is exceptional at a particular composite (such that it trumps other pieces of information on the score report) given that only four score levels are planned for the WorkFORCE Assessment for Job Fit score report.

Integrate FACETS (Index Score) With Scores on Other Assessments

9. **Compensatory fashion**

Combine scores on multiple assessments (e.g., FACETS with cognitive assessment) to come up with an overall assessment battery score. Battery score is used to determine whether a candidate is qualified for the job (or to be interviewed for the job). With this approach, a high score on one assessment in the battery can compensate for a low score on another assessment.

This approach is often used in the industry to account for adverse impact as a result of cognitive test scores. But as Sackett and Wilk (1994) and Sackett and Ellingson (1997) showed, this approach usually requires a need to weight noncognitive scores heavily over cognitive ability scores.

10. **Multiple hurdle fashion (sequential)**

Multiple assessments are administered, moving from early screens (e.g., personality test like FACETS) to later screens (e.g., structured interview) for the company. Applicants are required to achieve a passing score on each assessment in the process to be allowed to move on to the next assessment.

In general, this is the most common type of process flow for job selection processes in the industry.

11. **Multiple hurdle fashion (simultaneous)**

    Multiple assessments (e.g., personality test like FACETS with cognitive assessment) are administered at the same time. Applicants are required to achieve a passing score on each of these assessments to move on to the next stage of the selection process (e.g., interviews).

12. **Blend of compensatory and multiple hurdle fashion**

    Candidates are required to achieve a passing score on certain assessments in a battery (e.g., on FACETS) and also achieve a minimum score on the overall assessment battery. This approach ensures minimal skill levels in critical areas while allowing a candidate to compensate for a low score in a less critical area with a high score in another area.