Technical Report 1371

Enhancing the Validity of Rating-Based Tests

Peter J. Legree
Alisha M. Ness
Robert N. Kilcullen
U.S. Army Research Institute

Amanda J. Koch
Human Resources Research Organization

December 2018

United States Army Research Institute
for the Behavioral and Social Sciences

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

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

MICHELLE L. ZBYLUT, Ph.D.
Director

Technical review by

Mark C. Young, U.S. Army Research Institute
Kristophor G. Canali, U.S. Army Research Institute
Elissa M. Hack, U.S. Army Research Institute

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
December 2018	Final	09/01/2015 – 11/30/2018

4. TITLE AND SUBTITLE
Enhancing the Validity of Rating-Based Tests

5a. CONTRACT/GRANT NUMBER

5b. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S)
Peter J. Legree, Alisha M. Ness, Robert N. Kilcullen;
Amanda J. Koch

5c. PROJECT NUMBER
A790

5d. TASK NUMBER

5e. WORK UNIT NUMBER
311

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
6000 6th Street (Bldg. 1464 / Mail Stop: 5610)
Fort Belvoir, Virginia 22060-5610

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
6000 6th Street (Bldg. 1464 / Mail Stop: 5610)
Fort Belvoir, Virginia 22060-5610

10. SPONSOR/MONITOR'S ACRONYM(S)
ARI

11. SPONSORING/MONITORING
Technical Report 1371

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
ARI Research POC: Dr. Peter J. Legree, Selection and Assignment Research Unit

14. ABSTRACT
Profile similarity metrics (PSMs) can be computed for rating-based judgment tests, personality scales, and biodata inventories to supplement conventional measures and enhance scale validity. These metrics quantify: shape, the correlation between a respondent's rating profile and the scoring key; scatter, respondent tendency to use more or less of the available rating scale; elevation, respondent tendency to systematically provide high or low ratings; and delta, respondent tendency to provide high or low ratings relative to the key. Analyses conducted for three projects confirmed theoretical expectations that PSMs can be used to accurately model distance score variance and increment the validity of distance scores against performance outcomes. Project 1 utilized three judgment tests and demonstrated that shape and delta metrics predicted supervisor performance ratings (R = .33), while elevation and shape metrics predicted career intent (R = .25). Project 2 utilized conventional personality scales and showed that PSMs provided incremental validity beyond distance scores against performance outcomes and documented the stability of the validity gains using an independent cross sample. Project 3 evaluated the use of PSMs to score experimental 9-point personality in addition to conventional 5-point personality scales. Project 3 analyses demonstrated that PSMs provided incremental validity against performance outcomes beyond distance scoring for the combined personality battery (R = .54 vs. R = .47). The third project also documented construct validity between overlapping constructs for the 5-point and 9-point scales. These results redefine validity expectations for personality/judgment constructs and demonstrate the efficacy of PSMs procedures to broaden the scope of psychological domains for which accurate measurement is possible. The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) supported this research project.

15. SUBJECT TERMS
Profile Similarity Metrics, Personality, Situational Judgment

SECURITY CLASSIFICATION OF:			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE	Unlimited	61	Tonia S. Heffner 703-545-4408
Unclassified	Unclassified	Unclassified	Unclassified		

Standard Form 298 (Rev. 8-98)

Technical Report 1371

Enhancing the Validity of Rating-Based Tests

Peter J. Legree
Alisha M. Ness
Robert N. Kilcullen
U.S. Army Research Institute

Amanda J. Koch
Human Resources Research Organization

Selection and Assignment Research Unit
Tonia S. Heffner, Chief

December 2018

EXECUTIVE SUMMARY

Research Requirement:

Profile similarity metrics (PSMs) can be computed for rating-based scales to quantify: shape, the correlation between a respondent's rating profile and the scoring key; scatter, respondent tendency to use more or less of the available rating scale; elevation, respondent tendency to systematically provide high or low ratings; and delta, respondent tendency to provide high or low ratings relative to the key. Based on formulaic analyses, research hypotheses proposed that PSMs can be used to model distance score variance and to provide incremental validity beyond distance scores against performance outcomes.

Approach:

Analyses for three projects evaluated hypotheses that PSMs can be used to model distance score variance and increment the validity of distance scores against performance outcomes. The first project used data collected for three rating-based judgment tests that had been validated against supervisor performance ratings and self-report career intent. The second project used data collected for conventional personality scales that had incorporated conventional 5-point rating scales and been validated against U.S. Army Cadet Command (USACC) order of merit scores (OMS) for two separate cohorts. The third project used personality data collected for conventional personality scales incorporating 5-point rating scales and experimental personality scales incorporating 9-point rating scales that had been validated against USACC OMS.

Findings:

Highly consistent support was documented for the PSM research hypotheses across the analyses conducted for the three projects (i.e., PSMs can be used to model distance score variance and increment the validity of distance scores against performance outcomes). In addition, many of the estimated gains in scale validity were substantial.

The first project demonstrated that highly-efficient judgment tests could be developed and scored using PSMs that had modest levels of validity against supervisor performance ratings (R = .33) and career intent (R = .25). These scales required less than 10 minutes to administer, and the computed validities compare favorably to meta-analysis estimates of judgment test validity ($\overline{r}$ = .26 and ρ = .34; McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001).

The second project used a cross validation design and demonstrated that PSM scoring provided validity gains for conventional personality scales that were highly stable in a fully independent cross-sample using data that had been collected two-years later.

The third project demonstrated that PSM scoring algorithms provided incremental validity against performance outcomes beyond distance scoring for a battery of conventional

5-point and experimental 9-point personality scales (R = .54 vs. R = .47). In addition, PSM scoring provided support for expectations based on job analysis results and psychological models proposing that high potential cadets would excel at communication tasks and demonstrate higher levels of safety awareness.

Utilization and Dissemination of Findings:

The analyses demonstrated the potential validity gains from using PSMs to score rating-based judgment tests and personality inventories. The judgment test results show that valid and highly-efficient judgment tests can be developed at minimal cost by incorporating rating scales with a relatively large number of response options and by using PSMs to score the judgment tests. In contrast, the development of judgment tests using conventional measures has been expensive, and the resultant scales require substantial administration time for data collection and have often been associated with low validity estimates.

The personality battery validity (R = .54) exceeded validity estimates for most personality inventories as well as meta-analytic validity estimates for general cognitive ability. Therefore, this result suggests that the application of PSM scoring techniques to personality scales may provide a credible basis to challenge the dominance of general cognitive ability for U.S. Army personnel selection applications.

CONTENTS

Page

APPENDICES

LIST OF TABLES

Page

ENHANCING THE VALIDITY OF RATING-BASED TESTS

Introduction

In this paper, we describe and evaluate the use of profile similarity metrics (PSMs) to improve the psychometric properties of rating-based situational judgment tests (SJTs) and personality inventories that have used distance-based algorithms to compute conventional scale scores. While distance scores have often been explicitly computed for rating-based SJTs, conventional scores for most personality and biodata scales are formulaically redundant with distance metrics, $r = -1$. (See Appendix A.) Therefore, our results may carry implications for improving the validity of a wide range of scales that are frequently used for personnel selection.

From a formulaic perspective, PSMs provide multiple indices that assess the similarity of a respondent's pattern of ratings to a scoring standard. These metrics quantify: shape, the correlation between a respondent's rating profile and the scoring key; scatter, the tendency of a respondent to use more or less of the available rating scale; elevation, the tendency to systematically provide high or low ratings; and delta, the tendency to systematically provide high or low ratings relative to the scoring key. From a psychometric perspective and as detailed below, PSMs can be combined through regression procedures to understand the variance of distance scores that are computed for rating-based scales, and enhance scale validity against conceptually related criteria. In addition, shape scores can be computed using alternate keys to evaluate competing keying approaches that may enhance scale validity.

Interest in the PSM framework was initially associated with the development of scoring standards and algorithms for rating-based judgment tests that were created for emerging knowledge domains such as social intelligence, tacit driving knowledge, emotional intelligence, and leadership (Legree, 1995; Legree, Heffner, Psotka, Martin & Medsker, 2003; Legree, Kilcullen, Psotka, Putka & Ginter, 2010; Legree, Psotka et al., 2014). However, PSMs can be computed for many personality and biodata inventories, and we began to speculate that PSMs might be optimally weighted to enhance the validity of these scales against relevant outcomes.

Distance-Based Metrics

Many rating-based judgment tests and personality inventories have used distance-based or distance-related metrics to compare respondent item ratings to scoring key values and calculate scale scores. These metrics include:

1. The mean absolute difference between a participant's item ratings and the keyed values, $D = \sum|X_i - K_i|/n$ for item $i = 1$ to n (e.g., Barrick & Mount, 1991; Costa & McCrae, 1991; Cullen, Sackett & Lievens, 2006; Edwards, 1993; McHenry, Hough, Toquam, Hanson & Ashworth, 1990; Muros, 2008; Sternberg et al., 2006; Wagner & Sternberg, 1985);

2. The mean square item difference between the participant's ratings and the keyed values, $D^2 = \sum(X_i - K_i)^2/n$ for item $i = 1$ to n (e.g., Edwards, 1993; Motowidlo, Crook, Kell & Naemi, 2009; Sternberg & Wagner, 1993);

3. The square root of the mean square item difference, $D^1 = \sqrt{D^2}$ (Edwards, 1993); and

4. Endorsement ratios based on a proportion scoring algorithm (e.g., Mayer, Caruso & Salovey, 1999; Mayer, Salovey, Caruso & Sitarenios, 2003).

Although these measures have rarely been simultaneously evaluated, extant analyses suggest very high correlations among these metrics. For example, Edwards (1993) reports highly similar validity estimates for D, D^2, and D^1 distance metrics against an array of outcome criteria. High correlations have also been reported between endorsement ratios and distance scores for the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) subtests (r = -.89; Legree, Psotka et al., 2014). These observations suggest that these measures represent a class of "distance-based" metrics that provide highly redundant information. Evaluating this proposition is central to understanding the potential of PSMs to increment distance measures because PSMs can be formulaically derived from the D^2 equation as will be described in this report – as opposed to PSMs simply being supplementary variables that reflect intuitive expectations.

Distance-based algorithms quantify the overall "match" between a scoring key and a respondent's rating profile. However, the term "match" highlights an important ambiguity of distance scores in the context of rating-based scales. A superior distance score may reflect: similar shape between an individual rating profile and the elements in the scoring key, similar levels of elevation between the items in a rating profile and the scoring key, the variance of the item values in an individual's rating profile, or some combination of these effects.

Although distance measures carry intuitive appeal, they can be highly influenced by respondent tendencies to elevate their ratings relative to the key (delta in elevation effects), or to use more or less of the available rating scale (scatter effects). Figure 1 depicts profiles of ratings for three individuals across five items on a 9-point scale, as well as the scoring key used to assess the quality of these responses. The figure shows that distance scores may conflict with correlation-based (shape) metrics when they are used to rank-order individual test performance and illustrates that elevation and scatter effects may dramatically impact distance scores for rating-based scales. To interpret these values, superior performance is indicated by distance values approaching 0.0, but by correlation (shape) values near 1.0. According to the distance metrics, the respondents' test performance would be ranked:

Respondent A > Respondent C > Respondent B.

Yet according to the correlation metric, the respondents' test performance would be ranked:

Respondent A = Respondent B > Respondent C.

How can Respondent B's test performance change so markedly depending on the scoring metric being used? Although the shape of Respondent B's profile is very similar to that of the keyed profile (as evidenced by the high correlations), Respondent B's profile is elevated relative to the keyed profile (i.e., vertically inflated). Similar effects can be illustrated for individual rating profiles that contain too much or too little scatter relative to the keyed profile (i.e., within-person rating variance). Such simple differences in the elevation and scatter of respondent rating profiles can negatively impact distance-based metrics but will have little impact on

correlation-based metrics. Reflection upon these issues suggests that maximizing scale validity may require optimally weighting these various metrics.

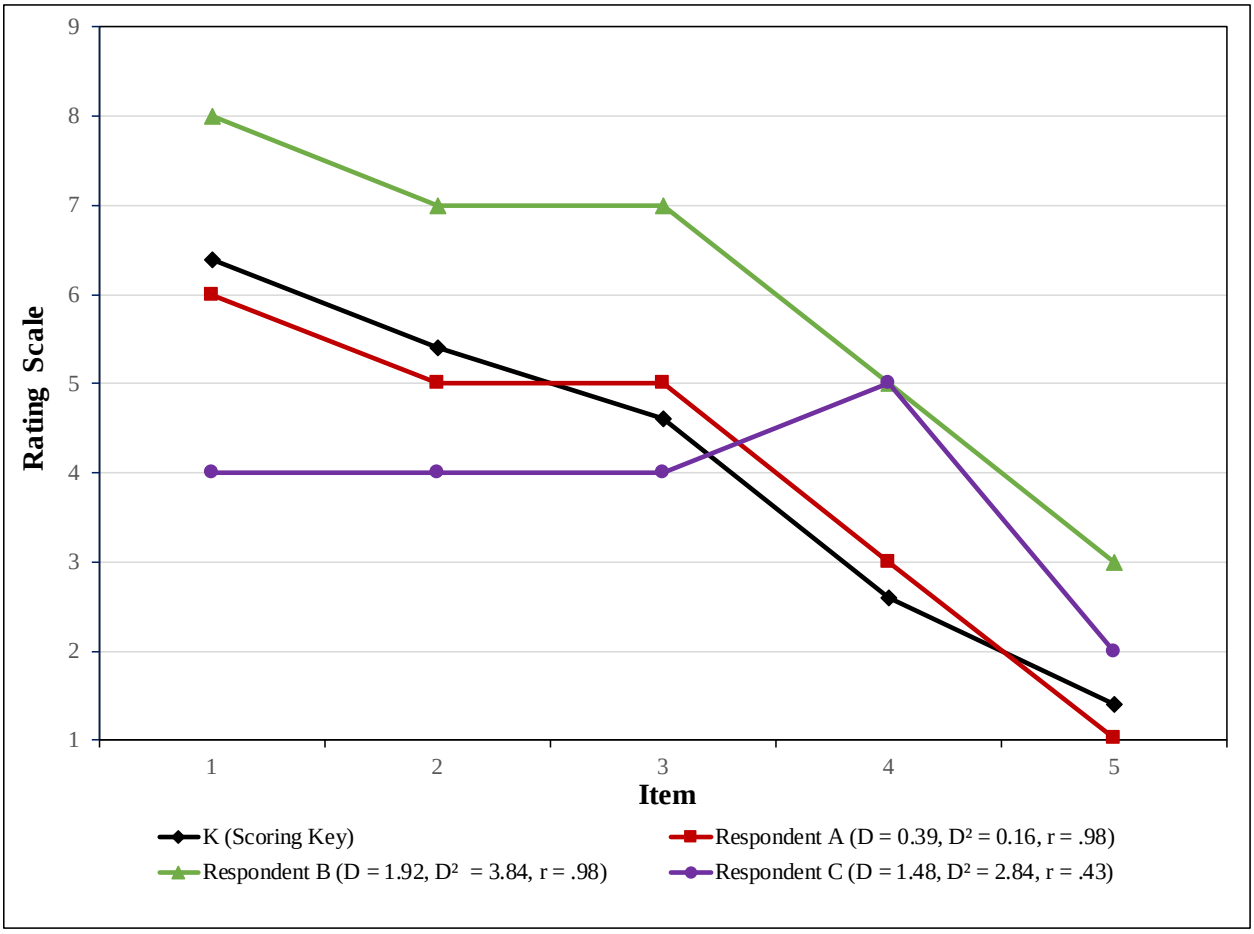

Figure 1. Scoring key and rating profiles for three respondents: Superior profiles indicated by lower distance (D), and higher correlations with the scoring key (*r*)

Profile Similarity Metrics as a Mathematical Framework

PSMs provide a useful and comprehensive scoring framework for rating-based judgment tests because they quantify these separate effects (Cronbach & Glaser, 1953; Legree, Psotka et al., 2014). To make this point clearly, we decompose the D^2 metric, which computes distance as the mean squared difference between items in a respondent rating profile and the scoring key. Understanding the D^2 formula is critical to recognizing the inherent limitations of using distance metrics for rating-based tests because the D^2 formula can be algebraically decomposed into separate metrics that quantify the shape, elevation, and scatter of respondent rating profiles. These metrics can also be used to optimize scale validity through regression procedures.

We define the D^2 metric as the mean squared difference between items in a respondent rating profile and the scoring key. Accordingly, each respondent's set of ratings for a judgment test is conceptualized as a rating profile vector, **X**, with *n* elements (i.e., item ratings). Likewise, the scoring key is represented as a scoring profile vector, **K,** also with *n* elements (i.e., X_i and K_i correspond to ratings for item *i* obtained from an individual and from the scoring key). The D^2 metric is then calculated as the mean squared difference between elements in the two arrays:

$$D^2 = \sum_{i=1}^{n} (X_i - K_i)^2 / n \quad (1)$$

Using algebraic substitutions that are detailed in Appendix B, Equation 1 can be represented using conventional statistical terms:

$$D^2 = \Delta_{Key}{}^2 + \frac{(n-1)(sd_x{}^2 + sd_k{}^2 - 2sd_x sd_k r_{xk})}{n} \quad (2)$$

In Equation 2, $\Delta_{Key}{}^2 = (X_{mean} - K_{mean})^2$ quantifies the squared difference between the respondent's mean rating and the mean keyed value. The remaining terms carry standard statistical meaning: $sd_x{}^2$ equals the variance of the elements in a respondent's rating profile, $sd_k{}^2$ equals the variance of the keyed values, and $r_{x,k}$ equals the correlation between a respondent's rating vector and the keyed values. According to Equation 2, superior D^2 scores (i.e., values approaching 0.0) will be associated with: delta terms, $\Delta_{Key}{}^2$, approaching 0.0; and shape values, $r_{x,k}$, approaching 1.0. However, the optimal value for the scatter term, $sd_x{}^2$, will vary across individuals, is dependent on the magnitude of the individual's shape term, and can be computed for an individual as: $sd_x = r_{x,k} sd_k$. (See Appendix B.) This result shows that respondent profiles with poor shape may have superior distances scores when rating scatter is minimized.

Equation 2 is important because it implies that the variance of distance-based metrics may be modelled as main effects using regression procedures and the following PSMs:

PSM 1. Shape, $r_{x,k}$, the correlation between a respondent's rating vector and the keyed vector (i.e., scoring key);

PSM 2. Scatter, $sd_x{}^2$, the respondent's rating variance;

PSM 3. Delta, $\Delta_{Key}{}^2 = (X_{mean} - K_{mean})^2$, the squared difference between the respondent's mean rating and the mean keyed value.

We use the terms "shape," "scatter" and "delta" to designate PSMs as individual difference metrics and minimize confusion with the analogous statistical terms when presenting results. Coupled with observations that distance-based metrics tend to be nearly redundant (Edwards, 1993; Legree, Psotka et al., 2014), we propose the following PSM hypotheses:

Hyp 1. Distance-based measures (e.g., *D* and D^2 metrics) computed using the same data will be nearly redundant, $r > .90$.

Hyp 2. Shape, delta, and scatter metrics will account for nearly all the variance in distance scores as main effects: $R_{Distance,Shape.Scatter.Delta} > .90$.

The first two hypotheses are largely formulaic, but their endorsement is critical to the proposition that all distance metrics may be accurately modelled using PSMs that are derived from the D^2 formula. Therefore, we incorporated a very high threshold into the two hypotheses, $R > .90$. Furthermore, endorsement of the first two hypotheses provides a strong foundation for exploring the use of PSMs to improve the validity of rating-based tests.

More importantly, there is no compelling reason to expect that distance metrics will reflect optimal weighting of shape, scatter, and delta for the purpose of optimizing scale validity. In fact, researchers have demonstrated that rating-based judgment test shape scores may have greater predictive and construct validity than corresponding distance scores, (Legree, 1995; Legree, Psotka et al., 2014; McDaniel, Psotka, Legree, Yost & Weekley, 2011; Weng, Yang, Lievens & McDaniel, 2018). However, research has not addressed the impact of simultaneously weighting the shape, scatter, and delta metrics to enhance scale validity. This reasoning implies a third hypothesis:

Hyp 3. Shape, delta, and scatter metrics will add incremental validity to distance scores against performance outcomes when distance scores reflect suboptimal weighting of these PSMs.

Hypothesis 3 addresses the possibility that distance-based scoring algorithms have systematically underestimated the validity of rating-based scales. For clarification, we expect that PSMs will increment scale validity unless regression weights computed by regressing distance scores onto PSMs mirror those weights obtained by regressing the criterion onto those same PSMs. Support for Hypothesis 3 carries practical implications for enhancing the utility of rating-based scales and theoretical implications for validating psychological models (e.g., optimally weighted PSMs for a conceptually relevant scale may validate construct expectations despite distance scores being uncorrelated with relevant outcomes).

Using PSMs to Refine Scale Keys

The above reasoning implicitly assumes that a high-quality key has been developed and is being used to compute distance scores. Equation 2 supports two intuitive expectations regarding the characteristics of high-quality keys, as well as one counter-intuitive implication. The first expectation corresponds to the beliefs that the key should have proper shape to maximize the shape term, $r_{x,k}$, for individuals who are high on the underlying construct (i.e., the the keyed values should allow the shape term to approach 1.0 for individuals who are high on the underlying construct). The second expectation corresponds to the belief that the key should be properly centered so the delta term, Δ_{Key}^2, will approach 0.0 for respondents who are high on the underlying construct.

With respect to the counter-intuitive result, the term representing the variance of the keyed values, sd_k^2, does not directly enter into the computation of the shape, delta, or scatter metrics. Therefore, it follows that a high-quality key is defined by its shape and centering, with its scatter, sd_k^2, being irrelevant.

These observations are important because key construction is often based on expert opinion or relatively simple models that may contain error. However, PSMs provide insight into these issues and may be used to optimally center the key and adjust key shape.

Adjusting key elevation. Equation 2 shows that the delta term, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$, is relevant to understanding the validity of distance scores computed using a conventional key. However, the potential validity of the delta term will be minimized if the scoring key is poorly centered (i.e., the delta term would be computed using the value, K_{mean}, as opposed to being computed using a value that optimizes the validity of the delta term, K_{opt}). To show that PSMs can be used to recenter the key, we define delta-optimal as:

$$\Delta_{opt}^2 = (X_{mean} - K_{opt})^2, \quad (3)$$

where $K_{opt} = K_{mean} + A$.

Using algebraic substitutions that are detailed in Appendix C, Equation 3 converts to:

$$\Delta_{opt}^2 = \Delta_{Key}^2 + A^2 + 2AK_{mean} - 2AX_{mean}, \quad (4)$$

where $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$, and A is a constant.

Equation 4 shows that the delta-optimal term represents the linear combination of the respondent delta and elevation terms (i.e., Δ_{Key}^2 and X_{mean}), and two constants (i.e., K_{mean} and A). Therefore, the variance of the delta-optimal term, Δ_{opt}^2, represents a perfect linear combination of the delta and elevation terms (i.e., $R_{\Delta Opt2,\Delta Key2.Xmean} = 1$). This result implies that in regression models, the elevation term will provide incremental validity beyond the delta term when the key is poorly centered. This reasoning focuses attention on an additional PSM and identifies a fourth hypothesis:

PSM 4. Elevation, X_{mean}, the respondent's mean item rating.

Hyp 4. Elevation will add incremental validity to the shape, delta, and scatter terms for the prediction of performance outcomes when the key is poorly centered.

Adjusting key shape. While key shape cannot be directly adjusted using PSMs, many scales are constructed with implicit expectations that respondent shape scores will correlate with relevant outcomes. Therefore, minimal correlations between shape scores and outcomes may indicate limitations regarding the shape of the key as opposed to the relevance of the construct.

This concern may be most relevant to improving the validity of personality scales because their keys are often constructed to reflect extreme values (e.g., "1" for reversed and "5" non-reversed items on a 5-point rating scale – see Appendix A). However, possessing high levels of positive traits may have negative effects beyond a specific threshold (Grant & Shwartz, 2011; MacCann, Ziegler & Roberts, 2012). For example, very high levels of self-control are also associated with obsessive-compulsive disorder (Tangney, Baumeister & Boone, 2004). In addition, this scoring approach does not acknowledge that an individual who is extremely high on one dimension may be lacking on other dimensions. Therefore, a scoring key that reflects

extreme responses for all items may not be the most beneficial in predicting desired outcomes. These observations suggest that improving the shape of conventional keys for personality scales may improve their utility.

While PSMs cannot be used to directly optimize key shape, consensus keys have been constructed for rating-based judgment tests based on expectations that rating errors will be distributed around the mean respondent rating for each response option (Legree, 1995; Mayer et al., 2003; McDaniel et al., 2011; Weng et al., 2018). Analyses have also demonstrated high levels of convergence between consensus and expert-based keys for judgment tests (Legree, Psotka, Tremble & Bourne, 2005). This convergence between consensus and expert keys for rating-based judgment tests suggests that consensus keying may represent a viable alternative for personality scales that have been keyed using conventional methods. This reasoning identifies a fifth PSM and a fifth hypothesis that are relevant to keying personality scales:

PSM 5. Shape-consensus, $r_{x,consensus}$, the correlation between a respondent's rating vector and the consensus key with each keyed element computed as the mean respondent item rating.

Hyp 5. Shape-consensus will add incremental validity to the shape, delta, scatter, and elevation terms for the prediction of performance outcomes when the conventional key has poor shape.

PSM Implications for Scale Design. The above formulas indicate that rating-based scale validity may be highly dependent on the psychometrics of the shape, delta, elevation, and scatter measures that can be computed for individual rating profiles. Following this rationale, we provide several suggestions for the design of rating-based scales in order to improve the psychometrics of the underlying PSMs.

First, we recommend attaching many options (e.g., response actions) to each scenario to improve the psychometrics of the shape metric for rating-based judgment tests. This approach can allow much more data to be collected per scenario, while simultaneously reducing overall test administration requirements. For example, a 5-scenario judgment test with 10 options per scenario will yield 50 data points, whereas a 10-scenario judgment test with 4 options per scenario will yield 40 data points. However, the 5-scenario test will decrease overall reading requirements because scenario descriptions are often lengthy. Therefore, we prefer the 5-scenario format from both information and test administration perspectives.

Second, we suggest providing large versus small rating scales (e.g., 11-point vs. 5-point scales) because small rating scales constrain the capacity of highly discerning respondents to register subtle differences in their opinions (e.g., Stevens, 1975). Therefore, the coarseness of small scales may limit the psychometrics of the shape, delta, elevation, and scatter metrics for some individuals. In addition, the impact of this constraint on the shape metric will be magnified for respondents who elevate or depress their ratings because the effective size of a small rating scale will be further reduced (e.g., only 3-points on a 5-point rating scale may be used by individuals who systematically elevate their ratings).

Finally, a mix of reversed and non-reversed items is required to compute shape scores using conventional keys (i.e., the key must contain variance so that individuals who are high on the underlying construct will select ratings both ends of the rating scale). This concern is relevant to our analyses because some of the data we analyzed had been collected for personality scales that did not contain any reversed items. Therefore, a balance of reversed items may increase the potential of PSMs to improve the scale validity of personality scales.

Current Project

We conducted three sets of analyses to evaluate the five PSM hypotheses and assess our expectations for scale design. For our first project, we leveraged the PSM framework to create three rating-based judgment tests with minimal test administration time requirements. These scales used abbreviated scenarios, paired many items with each scenario, and incorporated large rating scales to allow respondents to register subtle differences in opinion in order to maximize respondent variance on the shape, delta, scatter, and elevation metrics.

Our second project was designed to evaluate the utility of using PSMs to improve the validity of an established personality inventory against performance outcomes. These personality scales adopted a biodata approach, used a 5-point rating format, and had been systematically refined over a fifteen-year period. We used data that had been collected in 2013 to validate PSM-based scale scores and cross-validated the results using data that had been collected from an independent sample in 2015.

Our third project evaluated the utility of experimental personality scales that were created in accordance with the PSM framework to increment the predictive validity of established personality scales. Each experimental item consisted of two opposing statements. Respondents were asked to rate the extent to which the two items describe their behaviors and experiences using a large, 9-point rating scale to allow respondents to register subtle differences in their self-assessments. These items were distributed over nine scales, and each scale contained a near even mix of reversed and non-reversed items. Unlike the established personality inventory, the experimental scales had not been extensively evaluated or refined. We used regression procedures to evaluate the PSM hypotheses, estimate the validity of the experimental (9-point) personality scales, and assess their potential to increment the validity of the established (5-point) personality scales.

Project 1: PSMs for Rating-Based Judgment Tests

Analyses for the first project were designed to evaluate the PSM hypotheses using data collected for three rating-based judgment tests that were developed to predict job performance and career continuance criteria. While the first four hypotheses could be assessed with these data, the fifth hypothesis was not relevant because the judgment tests had been consensually keyed. Therefore, all references to shape scores imply "shape-consensus" scores in this section with the keyed values defined as the mean respondent rating for each option. The keying data were collected using a separate sample of officers.

Method

Participants. The validation sample consisted of 644 U.S. Army officers who volunteered to participate in the project. This sample included 215 Captains assigned as Company Commanders and 429 Lieutenants assigned as Platoon Leaders.

Design and Procedure. We incorporated large rating scales into the judgment tests to enable respondents to register subtle differences in their beliefs and understandings in accordance with the PSM framework. We expected that this response format would enhance the validity of the shape, delta, scatter, and elevation metrics against job performance and career continuance outcomes while using test administration time efficiently.

The scales were embedded into a larger project to assess their validity (Russell, Paullin, Legree, Kilcullen & Young, 2017). The judgment test predictor data had been collected from Captains who were Company Commanders and Lieutenants who were Platoon Leaders. The Captains and Lieutenants also provided career intent data. The performance rating data were collected from the direct supervisors of the officers who completed the judgment tests.

To evaluate the PSM hypotheses, we report analyses that are based on the full sample (Captains and Lieutenants) because its larger sample size provides greater stability. We also used the data to evaluate the overall efficacy of using the rating-based judgment tests to predict the performance ratings as well as the career intent criteria.

Measures. Each of the three rating-based judgment tests referenced very brief scenarios, and each test listed between 17 and 30 options per scenario for respondents to rate using a large rating scale (i.e., 9 or 10 point scales). Each option consisted of a short phrase, and each judgment test required five to ten minutes to administer. Table 1 contains example items for the judgment tests. The career intentions and supervisor performance rating data were used as outcomes for this project.

Table 1
Judgment Test and Personality Scale Example Items

LKT Characteristics Scale (Project 1)

Scenario:	1	2	3	4	5	6	7	8	9	10
How important are these traits to leadership?	Not-at-all Important								Extremely Important	
1. Patriotism	○	○	○	○	○	○	○	○	○	○
2. Curiosity	○	○	○	○	○	○	○	○	○	○

Rating-Based Consequences Test (Project 1)

Scenario:	1	2	3	4	5	6	7	8	9
What would be the results if people no longer need or want sleep?	Not Very Original			Somewhat Original					Highly Original
1. Get more work done	○	○	○	○	○	○	○	○	○
2. Alarm clock not necessary	○	○	○	○	○	○	○	○	○

CBEF (Projects 2 & 3)

Achievement Example Items

1. To what extent have you been willing to take on a difficult task if you could learn a lot from doing it?
   ○ 1 (*never*); ○ 2 (*seldom*); ○ 3 (*occasionally*); ○ 4 (*frequently*); ○ 5 (*often*).
2. To what extent has your main source of satisfaction come from school or work?
   ○ 1 (*never*); ○ 2 (*seldom*); ○ 3 (*occasionally*); ○ 4 (*frequently*); ○ 5 (*often*).

CPM (Project 3)

1.	I give my best effort when it's needed at work or school.	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	I enjoy giving my best effort at work or school regardless of the task.
2.	It's important to know when to cut your losses.	○ ○ ○ ○ ○ ○ ○ ○ ○ ○	It is important to finish something you've started.

Leader Knowledge Test Characteristics and Skills Scales. The Leader Knowledge Test (LKT) contained two scales that were designed to assess knowledge of characteristics and skills that are relevant to leader performance (Yukl, 2002). These scales reflected expectations that leaders gain knowledge regarding the importance of leader-relevant characteristics and skills through experience and reflection upon their experiences (Polanyi, 1966; Stenberg & Hedlund, 2002; Wagner & Sternberg, 1985).

We constructed the LKT scales by identifying 15 characteristics and 15 skills that are associated with effective leadership (Yukl, 2002). In addition, we identified 15 characteristics and 15 skills that are socially positive, but have not been linked to effective leadership. The 30 characteristics were assembled into the LKT Characteristics scale, and the 30 skills were assembled into the LKT Skills scale. To complete the LKT, respondents read 133 words to respond to the 30 items on the LKT Characteristics scale, and 129 words for the 30 items on the LKT Skills scale; the word count includes instructions.

Consequences Ratings Test. The Consequences Test was originally developed to measure creativity using a constructed response format (Christensen, Merrifield & Guilford, 1953). Although analyses have demonstrated modest validity for the Consequences Test against leadership outcomes (Zaccaro et al., 2015), the test is not practical for operational use because a panel of human experts is required to score each protocol.

Therefore, we developed a rating-based version of the test that described 5 scenarios and presented 17 options per scenario. Respondents rated the creativity of each alternative and the scoring algorithm quantified the quality of the creativity ratings. This test reflects expectations that the abilities to provide creative responses and assess the creativity of responses are correlated. This scale required respondents to read 421 words to complete 85 items – including instructions.

Supervisor Performance Ratings. Confidential supervisor ratings were collected for each participant using a 13-item questionnaire. The supervisor data had been previously analyzed to develop performance scales corresponding to: Management, Administration & Communication; Leadership & Personal Discipline; and Technical Task Competence (Russell et al., 2017). We computed an overall performance score by averaging these outcomes, which were highly correlated (all $r > .88$, all $p < .001$).

Career Intent. Respondents completed 3 items regarding their long-term career goals. These data were used to develop a career continuance outcome, Career Intent. Career Intent reflects intentions to remain in the Army.

Data Analysis. The following six metrics were computed for each judgment test: mean item distance, $D = \sum|X_i - K_i|/n$; mean item distance squared, $D^2 = \sum(X_i - K_i)^2/n$; shape = $r_{x,consensus}$; delta, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$; scatter = sd_x^2; and elevation = $\sum X_i/n$. The judgment test scoring keys were consensually derived as the mean rating for each item. Superior PSM scores are indicated by values for the shape metrics that approach 1.0, but by values approaching 0.0 for the distance and delta metrics. We also expected that superior scatter scores would be indicated by higher values because analyses have documented modest positive correlations between shape and scatter scores (e.g., Legree, Psotka et al., 2014).

Results

Descriptive Statistics. Table 2 reports reliability and validity estimates against supervisor performance ratings and self-reported career continuance expectations for each metric by scale. Reliabilities were acceptable for all the measures, although somewhat lower for the shape metrics. The bivariate correlations indicate that the most potent predictors of performance corresponded to the shape metric for each scale despite their lower reliabilities. This result underscores the importance of designing each judgment test to maximize the reliability of the shape metric for each scale. In addition, the delta and scatter metrics correlated with the supervisor performance ratings.

In contrast to the correlations with the performance outcome, the strongest predictors of Career Intent corresponded to the elevation metrics for the LKT Characteristics and Skills scales. Although not detailed in the Table 2, somewhat higher validities for the PSM and distance metrics were obtained for the Captain subsample.

PSM Hypotheses.

Foundational Hypotheses. The first hypothesis proposed that D and D^2 metrics would be highly correlated for each of the three rating-based judgment tests. As expected, the D and D^2 metrics approached redundancy for each scale (all $r > .97$, all $p < .001$). See Table 3.

The second hypothesis proposed that the distance scores for each of the judgment tests can be modelled as a composite of the shape, delta, and scatter effects. Therefore, we regressed the distance scores for each scale onto the corresponding shape, delta, and scatter metrics for each judgment test. The regression analyses confirmed Hypothesis 2 for each of the three scales (all $R > .93$, all $p < .001$). Table 3 provides regression results. Confirmation of the first two hypotheses indicates the use of either distance metric (i.e., D or D^2) is arbitrary and that PSMs may be used to model distance score variance for rating-based judgment tests.

Predictive Validity Hypotheses. The third hypothesis proposed that optimally weighting PSMs may provide incremental validity beyond distance scores for each criterion and judgment test. We assessed Hypothesis 3 by regressing each criterion on the distance scores (step 1), followed by the shape, delta, and scatter metrics (step 2), and the results are reported in Table 4 (Step 2).

The hierarchical regression models supported Hypothesis 3 for each judgment test against the performance outcome. In addition, the hierarchical regression procedure supported Hypothesis 3 for the LKT Skills scale against the career continuance outcome. These gains roughly doubled the predictive validity of each individual test against the performance criterion and provided a very large gain for the LKT Skills scale against the continuance outcome.

Table 2

Judgment Test Reliabilities[a], Validities, and Correlations

Metric	R_{xx}	Performance		Career Intention		Inter-correlation Matrix				
		r	Sig	*r*	Sig	D^2	Shape	Delta	Scatter	Elevation
LKT Characteristics (*n* = 634)										
Distance	.80	-.10	.013	.03	.407	.97	-.59	.68	.21	.16
Distance-squared	.78	-.11	.005	.03	.430		-.66	.71	.17	.11
Shape	.75	.18	.001	.02	.701			-.40	.36	.00
Delta	.82	-.09	.028	.02	.542				-.27	.11
Scatter	.80	.11	.010	.03	.512					-.24
Elevation	.88	-.03	.377	.15	.001					
LKT Skills (*n* = 631)										
Distance	.87	-.08	.042	.02	.593	.97	-.38	.71	.32	.03
Distance-squared	.85	-.07	.072	-.01	.762		-.41	.64	.39	-.10
Shape	.65	.17	.001	-.04	.370			-.26	.26	-.03
Delta	.90	-.11	.007	.12	.002				-.32	.40
Scatter	.85	.10	.011	-.10	.010					-.38
Elevation	.93	-.05	.196	.20	.001					
Consequences (*n* = 644)										
Distance	.96	-.10	.013	.09	.018	.98	-.65	.77	.32	-.33
Distance-squared	.97	-.12	.003	.09	.021		-.72	.79	.27	-.43
Shape	.91	.15	.001	-.10	.013			-.54	.14	-.50
Delta	.87	-.15	.001	.05	.187				-.27	.32
Scatter	.88	.07	.063	.02	.573					.04
Elevation	.92	-.08	.037	.01	.861					

[a]Coefficient alpha computed for the distance, distance-squared, shape, scatter and elevation metrics. Split-half estimate computed for the delta metric.

Table 3

Distance Scores for the Judgment Tests Regressed on the Corresponding PSMs

Judgment Test (n)	H1[a]	H2[b]			Shape[c]		Delta[c]		Scatter[c]	
	r_{D,D^2}	$R_{D,PSMs}$	*F* change	*Sig*	*β*	*r*	*β*	*r*	*β*	*r*
LKT Characteristic (635)	.97	.93	1429.04	.001	-.552	-.59	.617	.68	.575	.21
LKT Skills (632)	.97	.97	3921.96	.001	-.350	-.38	.817	.70	.692	.35
Consequences (644)	.98	.98	6412.65	.001	-.327	-.65	.745	.77	.568	.32

[a]Sample size ranged from 632 to 644, all correlations significant at $p < .001$.

[b]Model Statistics: ($df = 3, 628\text{-}640$).

[c]All β coefficients significant at $p < .001$.

Table 4

Supervisor Performance Ratings and Career Intent Regressed on PSMs for LKT Characteristics, LKT Skills, and Consequences Test[a]

Scale (n)	Distance (Step 1; Baseline)			PSMs: Shape, Delta, and Scatter (Step 2; H3)				Elevation (Step 3; H4)			
	R	*F* change	Sig	*R*	ΔR^2	*F* change	Sig	*R*	ΔR^2	*F* change	Sig
Supervisor Performance Ratings											
LKT Characteristics (635)	.10	6.20	.013	.19	.025	5.53	.001	.19	.001	0.46	.498
LKT Skills (632)	.08	4.16	.042	.19	.028	6.16	.001	.19	.000	0.20	.651
Consequences (644)	.10	6.21	.013	.21	.033	7.45	.001	.22	.006	4.01	.046
Career Intent											
LKT Characteristics (634)	.03	0.69	.407	.06	.002	0.42	.737	.19	.032	20.99	.001
LKT Skills (631)	.02	0.29	.593	.16	.026	5.52	.001	.21	.018	12.00	.001
Consequences (643)	.09	5.61	.018	.12	.005	1.12	.339	.12	.001	0.49	.486

[a]Degrees of freedom (*df*) lost at each regression step: 1 *df* at Step 1, 3 *df* at Step 2, 1 *df* at Step 3.

The fourth hypothesis proposed that re-centering the key may increase scale validity. We assessed Hypothesis 4 by entering the elevation metric at step 3 within each of the six regression models. The regression analyses supported Hypothesis 4 by demonstrating that the elevation term provided incremental validity against the Career Intent outcome for the LKT Characteristics and Skills scales, and against the performance outcome for the Consequences Rating Test. Moreover, the validity gains at step 3 were substantial for the LKT Skills and Characteristics scales. Table 4 provides results for the six regression models used to evaluate the predictive validity hypotheses.

We also modified the order of the regression steps to assess whether distance scores provide incremental validity beyond the PSMs for each judgment test. Table 5 summarizes the regression analyses and indicates that the distance scores provided significant incremental validity beyond the PSMs for two of the six regression models. However, the incremental validity estimates for the distance scores were relatively minor (i.e., all $\Delta R^2 \leq .018$). While these results confirm the expectation that the predictive validity of the judgment tests primarily reflects the PSMs, a better understanding of distance scores might further improve the validity of these instruments (i.e., by better understanding any predictive variance that is not assessed by the PSMs as main effects).

Additional Analysis. We extended the analyses by regressing each criterion onto only the relevant PSMs (i.e., shape, delta, scatter, and elevation if Hypothesis 4 was supported), and we report the β weights in Table 6. Comparison of the β weights reported in Tables 3 and 6 shows that distance scores represent sub-optimally weighted PSM composites that underestimate the validity of rating-based judgment tests. The regression weights indicate that the shape metrics are consistently the most potent predictor of performance, whereas the elevation metric was the strongest predictor of Career Intent for the LKT Characteristics and Skills scales. These results show that PSM composite scores may be differentially computed to predict performance or continuance outcomes.

SJT Validity Comparison. Although the above analyses support the PSM hypotheses, the use of rating based judgment tests with minimal encoding requirements in place of conventional SJTs raises the issue of the overall efficacy of this approach. Meta-analysis indicates that conventional SJTs have modest validity ($\bar{r}$ = .26 and ρ = .34; McDaniel et al., 2001). Therefore, we documented the combined utility of the three judgment tests by regressing the officer performance and career continuance outcomes onto the scale metrics. We conducted these analyses by entering the shape metrics in step 1; the delta metrics in step 2; the scatter metrics in step 3; and the elevation metrics in step 4. This order was followed because most maximum performance measures are shape scored, the delta metric references the scoring key and is analogous to a shape measure, while the scatter and elevation metrics represent descriptive statistics. Table 7 summarizes the regression results for the full sample, as well as for the Lieutenant and Captain subsamples.

Table 5

Supervisor Performance Ratings and Career Intent Regressed on PSMs for LKT Characteristics, LKT Skills, and Consequences Test

Scale	PSMs: Shape, Delta, and Scatter				Elevation					Distance				
	R	*F* change	*df*	Sig	*R*	ΔR^2	*F* change	*df*	Sig	*R*	ΔR^2	*F* change	*df*	Sig
Supervisor Performance Ratings														
LKT Characteristics	.19	7.65	3,631	.001	.19	.001	0.39	1,630	.531	.19	.000	0.10	1,629	.756
LKT Skills	.19	7.48	3,628	.001	.19	.000	0.06	1,627	.811	.19	.001	0.48	1,626	.490
Consequences	.18	6.76	3,640	.001	.18	.000	0.03	1,639	.860	.22	.018	12.24	1,638	.001
Career Intent														
LKT Characteristics	.05	0.43	3,630	.728	.16	.025	16.11	3,629	.001	.19	.008	5.44	3,628	.020
LKT Skills	.14	4.16	3,627	.006	.21	.024	15.68	3,626	.001	.21	.001	0.71	3,625	.400
Consequences	.11	2.38	3,639	.069	.12	.003	1.90	3,638	.168	.12	.001	0.42	3,637	.519

Table 6

Performance Outcomes Regressed on Shape, Delta, Scatter, and Elevation Metrics

Scale	Best PSM Model Statistics						Shape			Delta			Scatter			Elevation		
	R	R^2	Adj ΔR^2	*F* change	*df*	Sig	β	*r*	Sig	β	*r*	Sig	β	*r*	Sig	β	*r*	Sig
Supervisor Performance Ratings																		
LKT Char	.19	.035	.030	7.65	3,631	.001	0.16	.18	.001	-0.01	-.09	.802	0.04	.10	.343			
LKT Skill	.19	.034	.030	7.48	3,628	.001	0.14	.17	.001	-0.06	-.11	.180	0.05	.10	.255			
Consequences	.18	.031	.025	5.07	4,639	.001	0.10	.15	.041	-0.08	-.15	.107	0.04	.07	.351	-.01	-.08	.860
Career Intent																		
LKT Char	.16	.027	.021	4.36	4,629	.002	0.00	.02	.995	0.03	.02	.565	0.07	.03	.105	0.16	.15	.001
LKT Skill	.21	.043	.037	7.11	4,626	.001	-0.01	-.04	.734	0.04	.12	.321	-0.02	-.10	.648	0.18	.20	.001
Consequences	.11	.011	.006	2.38	3,639	.069	-0.10	-.10	.035	0.01	.05	.850	0.04	.02	.346			

Table 7

PSM Validity Estimates for the Full and Subsamples

Sample (*n*)	Step 1: All Shape			Step 2: All Delta				Step 3: All Scatter				Step 4: All Elevation			
	R	*F* Change	Sig	*R*	ΔR^2	*F* Change	Sig	*R*	ΔR^2	*F* Change	Sig	*R*	ΔR^2	*F* Change	Sig
Supervisor Performance Ratings															
Captains (212)	.24	4.16	.007	.33	.052	3.96	.009	.33	.003	0.236	.872	.35	.012	0.91	.439
Full Sample (629)	.22	10.31	.001	.23	.006	1.37	.251	.24	.002	0.52	.671	.24	.002	0.48	.695
Lieutenants (417)	.19	5.25	.001	.23	.014	2.00	.113	.24	.006	0.834	.476	.25	.007	1.067	.363
Career Intent															
Captains (212)	.16	1.79	.151	.24	.031	2.27	.082	.27	.014	1.02	.385	.31	.025	1.82	.145
Full Sample (628)	.12	3.23	.022	.17	.014	3.06	.028	.19	.008	1.73	.159	.24	.020	4.35	.005
Lieutenants (416)	.13	2.32	.075	.17	.011	1.55	.200	.18	.006	0.89	.447	.25	.029	4.19	.006

From a practical perspective, we were primarily concerned with estimating the predictive validity of the judgment tests against the supervisor performance ratings for the Captain subsample because the success of these individuals is believed critical to U.S. Army operational effectiveness (Paullin et al., 2014) and is reflected in DOD promotion policy that provides a more stringent promotion ratio for Captains, 76%, than for Lieutenants, 95% (Schirmer, 2016). The regression results computed for the Captain subsample indicate that the shape and delta metrics provide an impressive level of validity against the performance rating outcome, R = .33. This result also suggests that the content of the judgment tests aligns well with the job requirements of the captains. The lower validity estimate that was computed for these scales using the Lieutenant subsample, R = .19, is broadly consistent with the view that Lieutenant job requirements involve greater emphasis on face-to-face communications in place of indirect leadership skills (Paullin et al., 2014).

Regarding the Career Intent outcome, most officer loss occurs early in an officer's career. Therefore, predicting the Career Intent outcome for the Lieutenant subsample is critical. These regression results indicate that the judgment tests were predictive of Lieutenant Career Intent, R = .25, but not Captain Career Intent. Furthermore, much of the gain in incremental validity for the judgment tests against the career intent outcome was associated with the elevation terms for the LKT Skills and Characteristics scales as detailed in Table 6. This result may suggest a learned helplessness effect such that Lieutenants become disengaged from the military if they believe their attempts to engage in leadership activities are ineffective.

More generally, regression analyses demonstrated that: (a) the shape metrics were the most potent predictors of performance, (b) the delta metrics may supplement the shape metrics for the prediction of performance at the higher command level, and (c) the elevation metrics were important predictors of career intent.

Project 1 Summary

The regression analyses confirmed PSM hypotheses proposing that: the distance metrics, D and D^2, are nearly redundant; distance metrics represent linear composites of the shape, delta, and scatter metrics; and optimally weighting the shape, delta, scatter, and elevation metrics against specific criteria increments the validity of distance scores for the rating-based judgment test. Confirmation of these hypotheses supports the view that the PSM framework provides a potent method to optimize the validity of rating-based judgment tests against valued outcomes.

From a scale design perspective, the results showed that valid judgment tests may be created by describing brief scenarios, attaching multiple options to each scenario, and providing rating scales with a relatively large number of response categories. Despite the minimal administration requirements of these tests, the validity estimate of the rating-based judgment tests for the Captain subsample, R = .33, compared favorably to validity estimates for conventional SJTs that are based on meta-analysis ($\overline{r}$ = .26 and ρ = .34; McDaniel et al., 2001).

Project 2: PSMs to Increment Conventional Personality Scale Validity

For the second project, we used an established personality inventory to evaluate the PSM hypotheses and assess the utility of these metrics for incrementing scale validity against performance outcomes. Although validity expectations for these types of personality scales are based on the compelling rationale that past performance predicts future performance, personality scale validities tend to be modest (Barrick & Mount, 1991; Hogan, 2005; Hough & Oswald, 2000; McHenry et al., 1990; Schmidt & Hunter, 1998). Therefore, we reasoned that distance scoring may have suppressed the validity of these scales, and we reanalyzed a large dataset that had been collected to validate the personality battery against performance outcomes to evaluate the PSM hypotheses. We also used data from an independent sample to cross-validate the results for a subset of the personality scales.

Method

Participants. The primary sample consisted of 4,192 cadets in the U.S. Army's Reserve Officer Training Corps (ROTC) who participated in the Leader Development Assessment Course (LDAC) during the summer of 2013 and provided useable data. The cross-validation sample consisted of 4,283 ROTC cadets who participated in LDAC during the summer of 2015. The demographic composition of the primary and cross-validation samples were similar. Both samples were primarily male, 78%. Individuals in the two samples identified as: Caucasian, 82%; African-American, 11%; Asian, 7%; American Indian or Alaskan Native, 2%; and Native Hawaiian or Pacific Islander, 1%. Approximately 12% of the sample identified as Hispanic.

Design & Procedure. Our primary dataset contained personality data that had been collected from individuals who attended LDAC during the summer of 2013. Performance data were subsequently obtained for these participants and used to estimate scale validities. We used the 2013 dataset to conduct the primary analyses (i.e., to evaluate the five PSM hypotheses and estimate the level of incremental validity of PSM-based scale scores over distance scores for each personality scale). We also used this sample to create battery-level composite scores and estimate the gains in incremental validity that could be obtained by optimally weighting the PSM scale scores over distance scores.

To address concerns that PSM scoring algorithms may capitalize on sample specific variance, we cross-validated the 2013 scoring algorithms using data that were collected in 2015 from an independent sample. Unfortunately, 3 of the 10 scales that were administered in 2013 were not administered in 2015. In addition, PSM scoring did not increment the validity of 1 of the 10 personality scales. Therefore, we only cross-validated the PSM scale scores for the 6 common scales, as well as composite scores that were based on those common scales. Despite these limitations, this design represents a strong assessment of the possibility that PSMs capitalize on sample specific variance due to the 2-year delay between the data collections.

The Army ROTC cadets were administered the personality scales as a part of a battery of paper and pencil tests during their initial week at LDAC. The outcome criteria were collected

after the cadets had completed LDAC. The same data collection procedure was followed for the development sample in 2013 and the cross-validation sample in 2015.

Measures.

Cadet Background and Experiences Form (CBEF). The CBEF is a multiple-choice personality inventory that assesses past behaviors and experiences and is designed to predict officer performance and retention (Kilcullen, Robbins & Tremble, 2009). The CBEF contains approximately 120 items that use a 5-point Likert scale. We reviewed the versions of the CBEF that had been administered in 2013 and 2015, and we identified seven scales that could be analyzed for both samples using PSMs because they contained a mix of reversed and non-reversed items. We also identified three personality scales that were only administered to the 2013 sample. Table 1 contains example items to illustrate the CBEF item format, and these items were distributed over the ten scales described in Table 8.

Order of Merit Score (OMS). The OMS metric was provided by the U.S. Army Cadet Command (USACC) and represents our primary outcome measure. These scores reflect cadet performance in academic, military training, and physical fitness programs (e.g., college GPA, LDAC performance assessments, and Army Physical Fitness Test scores) as well as supervisor ratings of cadet leadership potential. OMS is an important outcome to predict because USACC awards ROTC scholarships to individuals who are likely to obtain high OMS scores, and the U.S. Army uses OMS to assign cadets to U.S. Army components and critical occupations. Therefore, we used OMS as the primary criterion to validate the CBEF personality scales.

Data Analysis. To evaluate the first four hypotheses, we computed the following six metrics for each personality scale using the conventional key: mean item distance, $D = \sum|X_i - K_i|/n$; mean item distance squared, $D^2 = \sum(X_i - K_i)^2/n$; shape = $r_{x,k}$; delta, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$; scatter = sd_x^2; and elevation = $\sum X_i/n$. The conventional scale keys correspond to an extreme value for each item (i.e., 5 for non-reversed items and 1 for reversed items). To evaluate Hypothesis 5, we computed shape-consensus scores: $r_{x,consensus}$, with the consensus keys based on the data collected from the primary sample.

Results

Descriptive Statistics. Table 9 contains descriptive information for the CBEF scale scores. The distance scale scores demonstrated acceptable levels of reliability (r_{xx} = .60 to r_{xx} = .82) with the exception of safety (r_{xx} = .41). Table 9 also documents that the ratio of reversed to non-reversed items varied widely over 10 scales.

PSM Hypotheses.

Foundational Hypotheses. The first hypothesis proposed that D and D^2 scores would be highly correlated for each of the 10 personality scales. As detailed in Table 10, D and D^2 scale scores approached redundancy for each scale (all $r > .92$, all $p < .001$). Therefore, Hypothesis 1 was supported.

Table 8

CBEF Scales and Definitions[a]

Scale	Definition
Administered to Developmental and Cross-Validation Samples	
Army Identification	Degree of identification with, and interest in being, a U.S. Army Soldier.
Fitness Motivation	Degree of enjoyment from physical exercise and willingness to stay physically fit.
Oral Communication	Degree of comfort with oral communication.
Stress Tolerance	Degree of emotional control and composure under pressure.
Tolerance for Injury	Degree of enjoyment from risky and hazardous activities.
Past Withdrawal	Degree of commitment and continuance in groups.
Written Communication	Degree of comfort with written communication.
Administered to Only the Developmental Sample	
Goal Orientation Continuance	Degree of motivation towards remaining in the Army.
Goal Orientation Performance	Degree of motivation towards achieving performance goals.
Safety	Degree of adherence to safety procedures.

[a]Refer to Kilcullen, Robbins & Tremble (2009) and Allen & Young (2012) for additional information regarding the constructs.

Table 9

Descriptive Statistics of CBEF Scales for Distance Scores

Scale	Mean[a]	SD	Reliability Coefficient α	Item Reversal Ratio
Administered to Developmental and Cross-Validation Samples				
Army Identification	1.09	0.58	.82	1:11
Fitness Motivation	1.23	0.53	.82	5:9
Oral Communication	1.06	0.40	.68	4:7
Past Withdrawal	2.87	0.44	.60	3:5
Stress Tolerance	1.79	0.48	.66	1:10
Tolerance for Injury	1.37	0.70	.69	1:4
Written Communication	1.72	0.65	.74	2:5
Administered to Only the Developmental Sample				
Goal Orientation Continuance	1.62	0.96	.91	1:6
Goal Orientation Performance	1.63	0.59	.74	1:6
Safety	1.38	0.51	.41	2:4

[a]Distance scale scores range from 0 to 4 with superior scores approaching 0.

Table 10

CBEF Distance Scores Regressed on PSMs

Scale	H1[a]	H2[b]			Shape[c]		Delta[c]		Scatter[c]	
	r_{D,D^2}	$R_{D,PSMs}$	*F* change	*Sig*	*β*	*r*	*β*	*r*	*β*	*r*
CBEF Scales Administered to Both Samples										
Written Communication	.96	.98	34327.58	.001	-.79	-.94	.26	.64	-.14	-.34
Fitness Motivation	.96	.99	49843.35	.001	-.64	-.91	.15	.42	-.40	-.82
Army Identification	.95	.98	27290.73	.001	-.41	-.86	.63	.93	-.05	-.18
Stress Tolerance	.96	.99	47636.84	.001	-.36	-.72	.76	.93	-.08	-.10
Tolerance for Injury	.95	.96	15143.18	.001	-.67	-.88	.42	.75	-.10	-.10
Past Withdrawal	.98	.98	42932.22	.001	-.64	-.84	.19	.17	.52	.77
Oral Communication	.93	.97	19447.11	.001	-.68	-.88	.01	.21	-.45	-.75
CBEF Scales Administered to Only the Developmental Sample										
Safety	.92	.95	13893.59	.001	-.63	-.82	.31	.54	-.37	-.61
Goal Orientation Performance	.95	.97	26831.85	.001	-.63	-.86	.51	.80	-.01	.07
Goal Orientation Continuance	.96	.98	33373.49	.001	-.41	-.90	.56	.93	-.15	-.53

[a]Sample size ranged from 4119 to 4192, all correlations significant at $p < .001$.

[b]Model Statistics: ($df = 3, 4115\text{-}4185$).

[c]All β coefficients significant at $p < .001$.

The second hypothesis proposed that distance scores for each of the 10 personality scales could be modelled as a composite of shape, delta, and scatter effects. Regression analyses supported Hypothesis 2 by documenting that distance scores for each of the 10 personality scales primarily represent a PSM composite formed by regressing the scale scores onto the shape, scatter, and delta metrics (all $R > .95$, all $p < .001$; see Table 10). These results demonstrate that distance scores represent a PSM composite for each personality scale.

Predictive Validity Hypotheses. The third hypothesis proposed that optimally weighting the shape, delta, and scatter metrics may provide incremental validity over distance scores against the performance outcome (OMS) for each of the personality scales. Hypothesis 3 was supported for 8 of 10 personality scales, and the regression results are reported in Table 11 (Step 2).

The fourth hypothesis proposed that re-centering the key may improve scale validity. Regression analyses supported Hypothesis 4 for 3 of the 10 scales as reported in Table 11 (Step 3). Although this result supports Hypothesis 4, we suspect that the coarse 5-point scale precluded more consistent support of this hypothesis.

The fifth hypothesis proposed that shape-consensus scores may increment shape metrics that are based on conventional keys. This hypothesis addresses the possibility that conventional keys may have relatively primitive shape that limits scale validity. Regression analyses supported this hypothesis for 7 of the 10 scales as reported in Table 11 (Step 4).

Scale and Composite Validity Analyses.

Scale Results. Validity gains were documented for 9 of the 10 scales based on the endorsement of the three PSM predictive validity hypotheses (e.g., the Written Communication and Fitness Motivation scale analyses supported Hypothesis 3 at Step 2 and Hypothesis 5 at Step 4). However, the inclusion of distance scores and PSMs and the use of multiple shape measures raise interpretation and multi-collinearity issues. To clarify the results, we re-computed PSM scores for each scale using only the delta, scatter, elevation, and either the shape metric, which utilized the conventional key, or the shape-consensus metric. These regression estimates are reported in Table 12 and can be directly compared to the distance validities reported at step 1 in Table 11 for each scale.

From a predictive perspective, the most substantial increment in scale validity was obtained for the Written Communication scale, $R = .32$ vs. $R = .17$. In addition, less substantial gains were documented for several scales including: Goal Orientation Continuance, $R = .11$ vs. $R = .04$; Safety, $R = .08$ vs. $R = .01$; and Fitness Motivation, $R = .31$ vs. $R = .28$. (Compare Tables 12 and 11.) Unlike the distance validities, these results support expectations based on psychological models and job analysis that higher performing officer cadets excel at communication tasks, are concerned with safety, and are goal oriented (Paullin et al., 2014).

Composite Results. We used hierarchical regression procedures to model the utility of PSM scoring algorithms for applied settings. Therefore, we regressed the performance outcome onto the CBEF distance scale scores in step 1. We then added the PSM scale scores in step 2.

Table 11

Developmental Sample: OMS Regressed on Distance, PSM, Elevation, and Shape-Consensus Metrics by Personality Scale[a]

Scale	Distance (Step 1; Baseline)			PSMs: Shape, Delta, and Scatter (Step2; H3)				Elevation (Step 3; H4)				Shape-Consensus (Step 4; H5)			
	R	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig
Written Communication	.17	125.75	.001	.25	.035	51.89	.001	.25	.000	1.89	.169	.33	.044	208.01	.001
Fitness Motivation	.28	342.82	.001	.30	.012	17.92	.001	.30	.000	1.91	.167	.32	.014	64.90	.001
Goal Continuance[b]	.04	5.96	.015	.11	.010	13.84	.001	.13	.005	19.50	.001	.14	.002	9.78	.002
Army Identification	.01	0.18	.667	.09	.007	10.02	.001	.09	.001	2.89	.089	.10	.002	9.22	.002
Tolerance for Injury	.07	18.87	.001	.07	.001	1.09	.354	.09	.002	8.39	.004	.10	.004	15.23	.001
Oral Communication	.12	63.40	.001	.13	.001	1.23	.299	.13	.000	0.20	.653	.13	.001	1.81	.178
Goal Performance[b]	.17	120.45	.001	.18	.005	6.96	.001	.18	.001	3.89	.049	.19	.002	8.46	.004
Past Withdrawal	.03	4.06	.044	.05	.002	2.81	.038	.06	.000	0.45	.500	.07	.001	5.40	.020
Stress Tolerance	.09	36.30	.001	.11	.003	4.57	.003	.11	.001	2.17	.141	.11	.000	1.58	.209
Safety[b]	.01	0.79	.375	.08	.007	9.55	.001	.09	.000	1.95	.163	.09	.000	0.28	.594

[a]Degrees of freedom lost at each regression step: 1 *df* at Step 1, 3 *df* at Step 2, 1 *df* at Step 3. Sample sizes ranged from 4179 to 4221.
[b]Not administered to the cross sample.

Table 12

Developmental Sample: OMS Regressed on Profile Similarity Metrics by Personality Scale[a]

Scale	Model Statistics						Shape		Delta		Scatter		Elevation[b]	
	R	ΔR^2	Adj ΔR^2	*F*-change	*df*	Sig	β	*r*	β	*r*	β	*r*	β	*r*
Shape Based on Consensus Keys														
Written Communication	.32	.105	.104	164.52	3,4217	.001	0.22	.23	-0.08	-.13	0.20	.22		
Fitness Motivation	.31	.098	.097	152.04	3,4217	.001	0.16	.26	-0.14	-.21	0.11	.24		
Goal Continuance[c]	.11	.013	.012	18.09	3,4205	.001	0.12	.01	0.16	.07	0.03	.00		
Army Identification	.10	.009	.009	13.24	3,4217	.001	0.08	.05	0.07	.01	0.06	.07		
Tolerance for Injury	.10	.010	.009	10.34	4,4179	.001	0.07	.08	-0.06[ns]	-.06	-0.02[ns]	.02	-0.09	-.07
Shape Based on Conventional Keys														
Goal Performance[c]	.18	.033	.032	47.36	3,4209	.001	0.13	.16	-0.08	-.13	0.06	.04		
Stress Tolerance	.11	.013	.012	18.02	3,4215	.001	0.09	.10	-0.04	-.08	-0.03[ns]	-.02		
Past Withdrawal	.04	.002	.001	2.80	3,4216	.038	-0.03	-.03	0.03	.03	-0.01[ns]	.01		
Safety[c]	.08	.006	.006	8.99	3,4201	.001	0.04	.03	-0.03	-.03	-0.07	-.05		

[a]All coefficients are significant ($p < .05$) unless otherwise noted.
[b]Elevation was included in the TI equation because H4 was supported.
[c]Not administered to the cross-validaton sample.
[ns]Not significant.

We also reversed the entry order of the PSM and distance scale scores to determine whether the distance scores would increment PSM scores at the composite level.

Both steps of each hierarchical model were statistically significant as reported in Table 13. However, the regression analyses documented that PSM scales scores provided a larger increment to the distance scale composite (R = .46 vs. R = .35; ΔR^2 = .09) than the distance scores provided to the PSM scale composite (R = .46 vs. R = .44; ΔR^2 = .02).

Table 13
Developmental Sample: Composite Validity Against OMS for the Nine Distance versus Nine PSM Scale Scores

Model Steps	R	R^2	Adj R^2	Change Statistics			
				ΔR^2	F	df	Sig
Model 1: Distance First, PSM Second							
1. Distance Scores	.35	.124	.122	.124	60.60	9,3851	.001
2. PSM Scores	.46	.213	.209	.089	48.34	9,3842	.001
Model 2: PSM First, Distance Second							
1. PSM Scores	.44	.193	.191	.193	102.09	9,3851	.001
2. Distance Scores	.46	.213	.209	.021	11.14	9,3842	.001

Cross-Validation Analyses. To evaluate the possibility that PSM scoring capitalizes on sample-specific variance, we used the 2013 sample to optimally weight the CBEF scales and composite (i.e., the consensual keys and the regression weights were based on the 2013 sample and applied to the 2015 sample). However, only 7 of the 10 scales were administered to both samples, and PSMs did not increment one of those scales. Therefore, the cross-validation composite was computed using the 6 common scales and the 2013 sample. We then cross validated the 2013 algorithms using the 2015 dataset. Table 14 summarizes the results.

Table 14
Cross-Validated Estimates for Composite, PSM and Distance Metrics

CBEF	n	Correlation PSM by Distance[a]	Validity Against OMS		
			Distance[a]	PSM	ΔSig[b]
Composite	4219	.78**	.32**	.40**	.001
Written Communication	4282	.62**	.19**	.31**	.001
Fitness Motivation	4283	.94**	.27**	.31**	.001
Stress Tolerance	4268	.84**	.15**	.17**	.033
Past Withdrawal	4280	.62**	.10**	.11**	.450
Tolerance for Injury	4238	.78**	.11**	.11**	.999
Army Identification	4276	.02	.03	.00	.161

[a]Followed the Steigler (1980) to test the difference between two dependent correlations with one variable in common. Program available at http://quantpsy.org/corrtest/corrtest2.htm.
[b]Distance scores were reflected so that all correlations would be positive.
**: p < .01 level (2-tailed).

Of critical importance regarding the cross-sample analyses, the six-scale PSM composite validity was substantially higher than the six-scale distance composite validity, r = .40 vs. r = .32. At the construct level, the validity of the Written Communication and Fitness Motivation PSM scale scores continued to have substantially higher scale validities than the corresponding distance scores (Written Communication, r = .31 vs. r = .19; Fitness Motivation, r = .31 vs. r = .27). Furthermore, the higher battery composite and scale validities were obtained for the PSM-based scores despite the 2-year delay between data collection for the developmental and cross-validation samples. While decreases for the 6-scale composite validities were observed between the developmental and cross-samples for both scoring algorithms, the magnitudes of these decreases were minimal and their consistency may represent minor sampling effects (PSM, $R_{\text{developmental}}$ = .43 vs. r_{cross} = .40; distance, $R_{\text{developmental}}$ = .33 vs. r_{cross} = .32).

Project 2 Summary

The regression analyses conducted using conventional personality data supported the two foundational hypotheses proposing that: (H1) the distance metrics, D and D^2, are nearly redundant; (H2) distance metrics represent linear composites of the shape, delta, and scatter metrics. In addition, results supported the three predictive validity hypotheses proposing that personality scale validity may be incremented by: reweighting the shape, delta, and scatter metrics (H3); adding the elevation metric (H4); and using the shape-consensus metric (H5).

Support for these hypotheses, coupled with the cross-validation analyses, is consistent with the view that the PSM framework provides a potent method to enhance the validity of rating-based personality scales against important performance outcomes (R = .44 vs. R = .35). We also emphasize that the composite validity estimate based on PSM scoring, R = .44, approaches validity estimates that are frequently associated with general cognitive ability ($\overline{r}$ = .51, Schmidt & Hunter, 1998). Finally, the validity gains for the individual scales carry implications for theory regarding psychological constructs that are expected to relate to cadet performance (e.g., Written Communication, Goal Orientation, and Safety).

Project 3: Optimizing Validity for Experimental Personality Scales

For the third project, we analyzed data for the Continuum Personality Measure (CPM; Kilcullen et al., 2013). The CPM differs from most conventional scales because each item has a large 9-point response continuum with anchor statements that reflect high and low standing on a single personality attribute. In addition, each scale had a near even mix of reversed and non-reversed items. In contrast, most widely-used personality scales (e.g., the NEO, Costa & McCrae, 1991) use smaller rating scales. The use of small rating scales is consistent with the view that large rating scales provide only minimal improvement in the psychometrics of distance scores, yet have greater administration requirements (Cox, 1980; Nunnally, 1978; Preston & Colman, 2000). Because preliminary analyses using distance scores for the CPM had resulted in modest scale validities, we speculated that distance scores for 9-point scales might mimic the action of poorly weighted PSMs and reduce scale validity. Therefore, we conducted analyses using the CPM to evaluate the PSM hypotheses and reexamine scale validity.

We also reasoned that using multiple methods to assess personality would boost the predictive validity of the personality measures. Because respondents had also completed the CBEF, we conducted analyses to estimate the incremental validity of the CPM scales beyond the CBEF scales. We conducted these analyses twice, using either PSM-based scores or distance scores for all scales. Based on the results from the second project, we speculated that the combined validity of PSM-based scales computed for the two batteries would exceed the standard for a large validity coefficient (r = .50; Cohen, 1988) and rival the validity of general cognitive ability ($\overline{r}$ = .51; Schmidt & Hunter, 1998). Finally, we compared CPM and CBEF scale validities for overlapping constructs to evaluate the potential of the CPM approach to provide alternate measures for CBEF scales.

Method

Subjects. The sample consists of 3,909 ROTC cadets in the U.S. Army who provided useable data and participated in LDAC during the summer of 2016. The sample was primarily male, 77%. Respondents self-identified as: Caucasian, 80%; African-American, 12%; Asian, 8%; American Indian or Alaskan Native, 2%; and Native Hawaiian, Pacific Islander or Multi-racial, 1%. Approximately, 13% of the sample identified their ethnicity as Hispanic.

Design & Procedure. The ROTC cadets were administered the CPM and the CBEF during their initial week at LDAC. The outcome criterion, OMS, was collected from the U.S. Army Cadet Command after the cadets completed LDAC. To assess the PSM hypotheses, we scored both batteries using PSMs and distance metrics. In addition, we did not conduct item analyses to eliminate poorly performing items.

Measures.

Continuum Personality Measure (CPM). The CPM contains 99 items that were designed to assess past behaviors and experiences. Each item required respondents to rate the extent to which two opposing statements describe their behaviors and experiences using a 9-point rating scale. Table 1 contains example items to illustrate the CPM format, and these items were distributed over nine domains that are described and identified in Table 15.

Cadet Background and Experiences Form (CBEF). This version of the CBEF was used as the established personality battery (Kilcullen et al., 2009). For the 2016 LDAC sample, the CBEF contained 103 items that were distributed over 13 scales with 8 scales containing a mix of reversed and non-reversed items, and 5 scales not containing any reversed items. Therefore, we conducted PSM analyses on only the 8 CBEF scales with reversed items, although we used all 13 scales to estimate CBEF composite validity. The CBEF scales are listed in Table 15.

Table 15
CPM and CBEF Scales and Definitions[a]

Scale	CPM	CBEF	Definition
Achievement Orientation	x	x	The willingness to give one's best effort and to work hard towards achieving difficult objectives.
Army Identification	x	x	Degree of identification with, and interest in being, a U.S. Army Soldier.
Cognitive Flexibility	x		Willingness to entertain new approaches to solving problems. Enjoys creating new plans and ideas. Initiates and accepts change and innovation.
Fitness Motivation	x	x	Degree of enjoyment from physical exercise and willingness to stay physically fit.
Hostility to Authority	x	x	Suspicious of the motives and actions of legitimate authority figures. Views rules and directives from authority as illegitimate.
Peer Leadership	x	x	Seeks positions of authority. Comfortable with being in charge of a group and accepts responsibility for the group's performance.
Self-Efficacy	x	x	Feeling that one has successfully overcome past work obstacles.
Stress Tolerance	x	x	Degree of emotional control and composure under pressure.
Tolerance for Ambiguity	x		Ability to tolerate work situations where the right goal or the correct path to the goal is vague and ill-defined.
Guilt Proneness		x	Tendency to experience negative feelings regarding one's actions involving specific wrong or foolish behaviors.
Past Withdrawal		x	Degree of commitment and continuance in groups
Tolerance for Injury		x	Degree of enjoyment from risky and hazardous activities
Written Communication		x	Degree of comfort with written communication
Shame Proneness		x	Tendency to make global attributions regarding one's self, which lead to negative feelings about the global self.
Lie		x	A response distortion scale designed to detect socially desirable responding.

[a]Refer to Kilcullen, Robbins & Tremble (2009) and Allen & Young (2012) for additional information regarding the constructs.

Order of Merit Score (OMS). As in the second project, OMS was used as the principal criterion variable. OMS is an important outcome to predict because USACC awards ROTC

scholarships to individuals who are likely to obtain high OMS scores, and the U.S. Army uses OMS to assign cadets to U.S. Army components and critical occupations.

Data Analysis. To evaluate the first four hypotheses, the following six metrics were computed for each personality scale using the conventional key: mean item distance, $D = \sum|X_i - K_i|/n$; mean item distance squared, $D^2 = \sum(X_i - K_i)^2/n$; shape = $r_{x,k}$; delta, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$; scatter = sd_x^2; and elevation = $\sum X_i/n$. The conventional scale keys correspond to an extreme value for each item (e.g., 9 for CPM non-reversed items and 1 for reversed items). To evaluate Hypothesis 5, we computed shape-consensus scores: $r_{x,consensus}$, with the consensus keys based on the data collected from the primary sample.

Results

Descriptive Statistics. Table 16 reports descriptive statistics for the CPM and CBEF scales. Reliabilities were acceptable for all measures, but generally higher for the CBEF scales.

Table 16
Descriptive Statistics of CPM and CBEF Scales for Distance Scores

Scale	Mean	SD	Reliability Coefficient α	Item Reversal Ratio
CPM Scales (9-Pt Rating)				
Achievement Orientation	0.50	0.84	.60	4:14
Army Identification	0.14	1.51	.61	2:3
Cognitive Flexibility	1.30	1.11	.54	1:5
Fitness Motivation	0.17	1.52	.88	6:6
Hostility to Authority	1.35	1.01	.76	6:7
Peer Leadership	0.17	1.03	.76	6:7
Self-Efficacy	0.45	1.03	.64	3:11
Stress Tolerance	0.74	1.01	.70	5:7
Tolerance for Ambiguity	1.16	1.28	.72	1:5
CBEF Scales With Reversed Items (5-Pt Rating)				
Army Identification	0.99	0.56	.82	1:10
Fitness Motivation	1.10	0.66	.84	3:4
Guilt Proneness	1.00	0.48	.69	4:5
Past Withdrawal	1.08	0.45	.63	3:5
Peer Leadership	1.43	0.54	.75	3:6
Stress Tolerance	1.78	0.50	.78	1:10
Tolerance for Injury	1.20	0.73	.73	1:4
Written Communication	1.64	0.69	.75	2:5
CBEF Scales Without Reversed Items (5-Pt Rating)				
Achievement Orientation	0.81	0.50	.71	0:9
Hostility to Authority	3.18	0.52	.53	0:4
Self-Efficacy	0.57	0.42	.73	0:6
Shame Proneness	2.23	0.51	.68	0:10
Lie[a]	0.10	0.16	.71	2:5

[a]Includes Lie. However, Lie was not PSM scored because it is not distance scored.

PSM Hypotheses.

Foundational Hypotheses. The first hypothesis proposed that distance-based scale scores would be highly correlated for each personality scale. As detailed in Tables 17 and 18, the D and D^2 scores approached redundancy for each CPM scale (all $r > .90$, all $p < .001$) and each CBEF scale (all $r > .92$, all $p < .001$). Therefore, Hypothesis 1 was supported for all CPM and CBEF scales.

The second hypothesis proposed that distance scores can be accurately modelled as main effects associated with the shape, delta, and scatter metrics. Therefore, we regressed the distance scale scores onto the corresponding PSMs for each of the CPM and CBEF scales (all $R > .93$, all $p < .001$). These results demonstrate that distance scores can be viewed as a PSM composite for each personality scale and support Hypothesis 2. Comparison of Tables 17 and 18 also indicates that the CPM distance scores primarily represent shape variance, while the CBEF distance scores represent a mix of shape, scatter, and delta effects. The result that CPM distance scores primarily represent shape variance may reflect the design of these scales to have a near-even mix of reversed and non-reversed items.

PSM Validity Hypotheses. The third hypothesis proposed that the shape, scatter, and delta metrics may add incremental validity beyond the distance metric for each CPM and CBEF scale. Hierarchical regression analyses supported this proposition for 6 of 9 CPM scales as well as 6 of 8 CBEF scales. See Tables 19 (CPM) and 20 (CBEF).

The fourth hypothesis proposed the elevation metric would provide incremental validity beyond the first three PSMs when key-elevation is poor. Therefore, we extended the regression analyses and added the elevation metric in the third step of the hierarchical regression models for each of the CPM and CBEF scales. Regression analyses supported this proposition for 7 of 9 CPM scales as well as 3 of 8 CBEF scales. See Table 19 (CPM) and Table 20 (CBEF).

The fifth hypothesis proposed that the shape-consensus metric may provide incremental validity beyond the first four PSMs when key-shape is poor. Therefore, we added the shape-consensus metric in the fourth step of the hierarchical regression models for each of the CPM and CBEF scales. Regression analyses supported this hypothesis for 4 of 9 CPM scales as well as 3 of 8 CBEF scales. The result that at least one of the two keying hypotheses was endorsed for 8 of the 9 CPM scales and 4 of the 8 CBEF scales suggests broad limitations with the accuracy of conventional personality keying procedures.

We also modified the order of the regression steps to determine whether distance would provide incremental validity beyond PSM scores for each scale. Although the distance scores provided significant incremental validity beyond the PSMs for several scales, the incremental validity estimates were consistently minor (i.e., all $\Delta R^2 \leq .003$; See Tables 21 & 22). These results support the view that nearly all of the predictive validity associated with each of these scales can be modelled as main effects associated with the PSMs.

Table 17

CPM Distance Scores Regressed on PSMs

Scale	H1[a]	H2[b]			Shape[c]		Delta[c]		Scatter[c]	
	r_{D,D^2}	$R_{D,PSMs}$	*F* change	*Sig*	β	*r*	β	*r*	β	*r*
Achievement	.90	.96	15907.76	.001	-.72	-.84	.46	.61	-.13	-.31
Army Identification	.95	.94	8846.87	.001	-.89	-.92	.07	.14	-.14	-.35
Hostility to Authority	.95	.95	11964.29	.001	-.93	-.94	.05	.05	.12	.19
Cognitive Flexibility	.95	.96	13858.41	.001	-.61	-.74	.62	.75	-.05	.07
Fitness Motivation	.96	.97	22348.56	.001	-.88	-.96	.02	.07	-.19	-.57
Peer Leadership	.92	.97	18800.94	.001	-.89	-.95	.04	.03	-.21	-.48
Self-Efficacy	.93	.97	18180.46	.001	-.74	-.91	.34	.69	-.11	-.20
Stress Tolerance	.94	.95	13299.69	.001	-.92	-.94	.13	.26	-.09	-.11
Tolerance for Ambiguity	.96	.93	8506.79	.001	-.86	-.91	.19	.37	.08	.18

[a]Sample size ranged from 3871 to 3892, all *r* significant at $p < .001$.

[b]Model Statistics: ($df = 3, 3712\text{-}3874$).

[c]All β weights significant at $p < .001$.

Table 18

CBEF Distance Scores Regressed on PSMs

Scale	H1[a]	H2[b]			Shape[c]		Delta[c]		Scatter[c]	
	r_{D,D^2}	$R_{D,PSMs}$	*F* change	Sig	β	*r*	β	*r*	β	*r*
Army Identification	.94	.97	24089.79	.001	-.46	-.85	.60	.90	-.07	-.21
Stress Tolerance	.95	.98	36667.20	.001	-.36	-.75	.74	.93	-.09	-.13
Guilt Proneness	.92	.99	52418.07	.001	-.59	-.86	.04	.25	-.55	-.84
Peer Leadership	.94	.96	15573.08	.001	-.75	-.91	.17	.45	-.27	-.59
Past Withdrawal	.92	.98	29480.38	.001	-.59	-.83	.15	.32	-.53	-.79
Written Communication	.96	.98	29342.74	.001	-.81	-.95	.24	.64	-.13	-.29
Fitness Motivation	.95	.99	42638.14	.001	-.62	-.89	.07	.24	-.48	-.84
Tolerance for Injury	.95	.96	16775.51	.001	-.60	-.88	.42	.77	-.20	-.38

[a]Sample size ranged from 3909 to 3909, all correlations significant at $p < .001$.

[b]Model Statistics: ($df = 3, 3864\text{-}3905$).

[c]All β weights significant at $p < .001$.

Table 19

OMS Regressed on Distance, PSM, Elevation, and Shape-Consensus Metrics by CPM Scale[a]

CPM Scale	Distance (Step 1; Baseline)			PSMs: Shape, Delta, and Scatter (Step2; H3)				Elevation (Step 3; H4)				Shape-Consensus (Step 4; H5)			
	R	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig
Achievement	.26	283.66	.001	.28	.011	15.24	.001	.28	.000	1.01	.315	.28	.000	0.45	.504
Army Identification	.03	2.46	.117	.05	.001	1.83	.139	.08	.005	18.95	.001	.08	.000	0.25	.617
Hostility to Authority	.01	0.47	.491	.13	.016	21.48	.001	.14	.004	14.31	.001	.17	.010	41.66	.001
Cognitive Flexibility	.01	0.51	.473	.07	.005	6.49	.001	.08	.001	3.68	.055	.13	.010	38.87	.001
Fitness Motivation	.17	117.36	.001	.17	.001	1.45	.227	.18	.002	8.40	.004	.21	.010	41.10	.001
Peer Leadership	.14	82.46	.001	.15	.001	1.47	.221	.17	.006	23.05	.001	.17	.000	0.97	.326
Self-Efficacy	.07	19.53	.001	.11	.007	8.47	.001	.12	.004	14.31	.001	.15	.008	31.39	.001
Stress Tolerance	.06	14.22	.001	.10	.007	8.80	.001	.13	.006	21.69	.001	.13	.001	3.28	.070
Tolerance for Ambiguity	.06	16.04	.001	.13	.012	16.21	.001	.15	.006	23.50	.001	.15	.001	3.29	.069

[a]Sample sizes ranged from 3716 to 3878. Degrees of freedom lost at each step: 1 *df* at Step 1, 3 *df* at Step 2, 1 *df* at Step 3, 1 *df* at Step 4.

Table 20

OMS Regressed on Distance, PSM, Elevation, and Shape-Consensus Metrics by CBEF Scale[a]

CBEF Scale	Distance (Step 1; Baseline)			PSMs: Shape, Delta, and Scatter (Step2; H3)				Elevation (Step 3; H4)				Shape-Consensus (Step 4; H5)			
	R	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig
Army Identification	.02	2.29	.131	.06	.003	4.48	.004	.06	.000	0.84	.358	.07	.000	1.72	.189
Stress Tolerance	.14	77.08	.001	.16	.005	6.80	.001	.16	.000	1.42	.233	.16	.000	0.43	.514
Guilt Proneness	.09	35.20	.001	.11	.004	5.45	.001	.11	.000	0.61	.433	.14	.007	29.67	.001
Peer Leadership	.17	109.92	.001	.17	.002	2.13	.094	.17	.001	2.14	.143	.17	.001	2.89	.089
Past Withdrawal	.08	22.21	.001	.11	.006	8.05	.001	.11	.001	4.48	.034	.11	.000	0.83	.363
Written Communication	.19	142.54	.001	.25	.029	39.84	.001	.26	.002	9.59	.002	.31	.032	137.03	.001
Fitness Motivation	.27	306.62	.001	.29	.012	16.73	.001	.30	.007	28.61	.001	.32	.014	59.46	.001
Tolerance for Injury	.11	45.87	.001	.11	.001	0.82	.480	.11	.001	2.56	.110	.12	.000	1.71	.191

[a]Sample sizes ranged from 3868 to 3909. Degrees of Freedom lost at each step: 1 *df* at Step 1, 3 *df* at Step 2, 1 *df* at Step 3, 1 *df* at Step 4.

Table 21

OMS Regressed on PSM, Elevation Shape-Consensus and Distance Metrics by CPM Scale [a]

CPM Scale	PSMs: Shape, Delta, and Scatter (Step1)			Elevation (Step 2)				Shape-Consensus (Step 3)				Distance (Step 4)			
	R	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig
Achievement	.28	110.01	.001	.28	.000	0.02	.893	.28	.000	.52	.315	.28	.001	3.28	.070
Army Identification	.04	2.16	.091	.08	.005	19.75	.001	.08	.000	.21	.647	.08	.000	0.73	.394
Hostility to Authority	.13	21.48	.001	.14	.003	13.36	.001	.17	.011	42.01	.001	.17	.000	0.91	.340
Cognitive Flexibility	.07	6.20	.001	.08	.001	5.02	.025	.13	.010	38.89	.001	.13	.000	0.03	.870
Fitness Motivation	.17	39.04	.001	.18	.002	8.10	.004	.20	.009	37.15	.001	.21	.002	8.74	.003
Peer Leadership	.14	27.12	.001	.16	.005	20.90	.001	.16	.000	1.24	.265	.17	.002	7.32	.007
Self-Efficacy	.10	13.16	.001	.12	.005	19.54	.001	.15	.008	30.95	.001	.15	.000	0.77	.380
Stress Tolerance	.09	9.59	.001	.12	.008	29.34	.001	.13	.001	2.82	.093	.13	.001	4.68	.031
Tolerance for Ambiguity	.13	21.00	.001	.14	.005	17.66	.001	.15	.001	3.88	.049	.15	.002	7.09	.008

[a]Sample sizes ranged from 3716 to 3878. Degrees of freedom lost at each step: 3 *df* at Step 1, 1 *df* at Step 2, 1 *df* at Step 3, 1 *df* at Step 4.

Table 22

OMS Regressed on PSM, Elevation Shape-Consensus and Distance Metrics by CBEF Scale [a]

CBEF Scale	PSMs: Shape, Delta, and Scatter (Step 1)			Elevation (Step 2)				Shape-Consensus (Step 3)				Distance (Step 4)			
	R	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig	*R*	ΔR^2	*F*	Sig
Army Identification	.06	4.18	.006	.06	.001	3.97	.046	.07	.000	1.67	.194	.07	.000	0.10	.747
Stress Tolerance	.16	32.61	.001	.16	.000	0.13	.718	.16	.000	0.46	.498	.16	.000	1.28	.259
Goal Performance	.11	16.86	.001	.11	.000	1.27	.259	.14	.007	29.41	.001	.14	.000	0.74	.398
Peer Leadership	.16	34.51	.001	.17	.003	13.16	.001	.17	.001	2.68	.102	.17	.000	1.78	.182
Past Withdrawal	.10	13.45	.001	.10	.000	0.00	.962	.10	.000	1.27	.260	.11	.003	10.14	.001
Written Communication	.25	88.79	.001	.26	.002	6.61	.010	.31	.031	135.63	.001	.31	.001	4.38	.037
Fitness Motivation	.29	119.38	.001	.30	.007	30.78	.001	.32	.013	56.90	.001	.32	.001	2.64	.104
Tolerance for Injury	.11	15.63	.001	.11	.001	3.90	.048	.12	.000	1.75	.186	.12	.000	0.05	.814

[a]Sample sizes ranged from 3868 to 3909. Degrees of freedom lost at each step: 1 *df* at Step 1, 3 *df* at Step 2, 1 *df* at Step 3, 1 *df* at Step 4.

Scale and Composite Validity Analyses.

Scale Validities. At least one of the three PSM validity hypotheses was confirmed for each of the 9 CPM scales and for 6 of the 8 CBEF scales. Moreover, many of these gains were substantial. To clarify interpretations and address multi-collinearity issues, we modelled scale validity based on the best 3 or 4 PSMs for each CPM (Table 23) and CBEF (Table 24) scale.

For the CPM battery, PSM scoring resulted in a mean scale validity of .16, while distance scoring provided a mean scale validity of .09. These mean validity estimates indicate that PSM scoring provided a 73% gain in mean validity over distance scoring for the CPM scales. (The validities for individual CPM scales are provided in Table 19 for distance scores and Table 23 for PSM scores.) Furthermore, PSM-based scores provided substantial incremental levels of validity beyond distance scores for the following CPM scales: Hostility to Authority, $R = .17$ vs. $R = .01$; Cognitive Flexibility, $R = .12$ vs. $R = .01$; Tolerance for Ambiguity, $R = .14$ vs. $R = .06$; Self Efficacy, $R = .14$ vs. $R = .07$. Smaller gains were obtained for: Stress Tolerance, $R = .11$ vs. $R = .05$; Fitness Motivation, $R = .20$ vs. $R = .17$; and Achievement Motivation, $R = .28$ vs. $R = .26$.

For the CBEF battery, PSM scoring resulted in a mean scale validity of .17, while distance scoring provided a mean scale validity of .13. These validity estimates indicate that PSM scoring provided a 26% gain in mean validity over distance scoring for the CBEF scales. (The validities for individual CBEF scales are provided in Table 20 for distance scores and Table 24 for PSM scores.) Furthermore, modest to substantial increases in scale validity were documented for the following CBEF scales using only the best 3 or 4 PSMs for each scale: Written Communication, $R = .31$ vs. $R = .19$; Fitness Motivation, $R = .31$ vs $R = .27$; Past Withdrawal, $R = .10$ vs. $R = .08$; and Guilt Proneness, $R = .13$ vs. $R = 09$.

Composite Validity. We used hierarchical regression analyses to estimate the predictive validity of composites against OMS based on either PSM or distance scale scores. For these analyses, we also included distance scores for four CBEF scales that were not PSM-scored because they did not contain reversed items (i.e., Achievement, Self-efficacy, Hostility to Aggression, Shame Proneness) and conventional scores for the Lie scale, which quantifies endorsement of socially desirable, yet implausible responses (cf. Reeder & Ryan, 2012).

We report two parallel hierarchical regression models in Table 25 that differ in the use of either conventional distance scores (Model 1) or PSM scores computed using the best 3 or 4 PSMs for each scale (Model 2). Model 1 was designed to estimate the validity of using distance scores for all the CBEF and CPM scales, while Model 2 was designed to estimate the validity of using PSMs to score all the scales that contained a mix of reversed items and distance scores for those CBEF scales that did not contain reversed items. Comparison of results for the two models addresses the overall value of using PSMs to score the CBEF and CPM scales.

Table 23

CPM: OMS Regressed on Best PSMs for Each Scale[a]

Scale	Model Statistics						Shape		Delta		Scatter		Elevation	
	R	ΔR^2	Adj ΔR^2	*F*-change	*df*	Sig	β	*r*	β	*r*	β	*r*	β	*r*
Shape Based on Consensus Keys														
Hostility to Authority	.17	.028	.027	28.02	4,3863	.001	0.09	.11	-0.03	-.12	-0.06	-.06	0.09	.12
Cognitive Flexibility	.12	.014	.014	18.34	3,3796	.001	0.10	.10	-0.05	-.05	-0.05	-.04		
Fitness Motivation	.20	.040	.039	40.71	4,3864	.001	0.19	.19	-0.01[ns]	-.04	0.04	.08	0.05	.02
Self-Efficacy	.14	.020	.019	20.09	4,3865	.001	0.13	.14	-0.10	-.02	0.00	.03	0.10	.00
Shape Based on Conventional Keys														
Achievement	.28	.079	.078	110.00	3,3874	.001	0.26	.28	-0.05	-.10	0.02[ns]	.08		
Army Identification	.08	.007	.006	6.57	4,3806	.001	0.04	.04	-0.02[ns]	.00	-0.03[ns]	-.01	0.08	.07
Peer Leadership	.16	.026	.025	25.67	4,3870	.001	0.13	.14	-0.10	-.03	0.01[ns]	.04	-0.10	-.04
Stress Tolerance	.11	.013	.012	12.32	4,3856	.001	0.07	.07	-0.20	-.05	0.01[ns]	.01	-0.18	.01
Tolerance for Ambiguity	.14	.021	.020	20.23	4,3841	.001	0.09	.08	-0.00[ns]	.07	-0.05	-.06	0.12	.09

[a]All β coefficients are significant ($p < .05$) unless noted.

[ns]Not significant

Table 24

CBEF: OMS Regressed on Best PSMs for Each Scale[a]

Scale	Model Statistics						Shape		Delta		Scatter		Elevation	
	R	ΔR^2	Adj ΔR^2	*F*-change	*df*	Sig	β	*r*	β	*r*	β	*r*	β	*r*
Shape Based on Consensus Keys														
Guilt Proneness	.13	.016	.015	21.17	3,3905	.001	0.11	.12	-0.05	-.06	-0.00[ns]	.05		
Written Communication	.31	.097	.096	104.62	4,3904	.001	0.22	.20	0.06[ns]	-.10	0.21	.21	-0.12	-.10
Fitness Motivation	.31	.095	.094	102.65	4,3897	.001	0.16	.23	-0.01[ns]	-.17	0.13	.24	0.14	.17
Shape Based on Conventional Keys														
Army Identification	.06	.004	.003	4.13	4,3897	.002	0.07	.04	-0.07[ns]	-.02	-0.04	-.02	0.10	-.01
Stress Tolerance	.16	.024	.024	32.61	3,3902	.001	0.09	.13	-0.09	-.13	-0.05	-.04		
Peer Leadership	.17	.029	.028	29.25	4,3902	.001	0.12	.15	-0.01[ns]	-.10	0.01[ns]	.07	-0.08	-.13
Past Withdrawal	.10	.010	.009	13.45	3,3902	.001	0.07	.07	-0.07	-.08	-0.01[ns]	.02		
Tolerance for Injury	.11	.013	.012	12.71	4,3863	.001	0.05	.10	0.00[ns]	-.10	0.00[ns]	.02	-0.08	-.10

[a]All β coefficients are significant ($p < .05$) unless noted.

[ns]Not significant.

Table 25

Incremental Validity for Two Models Against OMS

Regression	*R*	R^2	Adj R^2	Change Statistics			
				ΔR^2	F Change	*df*	Sig
Model 1: Distances for all scales – CBEF then CPM							
Step 1. CBEF Distance Scales	.33	.111	.109	.111	57.32	8,3690	.001
Step 2. CBEF Distance (Never PSM scored)	.45	.205	.203	.095	88.03	5,3685	.001
Step 3. CPM Distance Scales	.47	.226	.221	.020	10.64	9,3676	.001
Model 2: Mostly PSMs but Distance for scales that cannot be PSM scored – CBEF then CPM							
Step 1. CBEF PSM Scales	.43	.189	.187	.189	107.53	8,3690	.001
Step 2. CBEF Distance Scales (Never PSM scored)	.51	.262	.259	.073	72.56	5,3685	.001
Step 3. CPM PSM Scales	.54	.288	.284	.027	15.25	9,3676	.001
Model 2 Continuation: Distance scores for those scales that had been PSM scored							
Step 4. Remaining CBEF/CPM Distance Scales	.56	.313	.306	.025	7.84	17,3659	.001

Distance Regression Model. In model 1, we regressed OMS on distance scores for the 8 CBEF scales that could be PSM scored in step 1 (R = .33, p < .001). In step 2, we added distance scores for the 5 CBEF scales that could not be PSM scored (R = .45, p < .001). In step 3, we added distance scores for the 9 CPM scales (R = .47, p < .001). The results reported for model 1 provide a baseline to evaluate the use of PSMs to score these scales.

PSM Regression Model. In model 2, we regressed OMS onto PSM scores for the 8 CBEF scales that were PSM scored in step 1 (R = .43, p < .001). In step 2, we added distance scores for 5 CBEF scales that were not PSM scored because they did not contain reversed items (R = .51, p < .001). In step 3, we added PSM scores for the 9 CPM scales (R = .54, p < .001).

Model Comparison. At a broad level, the results obtained for the two hierarchical regression models were consistent. Significant gains in validity were obtained at each step for both models and the resultant validity estimates were substantial. However, scoring the 8 CBEF scales using PSMs resulted in a substantially higher validity at step 1 than the corresponding estimate using only distance scores (R = .43 vs R = .33). This result associates increases of 30% in validity for those 8 scales when scored using PSMs in place of distance metrics.

In step 2, the inclusion of distance scores for the 5 CBEF scales, which did not contain reversed items and were not PSM scored, increased the composite validity estimates for both batteries. However, the results for the PSM model continued to provide a substantial improvement over the distance model (R = .51 vs R = .45). This result associates a 13% gain in the validity of the PSM model over the distance model. In addition, the validity estimate for the PSM model can be characterized as large because it surpassed the .50 validity threshold (Cohen, 1988).

At step 3, the CPM provided incremental validity beyond the CBEF for both models. However, only the PSM model continued to surpass the .50 validity threshold (R = .54 vs R = .47). This result associates a 15% validity gain with the use of PSMs as opposed to distance metrics to score the CBEF and CPM. Therefore, the regression models demonstrated that the CPM provides useful incremental validity to the CBEF, especially when scored using PSMs. In addition, the validity estimate that was based on PSM scoring procedures, (R = .54) slightly exceeds validity estimates for general cognitive ability ($\bar{r}$ = .51; Schmidt & Hunter, 1998).

All Scale Scores. We extended the second regression model and added the distance scores for the CPM and CBEF scales that had been PSM scored. The results for this step are presented in the lowest panel of Table 25. The inclusion of these additional metrics resulted in a significant increase in incremental validity (R = .56, p < .001). This value represents the potential validity of these scales by using a highly complex scoring algorithm and suggests that refinements to the PSM model may further increment the validity of existing rating-based personality scales.

Table 26

Convergent and Divergent Validity by Construct and Response Format: CBEF vs CPM[a]

Construct	Battery by Scoring Algorithm	Scale Validity (OMS)	Convergent Validity[b] Among Measures by Construct				Divergent Validity[b] Against Other Constructs	
			CBEF		CPM			
			Dist	PSM	Dist	PSM	$\bar{r}$	SD*r*
Army Identification	CBEF-Distance	.02	1.00	.37	.66	.16	.19	0.13
	CBEF-PSM	.06	.37	1.00	.28	.24	.06	0.09
	CPM-Distance	.02	.66	.28	1.00	.26	.18	0.12
	CPM-PSM	.08	.16	.24	.26	1.00	.06	0.09
Fitness Motivation	CBEF-Distance	.27	1.00	.89	.74	.66	.17	0.14
	CBEF-PSM	.31	.89	1.00	.59	.53	.14	0.12
	CPM-Distance	.17	.74	.59	1.00	.83	.16	0.13
	CPM-PSM	.20	.66	.53	.83	1.00	.18	0.11
Peer Leadership	CBEF-Distance	.17	1.00	.95	.67	.57	.21	0.16
	CBEF-PSM	.17	.95	1.00	.63	.56	.20	0.15
	CPM-Distance	.15	.67	.63	1.00	.83	.23	0.15
	CPM-PSM	.16	.57	.56	.83	1.00	.21	0.13
Stress Tolerance	CBEF-Distance	.14	1.00	.89	.57	.38	.21	0.17
	CBEF-PSM	.16	.89	1.00	.48	.38	.19	0.14
	CPM-Distance	.06	.57	.48	1.00	.60	.18	0.17
	CPM-PSM	.11	.38	.38	.60	1.00	.13	0.09
Achievement	CBEF-Distance	.27	1.00		.39	.40	.16	0.15
	CPM-Distance	.26	.39		1.00	.91	.18	0.15
	CPM-PSM	.28	.40		.91	1.00	.21	0.13
Hostility to Authority	CBEF-Distance	.12	1.00		.29	.07	.10	0.13
	CPM-Distance	-.01	.29		1.00	.12	.11	0.15
	CPM-PSM	.17	.07		.12	1.00	.07	0.11
Self Efficacy	CBEF-Distance	.09	1.00		.50	.22	.21	0.18
	CPM-Distance	.07	.50		1.00	.28	.17	0.19
	CPM-PSM	.14	.22		.28	1.00	.16	0.10
Guilt Proneness	CBEF-Distance	.09	1.00	.78			.10	0.11
	CBEF-PSM	.13	.78	1.00			.09	0.08
Past Withdrawal	CBEF-Distance	.07	1.00	.66			.22	0.15
	CBEF-PSM	.10	.66	1.00			.17	0.11
Tolerance for Injury	CBEF-Distance	.11	1.00	.96			.17	0.14
	CBEF-PSM	.11	.96	1.00			.16	0.13
Written Communication	CBEF-Distance	.19	1.00	.58			.14	0.13
	CBEF-PSM	.31	.58	1.00			.12	0.11
Cognitive Flexibility	CPM-Distance	.02			1.00	.31	.01	0.12
	CPM-PSM	.12			.31	1.00	.03	0.10
Tolerance for Ambiguity	CPM-Distance	.06			1.00	.34	.13	0.15
	CPM-PSM	.14			.34	1.00	.08	0.08
Shame Proneness	CBEF-Distance	.03	1.00				-.20	0.15

[a]Sample size ranged from 3759 to 3908.

[b]Distance scores reflected so superior distances and PSM scores would be positively correlated.

Table 27
CPM & CBEF Validity Estimates Using PSMs and Distance Scores for Only Overlapping Constructs

Regression	*R*	R^2	Adj R^2	Change Statistics		
				F Change	*df*	Sig
CPM Scales For Common Constructs						
Distance Scores	.30	.091	.089	54.06	7,3783	.001
PSMs	.35	.119	.118	73.32	7,3783	.001
CBEF Scales For Common Constructs						
Distance Scores	.39	.154	.152	98.02	7,3783	.001
PSMs	.42	.174	.172	113.70	7,3783	.001

CPM (9-Point) by CBEF (5-Point) Convergent Validity. Table 26 (Col 3 to 7) reports PSM and distance score correlations and validities for CPM and CBEF scales organized by construct. This information generally demonstrates convergent validity among CBEF and CPM scales for corresponding constructs. While the CBEF based validities are generally higher than the CPM validities, it is noteworthy that several CPM scale validities slightly exceed the corresponding CBEF scale validities even though the CBEF scales had been extensively revised. The table also reports evidence for divergent validity by summarizing the distribution of correlations for each scale with non-corresponding construct scales (Table 26, Col 8 & 9).

Table 27 reports composite validity estimates for scales that are based on only the CPM and CBEF scales for the six overlapping content domains. Each composite is based on only distance or PSM scores. While the validity estimate for the CBEF-based composite was greater than the corresponding CPM composite, the CBEF scales had been repeatedly modified over the preceding fifteen-year period. Therefore, the results suggest that the CPM 9-point format may become competitive with the CBEF 5-point format if the scales were refined.

Project 3 Summary

The third project demonstrated that: (a) the PSM framework can be used to quickly prototype personality scales that have impressive levels of validity against performance outcomes; (b) the regression analyses provided broad support for the two foundational and three predictive validity hypotheses that were derived from the PSM framework; (c) personality scale validity estimates that were based on PSMs as opposed to distance scores increased on average by 73% for the CPM scales and 26% for the CBEF scales; and (d) optimizing PSM scores computed for two personality batteries that greatly differed in response format resulted in a very substantial level of validity against performance outcomes, $R = .54$, which exceeds the validity estimate for general cognitive ability based on meta-analyses ($\overline{r} = .51$; Schmidt & Hunter, 1998)

General Discussion

We described the PSM framework and systematically explored its implications for designing and scoring rating-based judgment tests and personality inventories. The analyses supported expectations that PSMs may provide incremental validity beyond distance scores for a broad array of judgment tests and personality scales. Because most rating-based scales have not used the PSM framework to compute scale scores, we believe that much existing data could be reanalyzed within the PSM framework to reevaluate theory and extend research findings. This approach may also be ideal to reassess the predictive potential of measures that have been associated with low levels of validity (e.g., interest inventories, Schmidt & Hunter, 1998).

PSM Hypotheses and Scale Validity

The PSM framework is based on formulaic derivations showing that distance metrics may be modelled as main effects associated with the shape, scatter, and delta metrics. More importantly, we recognized that scale validity may be improved by using regression models to optimally weight these metrics because there is no compelling reason to expect that distance metrics optimally weight these separate metrics to predict outcomes. This observation suggested that reweighting those PSMs against relevant outcome metrics might readily enhance the validity of many rating-based scales.

In addition, we provided formulaic derivations showing that the elevation metric may be used to enhance scale validity when a conventional scale key is poorly centered. We also showed that key shape may limit scale validity. Based on these observations, we hypothesized that the elevation and shape-consensus metrics may be used to further increment the validity of rating-based scales against specific criteria. Therefore, we proposed optimizing the validity of rating-based scales using the following PSMs:

- Shape, computed relative to the conventional scoring key: $r_{x,k}$;
- Scatter, respondent rating variance: sd_x^2;
- Delta, respondent rating elevation relative to the scoring key: $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$.
- Elevation, respondent rating mean: X_{mean};
- Shape-consensus, computed relative to a consensually-derived key: $r_{x,consensus}$.

We then conducted regression analyses using a variety of rating-based judgment tests and personality inventories to evaluate our hypotheses and expectations. The empirical analyses provided strong and consistent support for the PSM foundational and predictive validity hypotheses. These results confirmed our expectations that the D and D^2 metrics are nearly redundant and primarily represent variance associated with the shape, delta, and scatter metrics. In addition, PSM scoring provided incremental validity beyond distance metrics for all the judgment tests and most of the personality scales. These results support expectations that distance scores frequently represent a sub-optimized composite of shape, delta, and scatter effects.

Scale Validity

From an applied perspective, the gains in scale validity were substantial for both the judgment tests as well as the personality scales. Despite the short length and minimal administration requirements for the three judgment tests, PSM scoring resulted in modest validity estimates against supervisor performance ratings, R = .33, and career continuance intent, R = .25 (Project 1). In fact, the validity estimates for the judgment tests described in Project 1 compare favorably to meta-analytic validity estimates for SJTs ($\overline{r}$ = .26 and ρ = .34; McDaniel et al., 2001). Furthermore, we developed the judgment tests at minimal costs by leveraging psychological models as opposed to developing the scales using conventional SJT procedures (McDaniel & Nguyen, 2001).

For the personality scales, the second project demonstrated that PSM scoring resulted in gains in scale and composite validity that were largely maintained in a fully independent cross-sample using data that had been collected two years later. Encouraged by these results, we structured the third project to evaluate the potential of using PSMs to validate personality batteries that had incorporated 5-point and 9-point rating scales. Although we were confident that Project 3 analyses would provide support for the proposed hypotheses, we were impressed that PSM scoring resulted in a substantial validity estimate against OMS, R = .54. This estimate exceeds established guidelines for categorization as a "large validity" coefficient (0.50, Cohen 1988) as well as the meta-analytic estimates for the validity of *Psychometric g* (e.g., $\overline{r}$ = .51, Schmidt & Hunter, 1998).

We also emphasize that none of the validity estimates reported in this paper were corrected for range restriction or attenuation of reliability (predictor or criterion) and that item analyses were not conducted to eliminate poorly functioning items. Therefore, the above validity estimates are likely lower bounds on the range of validity estimates that could be obtained using PSM scoring procedures and revised scale items. From an applied perspective, these results support the view that past performance is a very potent predictor of future performance and redefine expectations regarding the potential validity of personality scales against an array of outcome metrics.

The results for the personality scales carry practical importance for personnel selection and classification for the U.S. Army because USACC awards ROTC scholarships to individuals who are likely to obtain high OMS scores. The regression analyses showing that personality scale validities approach validity levels commonly associated with general cognitive ability and may provide a credible basis to challenge the dominance of general cognitive ability for many personnel selection applications.

Refining Expectations for Personality Constructs

One reason we conducted these analyses is that narrowly defined personality scales frequently underperform expectations regarding their underlying constructs. For example, despite the conceptual importance of writing performance to the OMS outcome measure (Paullin et al., 2014), the CBEF Written Communication distance metric had a modest validity against OMS, r = .17. Likewise, safety awareness is conceptually important to the performance leaders

in the military as well as many civilian sectors, yet the distance metric for the CBEF Safety scale was not significantly related to the OMS outcome, $r = .01$.

These types of results are conceptually difficult to rationalize and can suppress interest in their additional investigation. However, the importance of some constructs is reaffirmed by the demonstration of much higher validities for PSM-based scores: Written Communication, $R_{PSM} = .33$ vs. $R_{distance} = .17$; Safety, $R_{PSM} = .08$ vs. $R_{distance} = .01$. In this way, PSM scoring has great potential to renew and reinvigorate interest in a broad range of constructs that are likely to be important to human performance, but for which valid measurement has been elusive.

Scale Design

The PSM framework carries broad implications for scale design that may conflict with guidance developed for distance metrics. From a PSM perspective, scales should be designed to improve the psychometric properties of the shape, elevation, and scatter metrics, as opposed to having a singular focus on distance measures. Across the three projects described in this paper, the empirical analyses demonstrated that:

- Incorporating large versus small rating scales (e.g., 9-point vs. 5-point scales) in judgment tests and personality scales improved the utility of these scales in accordance with PSM expectations.

- Attaching many options to judgment test scenarios may improve the psychometrics of these scales, while efficiently utilizing test administration resources.

- Incorporating a mix of reversed and non-reversed items within personality scales enhances the validity of personality scales through the items' impact on the underlying PSMs.

By leveraging consensually derived scoring standards, these innovations provide a potent approach that can be used to increase the breadth of domains for which accurate psychological assessment is possible.

Limitations & Response Distortion

Our analyses were based on research data, and respondents had been informed that their participation would not affect their Army careers. The use of research data may be an important consideration because an ongoing concern with the use of personality data in applied settings is the potential for respondents to use a range of response distortion strategies to improve their scores.

However, the use of PSMs to measure individual differences on these instruments may mitigates this concern because their effective use requires advanced understandings of the statistical issues that underlie the rationale for PSMs to score these scales as well as highly specific knowledge regarding the scoring algorithms for each scale. Unlike conventional personality scales for which faking instructions are relatively easy to enunciate, the relative weighting of the five metrics on which the PSM scale scores are computed would overwhelm the capabilities of most respondents to improve their scale scores. Regardless, the reanalysis of data

collected under operational conditions will likely provide insight regarding the use of PSMs in operational settings.

Another important limitation to our results is that the success of using consensual keying standards to increment scale validity suggests that there are likely more powerful approaches to optimize the shape of the scoring keys for rating-based scales. However, the PSM framework provides an excellent basis to explicitly focus on methods that may improve key shape by disentangling effects that relate to scatter, delta, and elevation.

References

Allen, M.T., & Young, M.C. (2012). *Longitudinal Validation of Non-cognitive Officer Selection Measures for the U.S. Army Officer Candidate School* (ARI Technical Report 1323). Fort Belvoir, VA; U.S. Army Research Institute for the Behavioral and Social Sciences.

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

Christensen, P.R., Merrifield, P.R., & Guilford, J.P. (1953). *Consequences Form A-I.* Beverly Hills, CA: Sheridan Supply.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Costa, P.T., & McCrae, R.R. (1991). *NEO Five-Factor Inventory Form S.* Odessa, FL: Psychological Assessments Resources, Inc.

Cox, E.P. (1980). The optimal number of response alternative for a scale: A review. *Journal of Marketing Research, 17,* 407-422.

Cullen, M.J., Sackett, P.R., & Lievens, F.P. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142-155.

Cronbach, L.J., & Gleser, G.C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456-473.

Edwards, J.R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46,* 641-665.

Goldberg, L.R. (n.d.) IPIP scale scoring instructions. Retrieved from http://ipip.ori.org/newScoringInstructions.htm.

Grant, A., & Shwartz, B. (2011). Too much of a good thing: The challenge and opportunity of the inverted U. *Perspectives on Psychological Science, 6,* 61-76.

Hogan, R. (2005). In defense of personality measurement: Old wine for new whiners. *Human Performance, 18,* 331-341.

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

Kilcullen, R.N., Gluszek, A., Legree, P.J., Repchick, K., Daza, A., & Brady, M.F. (2013, August). *New approaches to creating fake-resistant temperament measures.* Poster presented at the 121st annual conference of the American Psychological Association, Honolulu.

Kilcullen, R.N., Robbins, J., & Tremble, T. (2009). Development of the CBEF. In D.J. Putka (Ed.), *Initial Development and Validation of Assessments for Predicting Disenrollment of Four-year Scholarship Recipients from the Reserve Officer Training Corps* (ARI Study Report 2009-06). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P.J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence, 21,* 247-266.

Legree P.J., Heffner, T.S., Psotka, J., Martin, D.E., & Medsker, G.J. (2003). Traffic crash involvement: Experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology, 88,* 15-26.

Legree P.J., Kilcullen, R.N., Psotka, J., Putka, D.J., & Ginter, R.N. (2010). *Scoring Situational Judgment Tests Using Profile Similarity Metrics* (ARI Technical Report 1272). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P.J., Kilcullen, R.N., Putka, D.J., & Wasko, L.E. (2014). Identifying the leaders of tomorrow: Validating predictors of leader performance. *Military Psychology, 26,* 292-309.

Legree, P.J., Psotka, J., Robbins, J., Roberts, R.D., Putka, D.J., & Mullins, H.M. (2014). Profile similarity metrics as an alternate framework to score rating-based tests: MSCEIT reanalyses. *Intelligence, 47,* 159-174.

Legree P.J., Psotka, J., Tremble, T.R., & Bourne, D.R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulz & R.D. Roberts (Eds.), *Emotional Intelligence: An International Handbook* (pp. 155-180). Cambridge, MA: Hogrefe & Huber.

MacCann, C., Ziegler, M., Roberts, R.D. (2012). Faking in personality assessment. In M. Ziegler, C. MacCannn & R.D. Roberts (Eds.), *New Perspective on Faking in Personality Research* (pp. 309-329). Cambridge, MA: Hogrefe & Huber.

Mayer, J.D., Caruso, D.R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27,* 267-298.

Mayer, J.D., Salovey, P., Caruso, D.R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3,* 97-105.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel M.A., & Nguyen N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

McDaniel, M.A., Psotka, J., Legree, P.J., Yost, A.P., & Weekley, J.A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96,* 321-336.

McHenry, J.J., Hough, L.M., Toquam, J. L., Hanson, M.A., Ashworth, S. (1990). Project A results: The relationship between predictor and criterion domains. *Personnel Psycholo*gy, *43*, 335-354.

Motowidlo, S.J., Crook, A.E., Kell, H.J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single response situational judgment test. *Journal of Business and Psychology*, *24*, 281–287.

Muros, J.P. (2008). *Know the Score: An Exploration of Keying and Scoring Approaches for Situational Judgment Tests.* (Unpublished Doctoral Dissertation). University of Minnesota, Minneapolis, MN.

Nunnally, J.C. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Paullin, C., Legree, P.J., Sinclair, A.L., Moriarty, K.O., Campbell, R.C., & Kilcullen, R.N. (2014). Delineating officer performance and its determinants. *Military Psychology, 26,* 259-277.

Polanyi, M. (1966). *The Tacit Dimension*. New York: Doubleday.

Preston, C.C., Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Reeder, M.C., & Ryan, A.M. (2012). Methods for correcting faking. In M. Ziegler, C. MacCannn & R.D. Roberts (Eds.), *New Perspective on Faking in Personality Research* (pp. 131-150). Cambridge, MA: Hogrefe & Huber.

Russell, T.L., Paullin, C.P., Legree, P.J., Kilcullen, R.N., & Young, M.C. (2017). *Identifying and Validating Selection Tools for Predicting Officer Performance and Retention* (ARI Research Note 2017-01). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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 finds. *Psychological Bulletin, 124*, 262-274.

Schirmer, P. (2016). *Challenging Time in DOPMA Flexible and Contemporary Military Officer Management*. Arlington, VA: Rand Corporation.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87,* 245-251.

Sternberg, R.J. (2006). The rainbow project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34,* 321–350.

Sternberg, R.J., & Hedlund, J. (2002). Practical intelligence, *g,* and work psychology. *Human Performance, 15,* 143-160.

Sternberg, R.J., & Wagner, R.K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2,* 1-5.

Stevens, S.S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects.* Oxford, England: John Wiley & Sons.

Tangney, J.P., Baumeister, R.F., & Boone, A.L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72,* 271–322.

Wagner, R.K., & Sternberg, R.J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458.

Weng, Q., Yang, H., Lievens, P., McDaniel, M.A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior, 104,* 190-209.

Yukl, G. (2002). *Leadership in Organizations* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Zaccaro S.J., Connelly S., Repchick, K.M., Daza, A.I., Young M.C., Kilcullen R.N., Gilrane, V.L., Robbins, J.M., & Bartholomew L.N. (2015). The influence of higher order cognitive capacities on leader organizational continuance and retention: The mediating role of developmental experiences. *Leadership Quarterly, 26,* 342-358.

Appendix A. Equivalence of Conventional and Distance Metrics

Table A-1 lists all possible conventional and distance scores for 5-point reversed and non-reversed items that are used for scales such as the NEO Five-Factor Inventory Form S (Costa & McCrae, 1991), any International Personality Item Pool scale (Goldberg, n.d.), and the Cadet Background and Experiences Form (CBEF, see Project 2). The table illustrates that conventional and distance item scores will always add to a specific constant (e.g., "5" for the CBEF and IPIP scales, "4" for the NEO Five-Factor Inventory Form S scale). This relationship implies that the correlation between the conventional and distance item and score metrics is perfect, $r = -1$.

Table A-1
Conventional and Distance Scoring Algorithms.

Items Form	Conventional Algorithm		Distance Algorithm			Sum
	Respondent Rating	Score	Respondent Rating	Key	Score	
Non-reversed Items						
	1	1	1	5	4	5
	2	2	2	5	3	5
	3	3	3	5	2	5
	4	4	4	5	1	5
	5	5	5	5	0	5
Reversed Items						
	1	5	1	1	0	5
	2	4	2	1	1	5
	3	3	3	1	2	5
	4	2	4	1	3	5
	5	1	5	1	4	5

Appendix B. PSM Derivations from D^2 Formula

To decompose the D^2 metric, each respondent's set of ratings for a judgment test should be conceptualized as a rating profile vector, **X**, with *n* elements (i.e., item ratings). Likewise, the scoring key should be represented as a scoring profile vector, **K,** also with *n* elements (i.e., X_i and K_i correspond to ratings obtained from an individual and the scoring key for item *i*). The D^2 metric is then calculated as the mean squared difference between elements in the two arrays:

$$D^2 = \sum_{i=1}^{n} (X_i - K_i)^2 / n \quad (1)$$

Substituting $X_i = x_i + X_{mean}$ and $K_i = k_i + K_{mean}$, and simplifying the formula, isolates the squared difference in elevation between the respondent and scoring profiles in D^2, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$:

$$D^2 = \Delta_{Key}{}^2 + \sum_{i=1}^{n} (X_i - K_i)^2 / n,$$

Derivations are detailed in Table B-1.

The delta term in Equation B-2 (i.e., Δ_{Key}^2) shows that distance metrics will penalize any respondent whose rating profile is poorly centered (i.e., elevated or depressed) in comparison to the scoring key. Expanding Equation B-2 and invoking statistical terminology and substitutions, [$\sum x_i^2 = sd_x^2(n\text{-}1)$; $\sum k_i^2 = sd_k^2(n\text{-}1)$; $x_i = z_{xi}sd_x$; $k_i = z_{ki}sd_k$; and $\sum z_{xi}z_{ki} = r_{x,k}(n\text{-}1)$], decomposes D^2 into separate terms that quantify individual differences in rating profile scatter, and shape relative to the scoring key (i.e., sd_x^2, and $r_{x,k}$):

$$D^2 = \Delta_{Key}{}^2 + \frac{(n-1)(sd_x{}^2 + sd_k{}^2 - 2sd_x sd_k r_{xk})}{n} \quad (2)$$

Equation 3 shows that D^2 is directly related to the delta and correlation terms (i.e., Δ_{Key}^2 and $r_{x,k}$), but has a quadratic relationship with the variance term, sd_x^2. Moreover, D^2 formulaically reflects only individual differences among the elevation, scatter, and shape of each respondent rating profile, **X**, and the scoring profile, **K**. It can also be shown that respondent rating scatter (i.e., variance or sd_x^2), must be minimized as a function of the shape of an individual rating profile, $r_{x,k}$, to obtain an optimal D^2 score. Formulaically, this value can be computed for any shape score by differentiating D^2 with respect to sd_x, and solving the minimum value:

$$d(D^2)/d(sd_x) = (n - 1)(2sd_x - 2sd_k r_{x,k})/n,$$

And solving for sd_x (i.e., the square root of the scatter term):

$$sd_x = sd_k r_{x,k}.$$

This equation shows that a distance metric will penalize a respondent if the individual's rating profile contains either excessive or restricted levels of scatter, with the optimal level of scatter dependent on the shape of the individual response profile, $r_{x,k}$., and the scatter of the keying profile, sd_k. This derivative also provides the formulaic basis for understanding demonstrations that low-scoring respondents may improve conventional distance scores for many rating-based SJTs by avoiding extreme responses (Cullen et al., 2006)**.**

Table B-1
PSM Derivations

$D^2 = \sum(X_i - K_i)^2/n$ for item i = 1 to n	Distance squared formula (Equation 1 in text)
$= \sum((x_i + X_{mean}) - (k_i + K_{mean}))^2/n$	Substitutions center X and K: $x_i = X_i - X_{mean}$ thus $X_i = x_i + X_{mean}$; $k_i = K_i - K_{mean}$ thus $K_i = k_i + K_{mean}$
$= \sum(x_i + X_{mean} - k_i - K_{mean})^2/n$	Distributive Property
$= \sum(x_i - k_i + (X_{mean} - K_{mean}))^2/n$	Rearrange and group
$= \sum(x_i - k_i + \Delta)^2/n$	Substituting Δ for $X_{mean} - K_{mean}$
$= 1/n\sum(x_i - k_i + \Delta)^2$	Constant multiplication property of sums
$= 1/n\sum(x_i^2 + k_i^2 + \Delta^2 - 2x_ik_i + 2x_i\Delta - 2k_i\Delta)$	Binomial expansion
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + \sum 2x_i\Delta - \sum 2k_i\Delta)$	Expansion property of sums
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 2\Delta\sum x_i - 2\Delta\sum k_i)$	Constant multiplication property of sums
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 2\Delta 0 - 2\Delta 0)$	$\sum x_i = 0$ & $\sum k_i = 0$ because x & k are centered
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 0 - 0)$	Multiplication property of zero
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i)$	Additive property of zero
$= 1/n\sum\Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - \sum 2x_ik_i)$	Regrouping property of sums
$= \Delta^2 + 1/n\sum(x_i^2 + k_i^2 - 2x_ik_i)$	Summation of a constant: Substitutes $1/n\sum\Delta^2 = 1/n(n\Delta^2) = \Delta^2$
$= \Delta^2 + 1/n\sum(x_i - k_i)^2$	Provides binomial solution (Equation 2 in text)
$= \Delta^2 + 1/n\sum(x_i^2 + k_i^2 - 2x_ik_i)$	From two steps above
$= \Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - \sum 2x_ik_i)$	Expansion property of sums
$= \Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - 2\sum x_ik_i)$	Constant multiplication property of sums
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2\sum x_ik_i)$	Substitutions based on statistical formulas re variance: $\sum x_i^2 = sd_x^2(n\text{-}1)$ & $\sum k_i^2 = sd_k^2(n\text{-}1)$
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2\sum z_{xi}sd_xz_{ki}sd_k)$	Substitutions based on statistical formulas re z-scores: $x_i = z_{xi}sd_{xi}$ & $k_i = z_{ki}sd_k$
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2sd_xsd_k\sum z_{xi}z_{ki})$	Constant multiplication property of sums
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2sd_xsd_kr(n-1))$	Substitutions based on formulas re the product moment correlation: $r = \sum z_{xi}z_{ki}/(n - 1)$ thus $\sum z_{xi}z_{ki} = r(n - 1)$
$= \Delta^2 + (n-1)/n(sd_x^2 + sd_k^2 - 2sd_xsd_kr)$	Rearrangement of terms

Appendix C. Using PSMs to Adjust Key Elevation and Optimize Validity

Equation 2 shows that delta term, $\Delta_{Key}^2 = (X_{mean} - K_{mean})^2$, is central to understanding distance scores computed using a conventional key. However, the delta term cannot optimize scale validity if the scoring key is poorly centered (i.e., the delta term would be computed using the value, K_{mean}, and not be computed using the optimal value, K_{opt}). To adjust the delta term, we define delta-optimal and K-optimal as follows:

$$\Delta_{opt}^2 = (X_{mean} - K_{opt})^2 \text{ with } K_{opt} = K_{mean} + A \text{ and } \Delta_{Key}^2 = (X_{mean} - K_{mean})^2. \quad (3)$$

$$= X_{mean}^2 + K_{opt}^2 - 2X_{mean}K_{opt}$$

$$= X_{mean}^2 + (K_{mean} + A)^2 - 2X_{mean}(K_{mean} + A)$$

$$= X_{mean}^2 + K_{mean}^2 + A^2 + 2AK_{mean} - 2X_{mean}K_{mean} - 2AX_{mean}$$

$$= (X_{mean}^2 + K_{mean}^2 - 2X_{mean}K_{mean}) + A^2 + 2AK_{mean} - 2AX_{mean}$$

$$= (X_{mean} - K_{mean})^2 + A^2 + 2AK_{mean} - 2AX_{mean}$$

$$= \Delta_{Key}^2 + A^2 + 2AK_{mean} - 2AX_{mean} \quad (4)$$

Equation C-1 shows that the delta-optimal term represents a linear combination of the conventional delta and elevation terms: $R_{\Delta Opt2,\Delta Key2.Xmean} = 1.00$ (i.e., A and K_{mean} are constant across individuals for a specific sample). This observation implies that the elevation term may provide incremental validity beyond the conventional delta term when the key is poorly centered. The delta-optimal term, Δ_{Opt}^2, can be directly computed once regression analyses are used to identify and optimize the elevation and delta terms.