Research Note 2022-04

Identifying Keys and Scoring Algorithms to Enhance Personality Scale Validity

Peter J. Legree
U.S. Army Research Institute

Zach Traylor
Texas A&M University

Oren R. Shewach
Human Resources Research Organization

Benjamin S. Kerner
George Mason University

December 2021

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

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

Human Resources Research Organization

Technical Review by:

Jessica R. Carre, U.S. Army Research Institute
Melissa J. Glorioso, U.S. Army Research Institute

DISTRIBUTION:

This Research Note has been submitted to the
Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
December 2021	Interim	June 2020 to December 2021

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Identifying Keys and Scoring Algorithms to Enhance Personality Scale Validity	W911NF-19-C-0065
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) Peter J. Legree, Zach Traylor, Oren R. Shewach, and Benjamin S. Kerner	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 1011
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Suite 700 Alexandria, Virginia 22314	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg 1464/Mail Stop 5610) Ft. Belvoir, VA 22060-5586	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 2022-04

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

ARI Research POC: Peter J. Legree, Selection and Assignment Research Unit

14. ABSTRACT *(Maximum 200 words)*:

The U.S. Army Cadet Command uses personality scales that have been validated against the Cadet Order of Merit Score, to help award U.S. Army Reserve Officer Training Corps scholarships. To improve the utility of these measures, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is evaluating alternate procedures that can be used to key and compute scale scores for these measures. Analyses were conducted using a cross-validation design to identify near-optimal scoring keys to validate shape and dot product scores for twenty-one personality scales against the Cadet Order of Merit Score. Subsequent regression analyses compared the efficacy of using dot product scores, profile similarity metrics, and distance scores for individual scales and the battery. Scale results show substantial validity gains for most scales when using the more sophisticated scoring approaches. Battery validity estimates were high for profile similarity metrics scores and dot product scores adjusted for respondent scatter and elevation effects, while modest results were obtained for conventional distance scores (R_{psm} = .61 & R_{dot} = .61 vs R_{dist} = .45).

15. SUBJECT TERMS

Personnel selection, profile similarity metrics, dot product scores

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	43	Tonia S. Heffner 703-545-4408

Standard Form 298

Research Note 2022-04

Identifying Keys and Scoring Algorithms to Enhance Personality Scale Validity

Peter J. Legree
U.S. Army Research Institute

Zach Traylor
Texas A&M University

Oren R. Shewach
Human Resources Research Organization

Benjamin S. Kerner
George Mason University

Selection and Assignment Research Unit
Tonia S. Heffner, Chief

December 2021

IDENTIFYING KEYS AND SCORING ALGORITHMS TO ENHANCE PERSONALITY SCALE VALIDITY

EXECUTIVE SUMMARY

Research Requirement:

Personality scales are widely used in the military to predict performance because they have modest predictive validity, yet minimal adverse impact. Most rating-based personality scales are scored using a distance algorithm and a dominance key that reflects extreme responses for all items. To enhance the utility of these personality instruments, validation analyses using sophisticated keying and scoring algorithms were compared to results obtained using distance scores and dominance keys.

Procedure:

While rating-based personality scales are usually scored using distance metrics, researchers have evaluated the potential of more sophisticated algorithms to enhance scale validity. We compared validity estimates computed against the USACC Cadet Order of Merit Score (OMS) for 21 personality scales to evaluate the following keying and scoring methods for rating-based personality scales:

- Baseline Method: We used dominance keys composed of extreme values to compute distance scores for rating-based personality scales. This approach does not require a cross-validation design, but results depend on the quality of these keys.
- Optimal Integer Key Method: We specified all possible integer based keys to define the "nteger key space" for each scale. Each candidate keying vector was used to compute profile similarity metric (PSM) shape scores. This approach carries several limitations:
    - The method is computationally possible for only shorter personality scales.
    - Shape scores cannot be computed for those participants who select the same response for each scale item (i.e. response vectors lack variance; scatter = 0).
    - A cross-validation design is required to assess validity results for the generated keying vectors.
- Random Key Method: We generated 10,000 candidate keying vectors for each scale. Each candidate vector contained randomly generated values thereby defining a "random vector key space" for each scale. This approach can be applied to both long and short personality scales. Similar to the Optimal Integer Method:
    - Shape scores cannot be computed for individuals whose rating vectors did not vary over the scale items and lack variance (scatter = 0).
    - A cross-validation design is required to assess validity results.
- Item-Criterion Method: We developed item criterion scoring vectors based on the item-criterion correlations for each item. These keys were used to compute dot product and PSM shape scores. Advantages and limitations:
    - Dot product scores can be computed for all individuals including participants whose rating vectors lacked variance (scatter = 0).
    - A cross-validation design is required to assess validity results.

Correlation and regression analyses evaluated the validity of scale scores that were computed using the aforementioned keying vectors. These scale scores were computed as dot product scores, distance scores, and PSMs. We estimated validity coefficients at the scale and battery level using the above mentioned keying approaches and scoring metrics/algorithms.

Findings:

Preliminary correlational analyses using five short personality scales demonstrated near equivalent validities for shape scores based on keying vectors derived from the Optimal Integer and Random Vector procedures, and computed as item criterion vectors. Furthermore, the validity estimates for these shape scores generally exceeded those based on conventional distance scores across the five scales. These gains in scale validity were consistent: $\bar{r}_{dist} = .14$ versus $\bar{r}_{shape\text{-}oi} = .18$, $\bar{r}_{shape\text{-}rv} = .19$, and $\bar{r}_{shape\text{-}ic} = .18$.

Additional correlational analyses for 21 long and short personality scales demonstrated equivalent validities for: shape scores based on the random vector keys; shape scores based on the item criterion vectors; and dot product scores computed using the item-criterion vectors. Furthermore, the validity estimates for the shape metrics and the dot product scores generally exceeded those based on conventional distance scores for these scales. The gains in mean scale validity using the alternate procedures over distance scoring were similar to gains obtained for the five short personality scales: $\bar{r}_{dist} = .12$ versus $\bar{r}_{dot} = .17$, $\bar{r}_{shape\text{-}ic} = .18$, and $\bar{r}_{shape\text{-}rv} = .17$.

Subsequent regression analyses corrected the scale dot product scores for scatter and elevation effects. The dot product scores were selected for this purpose because they can be computed for respondents whose rating profiles lack variance (i.e., rating scatter = 0). These regression analyses demonstrated that adjusting dot product scores using the elevation and scatter metrics further enhanced the mean validity of these scales, $\overline{R}_{dot\text{-}adjusted} = .19$. This gain represents a 72% improvement in mean scale validity over the use of distance scores.

Regression procedures demonstrated that battery validity estimates were greatest when summary scores were either based on the PSM shape, scatter and elevation metrics, or were computed by adjusting the dot products scores using the elevation and scatter metrics. These validity estimates were substantial: $R_{psm} = .61$; and $R_{dot.adj} = .61$. In contrast, more moderate validity estimates were computed using only conventional distance scores, $R_{dist} = .45$.

Utilization and Dissemination of Findings:

Results support computing scale scores as dot product scores that are adjusted using scatter and elevation metrics. While equivalent validity gains can be obtained by using PSMs, the PSM approach is not suitable for operational use because shape scores cannot be calculated for individuals whose rating profiles lack variance. Guidance is provided to construct conventional personality batteries for planned research activities.

IDENTIFYING KEYS AND SCORING ALGORITHMS TO ENHANCE PERSONALITY SCALE VALIDITY

CONTENTS

Page

TABLES

Identifying Keys and Scoring Algorithms to Enhance Personality Scale Validity

Introduction

Personality scales are widely used to predict performance in military and civilian settings because they have modest predictive validity, yet minimal adverse impact (Hogan, 2005; Hough & Oswald, 2000; Ones & Anderson, 2002; Putka, 2009; Schmidt & Hunter, 1998). Most personality scales use a conventional scoring algorithm that computes scale scores as the mean item rating with some ratings reversed for item direction (e.g., Goldberg, n.d.; Likert, 1932; Putka, 2009).

To improve the validity of these personality scales, we evaluated various methods to create scale keys, as well as alternate procedures to calculate scale scores using the newly created keys. To conduct these analyses, we used the USACC Cadet Order of Merit Score (OMS) as the principal criterion for these analyses. The Cadet OMS is ideal for this purpose because it predicts long-term officer performance outcomes and is important to the USACC branch assignment algorithm (Legree, Purl, Kilcullen & Young, 2019). In addition, the Cadet OMS has been used to validate personality scales to award ROTC scholarships to individuals who are likely to perform well in pre-commissioning programs (Graves, Green & Young, 2021).

Profile Similarity Metrics, Distance Scores and Related Algorithms

From a formulaic perspective, conventional personality scores are redundant (r = -1) with distance scores that are computed using a conventional key composed of extreme values (i.e., "1" for reversed items and "5" for non-reversed items; Legree, Kerner & Shewach, 2021). The equivalence of conventional and distance scores for personality scales is important because the distance squared formula can be algebraically transformed to identify profile similarity metrics (PSMs). To compute PSMs, each set of respondent raw ratings is conceptualized as a "rating vector," and the following PSMs can be easily computed (Legree, Ness, Kilcullen & Koch, 2018):

PSM 1. *Shape*, the correlation between each respondent's rating vector, **X**, and the scale's keying vector, **K**: Shape = $r_{x,k}$.
PSM 2. *Scatter*, the variance of each respondent's rating profile: Scatter = sd_x^2.
PSM 3. *Elevation*, the respondent's mean item rating: Elevation = $X_{mean.}$

It is important to recognize that the shape metric is the only PSM that requires a scoring key for its computation. However, shape scores cannot be computed for respondent rating vectors that lack variance. This issue will occur when a respondent assigns the identical rating to all scale items. Nevertheless, the use of shape scores allows researchers to focus on the identification and evaluation of alternate methods designed to improve the keying process (Legree et al., 2018; Legree et al., 2021). In contrast, the scatter and elevation metrics are computed as descriptive statistics, are not dependent upon the scoring key, and can be computed for all respondent rating vectors.

Previous analyses have demonstrated that PSMs account for nearly all distance score variance and can be reweighted to increase the predictive validity of personality scales against valued criteria. Moreover, these validity gains are stable and substantial (Legree et al., 2018)

Analyses have also shown that these metrics can be applied to understand sources of variance underlying alternate scoring algorithms for rating-based scales and used to improve the scale psychometrics. Specifically, PSMs have been used to model variance underlying the "Proportion Correct" scores that are routinely computed for the Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT; Legree et., 2014). Proportion Correct scores are computed as endorsement ratios obtained from a keying sample (i.e., Proportion Correct scores are not transformed Percent Correct scores). Those results demonstrated that PSMs account for nearly all MSCEIT proportion correct score variance (i.e., Multiple *R*'s for the six MSCEIT rating-based scales ranged from .94 to .99; Legree et al., 2014). In addition, those analyses showed that the MSCEIT branch scores are confounded by scatter and elevation effects. By controlling these effects, subsequent analyses demonstrated that the MSCEIT is highly *g* loaded. Collectively, these results show that PSMs can provide insight into the sources of variance underlying scoring algorithms for rating-based scales, even when those algorithms are not explicitly based on a distance formula.

PSMs and Optimal Integer Vector Keys

Previous PSM analyses have also demonstrated that much higher scale validity estimates can be obtained by using a cross validation design to search the "Integer Keying Space" for each scale and identify the optimal integer vector to compute PSM shape scores (Legree et al., 2021). This process is labor intensive and required:

1. Specifying all possible integer response vectors as candidate keys for each personality scale. This space is referred to as the Integer Keying Space.
2. Computing shape scores for each possible candidate keying vector for training samples.
3. Identifying the optimal integer vector from a training sample and cross-validating the shape scores using the identified vector as the key.

A double cross-validation design was incorporated into this process so that all the analyses could be conducted twice. Specifically, the full sample was divided into two subsamples. Each subsample was used as a training sample and then served as a validation sample for the other sample (e.g., Sample 1 corresponded to Training Sample #1 and Validation Sample #2). This design also allowed quantifying the stability of the validity estimates across the candidate keying vectors computed for the two training samples.

Subsequent regression analyses showed that the validity for four of the five personality scales was improved by using the optimal integer keys to compute PSM scale scores (Legree et al., 2021). Moreover, regression analyses demonstrated that the composite validity of this five-scale personality battery increased substantially and significantly by using PSMs as opposed to conventional distance scores: $R_{PSM} = .48$ vs $R_{distance} = .32$.

In theory, this process could be applied to identify integer keying vectors for any rating-based scale. However, the size of the Integer Keying Space (i.e., the number of possible integer-based vectors) increases exponentially with the size of the rating response scale and the length of the scale. Specifically, the size of the key space is provided by: $o^i - i$, where o equals the number of rating options and i equals the number of scale items.

Therefore, the Integer Keying Space for a relatively short personality scale can be exhaustively searched. For example, a seven-item personality scale that utilizes a five-point response scale yields a key space containing the 78,120 possible integer keys (i.e., $78{,}120 = 5^7 - 5$)[1]. This procedure can be followed to identify the optimal integer key for shorter personality scales because evaluating 78,120 candidate keys represents a manageable number of calculations.

However, the size of the Integer Keying Space becomes unmanageable for scales that contain many items and incorporate large rating response scale. For example, and as detailed below, we collected data for an 18-item scale using nine rating points. The Integer Keying Space for that 18 item scale is enormous: $9^{18} = 1.5 \times 10^{17}$ candidate keys. Therefore, this approach was not feasible for all of our longer scales, and we used this keying approach only for short personality scales containing up to nine items and incorporating a five-point Likert rating scale.

PSMs and the Random Vector Keys

As previously proposed, the optimal integer keying approach may be modified to identify superior keying vectors for large rating-based personality scales (cf. Legree et al., 2021). This approach would require:

1. Generating 10,00 potential keying vectors for each scale so that each potential keying vector contained a set of uniformly-distributed random numbers (i.e., real numbers). This approach sets the size of Random Vector Keying Space to 10,000 for each scale regardless of its scale length.
2. Computing shape scores for each possible candidate keying vector for each training sample.
3. Identifying the optimal random vector key from each training sample and cross-validating the identified keys using the hold-out sample.

Similar to the Optimal Integer keying approach, the proposed Random Vector method should utilize a double cross-validation design in order to efficiently use the full sample and evaluate the stability of the shape score validity estimates for the 10,000 keying vectors across the two training samples. However, this method cannot guarantee the identification of the "best" possible scoring key in the Random Vector space, which is infinite in size. Nevertheless, the use of 10,000 random number vectors helps ensure that very "good" keys are identified for most personality scales.

PSMs and Dot Product Scores

One of the more interesting empirical scoring calculi for rating-based personality scores has been described as "Item-Level Correlation Scores" (cf. Cucina et al., 2019). These scale scores are computed by weighting respondent item ratings by their corresponding item criterion correlations and then averaging those item scores to compute scale scores. To provide clarity and using conventional algebraic terminology, these scores correspond to dot product (or scaler) scores computed between vectors containing respondent ratings and the item criterion

[1] Keying vectors must contain variance to compute shape scores. Therefore, the 5 integer vectors lacking variance were not included in the analyses.

correlations. We emphasize that unlike shape scores, dot product scores can be computed for rating response vectors lacking variance, (i.e., scatter = 0).

It is also important that empirical analyses have demonstrated that the validities of dot product scale scores are often superior to validities computed using other empirical scoring calculi (Cucina et al., 2019). Much like distance and proportion correct scores, this pattern of results suggests that dot product scores can be influenced by scatter and elevation effects in ways that can limit or enhance scale validity (Legree et al., 2018; Legree et al., 2024).

This possibility appears likely because high levels of respondent scatter will result in more extreme dot product scores when scales contain a near even mix of reversed and non-reversed items, while extreme levels of respondent elevation will result in extreme dot product scores for scales containing a very uneven mix of reversed and non-reversed items. It follows that respondent rating scatter and elevation tendencies may compromise or enhance scale validity when dot product scores are computed.

This reasoning suggests that much of the variance underlying dot product scores can be modelled by using PSMs, albeit with shape scores computed using keying vectors defined by the item criterion vectors that were used to compute the dot product scores. Therefore, we computed and incorporated dot product scores in our analyses. Similar to results for distance scores, we reasoned that scale validity estimates for dot product scores could be improved by directly incorporating scatter and elevation metrics using regression models. The shape metric was also used in some of our analyses, but would be less suitable for operational applications because some respondents may provide rating vectors that lack variance. However, we caution the utility of this method rests on the untested assumption that item criterion vectors represent a viable source for high quality keys.

Research Approach

We designed and conducted a series of analyses to evaluate expectations that are based on the aforementioned theory and results. As detailed below:

1. For five short personality scales, we evaluated the convergence of validity estimates for shape scores that were computed using conventional, optimal integer, random vector, and item criterion keying vectors.
2. For 21 personality scales, we evaluated propositions that:
    a. Dot product score variance can be accurately modelled using the PSM scatter, elevation, and shape metrics.
    b. Personality scale validity estimates can be improved by using regression procedures to optimally weight variance associated with the dot product, scatter, and elevation metrics.
3. For the 21 scale personality battery, we estimated the validity of overall battery scores using the various scoring algorithms and keying procedures that are described above. We also conducted analyses to identify core personality scales that should be included in shorter batteries to maintain validity results.

Method

Participants

Personality data were collected from 3,909 ROTC cadets who volunteered to participate in a research project during the summer of 2016. The research sample was predominantly male, 77%. These participants self-identified as: Caucasian, 80%; African-American, 12%; Asian, 8%; American Indian or Alaskan Native, 2%; and Hawaiian or Pacific Islander, 1%. Approximately, 13% of the sample identified as Hispanic. These data have been previously analyzed (Legree et al., 2018; Legree et al., 2021).

Procedure

The ROTC cadets were administered the personality scales during their initial week at the annual USACC Field Training Exercise. The criterion data (i.e., OMS) were collected from the USACC after the cadets had completed the ROTC pre-commissioning program.

Measures

Order of Merit Score (OMS)

Order of Merit Scores (OMS) were provided by the U.S. Army and served as the criterion measure for these analyses. The OMS is intended to quantify overall cadet performance in college courses and military exercises. OMS is critical to predict because it is used by the U.S. Army to help assign cadets to occupations (Legree, Kilcullen, Putka, & Wasko, 2014). In addition, OMS has been validated against in-unit officer performance metrics using a longitudinal design (Legree et al., 2019), and it has been used to validate personality measures that are used to award ROTC on-campus scholarships (Graves et al., 2021).

Personality Scales

We used data collected for two personality batteries to assess of our expectations and hypotheses. These personality batteries differed primarily in their use of either a 9-point or a 5-point Likert response scale. In addition, each battery contained scales for at least 9 of the 14 constructs listed in Table 1.

Cadet Background and Experiences Form (CBEF – 5 Point Rating Scales). The CBEF was used as the established personality battery (Kilcullen et al., 2009). It contained 104 items that were distributed over 12 scales with 9 scales containing a mix of reversed and non-reversed items, as well as 3 unidirectional scales that did not contain any reversed items. The CBEF scales are listed by construct in Table 1, and example items are provided in Table 2.

Continuum Personality Measure (CPM – 9 Point Rating Scales). The CPM contained 107 items, and each item required respondents to rate the extent to which two opposing statements describe their behaviors and experiences using a 9-point rating scale. The CPM scales are also listed in Table 1, and example items are provided in Table 2.

Table 1. *Personality Constructs, Scale Coverage, Item Number, and Definitions*[a]

Construct	Scale	CPEF	CPM	Definition
Achievement Orientation	Ach	12	18	The willingness to give one's best effort and to work hard towards achieving difficult objectives.
Army Identification	AI	11	5	Degree of identification, and interest in being, a U.S. Army Soldier.
Cognitive Flexibility	CF	-	6	Willingness to entertain new approaches to solving problems. Enjoys creating new plans and ideas. Initiates and accepts change and innovation.
Fitness Motivation	FM	7	12	Degree of enjoyment from physical exercise and willingness to stay physically fit.
Guilt Proneness	GP	9	-	Tendency to experience negative feelings regarding one's actions involving specific wrong or foolish behaviors.
General Self Efficacy	GS	6	11	Feeling that one has successfully overcome past work obstacles.
Hostility to Authority	HA	9	13	Suspicious of the motives and actions of legitimate authority figures. Views rules and directives from authority as illegitimate.
Injury Tolerance	IT	5	-	Degree of enjoyment from risky and hazardous activities
Peer Leadership	PL	9	13	Seeks positions of authority. Comfortable with being in charge of a group and accepts responsibility for the group's performance.
Past Withdrawal	PW	8	-	Degree of commitment and continuance in groups
Stress Tolerance	ST	11	13	Degree of emotional control and composure under pressure.
Shame Proneness	SP	10	-	Tendency to make global attributions regarding one's self that lead to negative feelings about the global self.
Tolerance for Ambiguity	TA	-	16	Ability to tolerate work situations where the right goal or the correct path to the goal is vague and ill-defined.
Written Communication	WC	7	-	Degree of comfort with written communication

[a]Definitions retrieved from Legree, Ness, Kilcullen & Koch (2018).

Table 2. *Personality Scale Example Items*

CBEF

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

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.

Design

Overview. We used a double cross-validation design to conduct our analyses to preserve sample size. To assess our expectations, we scored both batteries using dot product (scaler) scores, PSMs and distance metrics. We did not conduct item analyses to eliminate poorly performing items. To conduct the cross-validation design, the full sample was randomly divided into Sample 1 and Sample 2. We subsequently ran all analyses twice by using: Sample 1 as the training sample and the Sample 2 as the validation sample; and reversing the roles of these samples to repeat the training and validation analyses. Therefore, the scoring vectors used to calculate the dot product and shape metrics for each validation sample were computed using the vectors identified in the other training sample.

Optimal Integer and Random Vector Key Comparisons:

We used the five shorter personality scales to compare the validity estimates for shape scores based on the conventional keys as well as keys derived using the Optimal Integer, Random Vector, and Item Criterion procedures. This approach allowed the comparison of validity estimates for shape scores that were computed using the conventional keys as a baseline to evaluate gains in validity for shape scores computed using the more complex keying procedures. We focused on the validities of shape scores because they are most relevant to optimizing scale validity from a PSM perspective (i.e., shape scores are the only PSM that reference the scoring key).

It is important to recognize that the Optimal Integer and Random Vector methods cannot be expected to identify the best keys. This cautionary statement reflects the expectation that the best key for these scales will contain decimal values (i.e., real numbers). Therefore, the optimal integer key will not perform as well as the best real number key. In addition, the random vector keying approach cannot exhaustively search the real number space, which is infinite in size. Nevertheless, the Optimal Integer and Random Vector procedures can be expected to identify "good keys" for most scales given the semi-exhaustive nature of these approaches.

Comparisons of the Random Vector and Item Criterion Keys

We used data for all 21 personality scales to estimate the stability and magnitudes of shape score validity estimates obtained using the Random Vector method. For these analyses, we compared validity estimates for the following measures:

- Shape scores based on the random vector keys;
- Shape scores based on the item criterion keys;
- Dot product scores computed using the item criterion keys;
- Distance scores computed using the conventional keys.

These analyses were primarily structured to assess the validity of shape scores based on the random vector and item criterion keys, and to compare results of dot product scores and shape scores computed using the item criterion keys.

Scale and Battery Validity Estimates

Our analyses were broadly structured to evaluate alternate scoring and keying procedures that could be used improve the validity of rating-based personality scales. We were also focused on identifying methods that could be easily implemented, used to compute scores for all respondent rating vectors, while also ensuring near optimal validity results (e.g., adjusting dot product scores using the scatter and elevation metrics).

At the scale level, we evaluated the use of PSMs to model the variance underlying dot product scores. We also modelled potential validity gains by using regression procedures to simultaneously weight the dot product, elevation and scatter metrics. We excluded shape scores from these analyses to preserves sample size because shape scores cannot be computed for participants whose rating vectors lack scatter.

At the battery level, we compared validity estimates based on: shape scores computed using the item criterion vectors adjusted for scatter and elevation effects; dot product scores adjusted for scatter and elevation effects; and distance scores adjusted for scatter and elevation effects. We also conducted analyses for the separate CBEF and CPM sub-batteries that incorporated either 5-point or 9-point rating response formats. Finally, we conducted analyses to identify a small set of scales for inclusion in future research projects.

At a more focused level, we suspected that the use of these alternate procedures would enhance the validity of most rating-based personality scales and that the validity estimates for scales adopting the 9-point response format would more closely match results obtained for scales using the standard 5-point format.

Results

Our keying analyses utilized a cross-validation design to identify candidate keys and associated scoring algorithms that could be used to improve the validity of personality scales. We then conducted analyses for each personality scale to assess the potential of alternate keying vectors to improve scale validity. We were principally interested in comparing validity estimates that were based on: conventional distance scores; dot product scores that were adjusted for only elevation and scatter effects; and scores based on the shape, elevation and scatter metrics. Finally, we conducted analyses for the personality batteries to evaluate the overall utility of these various scoring algorithms and keying procedures.

Keying Analyses

Preliminary keying analyses were structured to assess the capacity and equivalence of using the Random Vector, Optimal Integer, and Item Criterion procedures to identify candidate keying vectors for the computation of shape scores and related metrics, which were then validated using the validation samples. Because the Random Vector procedure has not been previously used to key personality scales, we detail results for this procedure in Table 3. Results for the Optimal Integer procedure have been previously described (Legree et al., 2021).

Subsequent keying analyses were structured to assess the level of equivalence among the random vector, optimal integer, and item criterion keying vectors. However, two rounds of analyses were conducted because it is not practical to use the Optimal Integer procedure for large personality scales. More specifically:

- For five shorter personality scales, we compared validity estimates for shape scores that were computed using the optimal integer, random vector, item criterion and dominance (i.e., conventional) keying vectors. These analyses were structured to evaluate the utility of these keying procedures to improve the validity of shape scores using these vectors.
- For all 21 personality scales, we focused analyses on comparing: distance scores computed using conventional keys; dot product scores based on the item criterion vectors; shape scores computed using the item criterion vector keys; and shape scores computed using the random vector keys.

Random Vector Keying Descriptives

The Random Keying procedure required creating 10,000 vectors of random numbers for each scale. The individual vector values were generated to be uniformly and randomly distributed over the range (0,1). This range is adequate because the 10,000 candidate vectors were only used to calculate shape scores for each respondent in the two training samples.

This approach allowed us to compute the descriptive statistics describing the distribution of the 10,000 training candidate vector validity estimates within each training sample for each scale. In addition, this approach allowed us to correlate the 10,000 validity estimates computed using the two training samples to assess the stability of these correlations.

Table 3 summarizes validity results for the Random Vector procedure. As expected, the mean candidate vector validity for each scale was centered very close to zero for each of the 21

scales (ranging from -.002 to +.002). Importantly, negative kurtosis estimates were obtained for each distribution (-1.30 up to -.37), which indicates the distribution tails were sharply curtailed. In addition, the standard deviation of these distributions ranged from 0.04 to 0.11. Finally, the stability estimates indicate that the validity estimates were highly stable across the two training samples for nearly all the scales.

Although the number of random vectors that could be generated for any rating-based scale is infinite, these results indicate that very "good" keying vectors were generated for nearly all the scales. For example, the keys identified for the two 5-Point Achievement scales were approximate four *SD*s above the mean. In addition, the Random Vector procedure resulted in the same random vector being identified as the key for several scales in both training samples.

As noted above, the stability estimates for all the scales exceeded .83, except for the 5-point GS-u (r = .30) and HA-u scales (r = .30). These scales are unusual because they represent 2 of 3 scales designed to contain only unidirectional items (i.e., they lacked any reversed items). We suspect that the conventional key for the 3rd unidirectional scale, SP (Shame Proneness) may have been poorly designed and should have contained a mix of extreme values.

Table 3. *Random Vector Development Statistics (Training Samples)*

Scale	*SD*		Kurtosis		Stability[a]	Sample 1		Sample 2	
	Sample 1	Sample 2	Sample 1	Sample 2		Max Value	Validity in S2	Max Value	Validity in S1
				5-Pt Scales					
Ach	0.07	0.06	-0.54	-0.51	.97	.35	.28	.32	.33
AI	0.05	0.04	-0.75	-0.72	.83	.12	.08	.10	.10
FM	0.13	0.12	-1.03	-0.98	.99	.27	.26	.26	.27
GP	0.05	0.06	-0.67	-0.72	.82	.14	.13	.15	.11
GS-u	0.05	0.02	-0.69	-1.10	.30	.11	.02	.05	.03
HA-u	0.06	0.07	-0.47	-0.72	.74	.17	.16	.18	.14
IT	0.06	0.06	-1.30	-1.22	.91	.11	.10	.12	.11
PL	0.10	0.09	-0.93	-0.82	.95	.21	.21	.21	.21
PW	0.06	0.04	-0.82	-0.81	.90	.15	.09	.11	.12
SP-u	0.07	0.06	-0.54	-0.51	.95	.20	.16	.17	.20
ST	0.09	0.07	-0.58	-0.56	.90	.24	.17	.19	.21
WC	0.11	0.10	-0.98	-1.04	.97	.25	.22	.23	.25
				9-Pt Scales					
Ach	0.11	0.10	-0.65	-0.65	.94	.29	.24	.27	.27
AI	0.04	0.04	-0.80	-1.03	.90	.09	.08	.09	.08
CF	0.04	0.07	-0.86	-1.03	.82	.09	.12	.14	.07
FM	0.11	0.09	-1.09	-1.03	.98	.23	.19	.21	.21
HA	0.05	0.05	-0.51	-0.46	.88	.16	.14	.16	.13
PL	0.08	0.08	-0.82	-0.82	.97	.21	.21	.21	.21
SE	0.06	0.06	-0.55	-0.59	.83	.17	.13	.17	.14
ST	0.07	0.09	-0.37	-0.62	.93	.20	.21	.24	.18
TA	0.06	0.04	-1.15	-1.19	.97	.13	.07	.08	.11

Note: Mean sample validity estimates ranged -0.002 to +0.002 for each scale. Scales ending in "u" were designed to contain only unidirectional items.

[a] Stability values reflect the correlation between the 10,000 validity estimates in samples 1 and 2.

Validity Estimates for the Optimal Integer, Random Vector and Item Criterion Procedures: Short Personality Scales

Table 4 summarizes correlational estimates for shape scores computed using the four primary keying vectors by scale within the two training samples (i.e., these correlations are technically not validity estimates). These four keys correspond to the conventional key, or were derived from and the Optimal Integer, Random Vector and Item Criterion procedures using the training samples.

Inspection of the quasi-validity estimates for the training samples indicates that mean validity gains were obtained for each of the alternate keying procedures relative to the mean validity estimates obtained using the conventional keys. In addition, validity gains were slightly higher for the optimal integer and random vector keys than the item criterion keys, but largely comparable. The result that the item criterion keys provided very high but not optimal results, may appear incongruent with general expectations regarding correlational methods. However, item criterion correlations are not computed in a way that would automatically optimize these keying vectors for the purpose of calculating shape scores. Nevertheless, the item criterion vectors consistently provided keys that were superior to conventional keys from a validity perspective.

Table 4. *Shape Validity Estimates for Short Personality Scales by Keying Approach for Training Samples*

5-Point Scales	Conventional	Optimal Integer	Random Vector	Item Criterion
		Training Sample 1		
FM[a]	.232	.274	.274	.253
IT	.104	.113	.114	.108
PL[b]	.164	.228	.214	.206
PW	.073	.147	.148	148
WC[c]	.161	.247	.247	.230
		Training Sample 2		
FM[a]	.219	.261	.259	.243
IT	.089	.115	.115	.105
PL[b]	.138	.215	.213	.193
PW	.075	.112	.110	.112
WC[c]	.172	.223	.226	.216
Mean Validity	**.143**	**.189**	**.188**	**.178**
Percent Gain		**32%**	**32%**	**25%**

All coefficients would be significant at $p < .001$, if inferential statistics were appropriate.

[a] Analyses identified the same random vector across the two training samples for FM.

[b] Analyses identified the same random number vector across the two training samples for PL.

[c] WC shape scores computed using the random keys vectors were highly correlated across the two training samples (i.e., the keys had very similar shape, and the two sets of shape scores were nearly redundant ($r = .992$). Validity estimates may appear to not shrink when rounded to the 3rd decimal point.

Table 5 provides the corresponding validity estimates for the validation samples. Inspection of the validities and the inferential statistics indicates that the gains in shape score validity estimates from the training samples were largely maintained in the cross-validation samples. Furthermore, the validity gains for the alternate procedures were substantial in comparison to the validity estimates for the conventional keys, for which mean validity gains ranged from 25 to 30%. Although largely comparable, slightly higher validities were associated

with the optimal integer ($\overline{R}$ = .184), and random vector keys ($\overline{R}$ = .186) than the item criterion keys ($\overline{R}$ = .178). These results suggest that the Random Vector and Item Criterion methods are reasonable approaches to key larger scales for which the Optimal Integer procedure is not feasible.

Steiger's *Z* procedure was used to compare the difference in magnitude of validity coefficients for shape scores based on the conventional key to the shape scores based on the optimal integer, random vector, and item criterion keys. Steiger's *Z* procedure provides a direct comparison of these validities computed for these correlated predictors (cf. Steiger, 1980; Lee & Preacher, 2013).

These results also document that shape score validity estimates based on the item criterion vectors appear comparable to the shape score validity estimates computed using keys derived from the Optimal Integer and Random Vector procedures. This observation is potentially important because the computation of item criterion vectors is straightforward, and as mentioned above, dot product scores use these vectors to compute scores for all participants. In contrast, shape scores that cannot be computed for participants whose rating vectors lack variance for any scale (i.e., scatter = 0).

Table 5. *Shape Score Cross-Validated Estimates for the Short Personality Scales*[a]

Scale	Shape-Conv Validity	Validity: Shape-OI v Shape-Conv				Validity: Shape-RV v Shape-Conv				Validity: Shape-IC v Shape-Conv				*n*
		Shape-OI Validity	$r_{shp\text{-}OI,shp\text{-}conv}$	*Z*	ΔSig[b]	Shape-RV Validity	$r_{shp\text{-}rv,shp\text{-}conv}$	*Z*	ΔSig[b]	Shape-IC Validity	$r_{shp\text{-}ic,shp\text{-}conv}$	*Z*	ΔSig[b]	
						Validation Sample 1								
FM	.219	.255	.789	-2.53	.011	.259	.831	-3.14	.002	.240	.982	-5.03	.001	1954
IT	.089	.101	.929	-1.41	.159	.107	.923	-2.03	.043	.094	.996	-2.47	.013	1937
PL	.138	.203	.780	-4.04	.001	.213	.591	-3.74	.001	.180	.943	-5.57	.001	1954
PW	.075	.102	.671	-1.48	.140	.094	.411	-0.78	.437	.102	.633	-1.40	.162	1954
WC	.172	.214	.574	-2.06	.040	.225	.681	-3.00	.003	.220	.881	-4.44	.001	1955
						Validation Sample 2								
FM	.232	.271	.826	-3.02	.003	.274	.846	-3.46	.001	.252	.981	-4.67	.001	1948
IT	.104	.107	.826	-0.23	.822	.103	.751	0.06	.950	.106	.954	-0.29	.771	1931
PL	.164	.225	.691	-3.51	.001	.214	.646	-2.68	.007	.204	.926	-4.68	.001	1953
PW	.073	.121	.638	-2.51	.012	.125	.674	-2.86	.004	.132	.763	-3.81	.001	1952
WC	.161	.237	.820	-5.73	.001	.247	.708	-5.10	.001	.218	.928	-6.77	.001	1954
Mean	**.143**	**.184 (29% Gain)**				**.186 (30% Gain)**				**.178 (25% Gain)**				

[a] All validity coefficients significant at $p < .001$.

[b] Significance of change in validity coefficients was based on Steiger's *Z* procedure and provides direct comparisons to the validities of shape scores based on the conventional keys (Steiger, 1980; Lee & Preacher, 2013). All *p*-values < .001 are rounded to .001.

Validity Estimates for the Random Vector and Item Criterion Procedures: All Scales

For the full set of 21 personality scales, we report cross-validated estimates for conceptually important variables in Table 6. These measures correspond to: distance scores computed using conventional keys; dot product scores computed using item criterion vectors; shape scores based on the item criterion vectors; and shape scores based on the random vector keys. These scoring algorithm and keying combinations were chosen to ensure representation of methods that: have been traditionally used (distance); can be computed for respondents with rating vectors lacking scatter (dot product); provide insight into dot product score variance (shape scores computed using the item criterion vectors); and assess the consistency in validity estimates for shape scores that are based on either the item criterion or the random vector keys.

Table 6 reports validity estimates for scales and scoring methods. We first compared the distance validities to the validities for the dot product scores, the shape scores based on the item criterion vectors, and the shape scores based on the random key vectors. We assessed the mean differences across scoring methods using Paired Sample *t*-tests. All comparisons involving the distance validities favored the use of alternate scoring methods (all $t(41) > 5.77$; all $p < .001$). Furthermore, the validities for dot product, shape scores based on the item criterion vectors, and shape scores based on the random vector keys were substantially greater than the corresponding validities based on distance scores: $\bar{r}_{dist} = .11$ versus $\bar{r}_{dot} = .17$, $\bar{r}_{shape\text{-}ic} = .16$, and $\bar{r}_{shape\text{-}rv} = .16$.

Next, we used Steiger's *Z* statistic to compare validities for dot product and shape scores computed using the item criterion vectors. This comparison explores the possibility that dot product score validities may be affected by scatter and elevation to either enhance or limit their validity over scale validities for shape scores. These analyses identified approximately seven scales in each validation sample for which the validity estimates between the dot product scores and shape computed using the item criterion vectors differed significantly.

However, the direction of these validity differences did not consistently favor either the dot product or the shape metric. For example, analyses conducted using Validation Sample 1 documented significant differences between the validity correlations that favored the dot product metric for four scales and the shape metric for four scales. Similarly, analyses conducted using Validation Sample 2 documented significant differences between the validity correlations that favored the dot product metric for four scales and the shape metric for three scales.

In addition, the direction and magnitude of these differences over the two samples appeared quite consistent. To quantify this consistence, we correlated the differences between the validities over the two samples. This correlation indicates that the magnitudes and direction of these differences over the 21 scales was extremely consistent, $r = .82$, $p < .001$.

The results comparing the dot product and shape scores are similar to those for distance analyses showing that distance score variance reflects the combined action of shape, scatter, and elevation metrics (Legree et al., 2018). As described above, the result supports expectations that (1) dot product scores can be modelled using the scatter, elevation, and shape metrics, and (2) dot product score validities can be improved by optimally weighting the scatter and elevation metrics using regression procedures (i.e., we expect that the scatter and elevation metrics may be poorly weighted by the dot product scores for the purpose of maximizing scale validity).

Table 6. *Comparison of Bivariate Validities for Distance, Dot Product, and Shape Metrics Computed Using the Item Criterion and Random Vector Keys*

Scale	Validation Sample 2								Validation Sample 1							
	Validity Estimates				Steiger's z Test for Dependent Correlations: Dot vs ShpIC[b]				Validity Estimates				Steiger's z Test for Dependent Correlations: Dot vs ShpIC			
	Dist	Dot	ShpIC	ShpRV	$r_{dot,sh\text{-}ic}$	Dot – IC	*z*	Sig[c]	Dist	Dot	ShpIC	ShpRV	$r_{dot,sh\text{-}ic}$	Dot – IC	*Z*	Sig[c]
All 5-Point Scales																
Ach	.30	.35	.34	.33	.88	.008	0.97	*ns*	.27	.31	.31	.28	.88	.005	0.97	*ns*
AI	.03[a]	.09	.11	.10	.86	-.024	-1.68	*ns*	.02[a]	.08	.08	.08	.97	.001	1.68	*ns*
FM	.27	.30	.25	.27	.87	.047	4.51	.001	.27	.29	.24	.26	.85	.051	4.19	.001
GP	.10	.11	.14	.11	.81	-.024	-0.41	*ns*	.09	.13	.12	.13	.78	.004	1.69	*ns*
GS-u	.10	.11	.04[a]	.03[a]	-.118	.070	1.89	.059	.06	.06	.01[a]	.02[a]	.005[a]	.062	1.60	.109
HA-u	.08	.11	.12	.14	.41	-.010	-0.85	*ns*	.12	.16	.12	.16	.44	.047	1.99	.047
IT	.12	.12	.11	.10	.85	.011	0.66	*ns*	.10	.10	.09	.11	.86	.010	0.83	*ns*
PL	.18	.21	.20	.21	.86	.006	1.93	.054	.15	.19	.18	.21	.86	.005	0.42	*ns*
PW	.08	.12	.13	.13	.77	-.014	-1.47	*ns*	.08	.10	.10	.09	.82	.001	1.47	*ns*
SP-u	.03[a]	.20	.22	.20	.81	-.021	-1.26	*ns*	.03	.18	.18	.16	.81	.003	1.26	*ns*
ST	.15	.22	.24	.21	.74	-.016	-0.81	*ns*	.13	.19	.19	.17	.70	.004	0.84	*ns*
WC	.18	.26	.22	.25	.90	.041	4.08	.001	.19	.28	.22	.23	.88	.055	5.09	.001
All 9-Point Scales																
Ach	.27	.27	.31	.27	.80	-.039	-2.93	.003	.25	.25	.28	.24	.88	-.026	-2.42	.015
AI	.02[a]	.11	.08	.08	.78	.031	1.99	.047	.027[a]	.113	.083	.076	.679	.030	1.66	.097
CF	.05	.02[a]	.07	.07	.88	-.045	-4.44	.001	-.01[a]	.061	.105	.12	.429	-.044	-1.81	.071
FM	.18	.19	.21	.21	.93	-.022	-2.40	.016	.16	.176	.178	.19	.933	-.002	-.240	*ns*
HA	.01[a]	.17	.15	.13	.74	.022	1.24	*ns*	.02[a]	.14	.14	.14	.69	.001	0.06	*ns*
PL	.04[a]	.20	.20	.21	.84	.000	0.00	*ns*	.05	.18	.17	.21	.89	.002	0.19	*ns*
SE	.06	.15	.16	.14	.89	-.007	-0.95	*ns*	.08	.14	.16	.13	.85	-.017	-1.39	*ns*
ST	.06	.18	.21	.18	.84	-.031	-2.38	.017	.05	.22	.26	.21	.88	-.039	-3.61	.001
TA	.08	.14	.11	.11	.81	.035	2.15	.032	.05	.09	.07	.07	.82	.023	3.70	.001
Mean	**.114**	**.173**	**.172**	**.166**					**.104**	**.164**	**.156**	**.156**				

ShpIC = Shape, Item Criterion. ShpRV = Shape, Random Vector. Scales ending in "u" were designed to contain only unidirectional items.

[a] All validity coefficients significant at $p < .05$ unless indicated.

[b] Steiger's *Z* procedure was used to compare the difference in magnitude of validity coefficients for dot product and shape metric based on the item criterion vectors by scale and based on dependent correlations. Steiger's *Z* procedure provides a direct comparison of these correlated predictors (Steiger, 1980; Lee & Preacher, 2013).

[c] All *p*-values < .001 were rounded to .001.

Scale Analyses

Based on expectations that PSMs can be used to model the validity underlying a variety of singular scoring procedures, as well as the preliminary keying results outlined above, we proposed the following two hypotheses:

Hyp 1. Shape, delta, and scatter metrics will account for nearly all the variance in dot product scores: $R_{Distance,Shape.Scatter.Delta} > .80$.

Hyp 2. Scatter and elevation scatter metrics will add incremental validity to dot product scores against performance outcomes because dot product scores may represent suboptimal weighting of these PSMs.

The first hypothesis references the .80 value because this point is commonly viewed as an upper bound in correlational research. More generally, we are evaluating whether the vast majority of dot product score variance can be modelled as main effects using the shape, scatter and elevation metrics.

The second hypothesis focuses on the potential of using the scatter and elevation metrics to provide incremental validity beyond the dot product scores against outcome criteria. More generally, we are evaluating whether the scatter and elevation metrics can be used to adjust the dot product scores to limit the possibility that the dot product scores represent poorly weighted composites corresponding to sources of variance underling the shape, scatter and elevation metrics.

Regression of Dot Product Scores on PSMs

The first hypothesis proposed that the shape (computed using the item criterion vectors), elevation, and scatter metrics will account for nearly all the variance in dot product scores. This result was assessed using a hierarchical regression procedure. In Step 1 of each regression model, we regressed the dot product on the shape scores, and in Step 2 we added the elevation and scatter metrics. Separate regression models were computed for each of the 21 scales using each validation sample (i.e., 42 analyses).

The regressions results supported Hypothesis 1 for each scale and validation sample. In fact, the lowest observed Multiple *R* at Step 2 was equal to .87. Furthermore, the mean Multiple *R* at Step 2 was extremely high for Sample 1 ($\overline{R}$ = .945), and Sample 2 ($\overline{R}$ = .955).

The results also indicated that the elevation and scatter metrics accounted for substantial dot product score variance beyond the shape metrics. This can be observed by comparing the mean multiple correlations at Steps 1 and 2 for each validation sample: Validation Sample 1, $\overline{R}_{step1}$ = .780 vs. $\overline{R}_{step2}$ = .945; Validation Sample 2, $\overline{R}_{step1}$ = .757 vs. $\overline{R}_{step2}$ = .955. Therefore, the analyses supported Hypothesis 1, and we conclude that the shape, scatter and elevation metrics account for most of the dot product score variance as expected. Table 7 reports details.

Table 7. *Dot Product Scores Regressed on: Shape Metrics (Step 1); and Elevation and Scatter Metrics (Step 2) by Scale*

Scale	Validation Sample #2							Validation Sample #1						
	Step 1: Shape	Step 2: Scatter and Elevation						Step 1: Shape	Step 2: Scatter and Elevation					
	$R_{step\text{-}1}$	$R_{step\text{-}2}$	R^2	Adj R^2	ΔR^2	F-change	Sig[a]	$R_{step\text{-}1}$	$R_{step\text{-}2}$	R^2	Adj R^2	ΔR^2	F-change	Sig[a]
All 5-Point Scales														
Ach	.880	.985	.971	.971	.196	6635	.001	.876	.986	.973	.973	.206	7459	.001
AI	.856	.979	.959	.959	.227	5410	.001	.972	.981	.962	.961	.016	418	.001
FM	.865	.981	.962	.962	.214	5485	.001	.854	.980	.960	.960	.231	5619	.001
GP	.812	.980	.961	.961	.302	7585	.001	.782	.983	.965	.965	.354	9987	.001
GS-u	.118	.997	.994	.994	.981	144944	.001	.005[ns]	.997	.993	.993	.993	118814	.001
HA-u	.414	.986	.971	.971	.800	27017	.001	.442	.980	.961	.961	.766	19133	.001
IT	.851	.983	.966	.966	.242	6935	.001	.862	.984	.968	.968	.225	6894	.001
PL	.856	.985	.970	.970	.238	7869	.001	.859	.983	.966	.966	.228	6593	.001
PW	.765	.983	.966	.966	.381	11033	.001	.817	.975	.951	.951	.284	5597	.001
SP-u	.811	.956	.914	.914	.256	2904	.001	.810	.959	.921	.920	.264	3245	.001
ST	.737	.989	.979	.979	.435	19849	.001	.697	.988	.976	.976	.491	20040	.001
WC	.902	.977	.955	.955	.141	3084	.001	.878	.976	.953	.953	.183	3827	.001
All 9-Point Scales														
Ach	.795	.865	.749	.748	.117	452	.001	.879	.907	.823	.823	.049	270	.001
AI	.782	.921	.848	.847	.235	1477	.001	.679	.915	.838	.838	.377	2234	.001
CF	.882	.905	.818	.818	.040	207	.001	.429	.982	.965	.965	.781	21054	.001
FM	.932	.955	.912	.912	.044	489	.001	.933	.955	.912	.912	.041	453	.001
HA	.740	.855	.730	.730	.182	651	.001	.690	.874	.764	.763	.288	1175	.001
PL	.842	.865	.748	.748	.039	148	.001	.887	.911	.831	.830	.044	251	.001
SE	.891	.920	.845	.845	.052	323	.001	.852	.939	.882	.882	.156	1277	.001
ST	.842	.913	.834	.834	.124	720	.001	.881	.931	.867	.867	.092	667	.001
TA	.809	.872	.760	.760	.106	425	.001	.821	.870	.756	.756	.082	321	.001
Mean	**.780**	**.945**		**.890**				**.757**	**.955**		**.910**			

Note: Sample size ranged from 1629 to 1954. All models sig at $p < .001$ at Step 1 unless noted. All models sig at $p < .001$ at Step 1 unless noted.
[a] All p-values <.001 were rounded to .001.

Regression of OMS on Dot Product, Scatter Metrics, and Elevation Metrics

Hypothesis 2 proposed that the scatter and elevation metrics will increment the validity of the dot product scores for many scales. Hypothesis 2 is similar to regression results showing that (1) distance score validity can be modelled using PSMs, and (2) scale validity can be enhanced by combining distance scores with PSMs in regression models. More generally, Hypothesis 2 reflects the expectation that scale validity can be enhanced by using regression procedures to properly weight the dot product, scatter, and elevation metrics because dot product scores can be influenced by scatter and elevation effects in ways that may limit scale validity.

This expectation was assessed using hierarchical regression models for each scale and sample. Within each regression model, we regressed the OMS criterion on the dot product scores at Step 1, and added the elevation and scatter metrics at Step 2. Separate regression models were computed for each of the 21 scales using each validation sample (i.e., 42 analyses). Given the results from Hypothesis 1 showing that nearly all dot product score variance can be accurately modelled using PSMs, we could justify using 1-tail inferential statistics to assess the presence of these gains. However, we report 2-tail inferential statistics to evaluate the significance of the validity gains in these models as to enhance confidence in our conclusions.

The regressions results supported Hypothesis 2 and indicated the elevation and scatter metrics can be used to enhance dot product scale validity for many of these personality scales (i.e., 67% of the regression models provided significant gains at Step 2). Across the 21 scales, these gains resulted in an 10% percent gain being associated with the inclusion of scatter and elevation metrics to enhance dot product scale validity (i.e., $\overline{R}_{step2} = .189$, $\overline{R}_{step1} = .169$ results in an 12% validity gain). Table 8 reports details for the regression analyses and contains bivariate validity estimates for the distance scores.

The validity gains for the dot product scores adjusted for the scatter and elevation metrics can also be compared to the simple distance validity estimates contained in Table 8 ($\bar{r}_{dist} = .105$). It follows that the combined use of the dot product, scatter and elevation metrics provides an 80% gain relative to using simple distance scores for the 21 personality scales. Inspection of the scale distance and dot product adjusted validities suggests that greater validity gains may be associated with scales having greater distance validities.

As a post-hoc assessment of this observation, we correlated the distance validity estimates with the magnitudes of the validity gains of the dot product adjusted for scatter and elevation scores over the validity estimates for the distance scores. (Distance validity estimates are reported in Table 8.). This correlational analysis demonstrated a strong relationship between the magnitude of the simple distance scores and the magnitude of the validity gains over the distance scores, $r = .55$, $p < .001$. It follows that the dot product adjusted scores act to multiply validity gains (i.e., scales with lower distance validities are enhanced minimally in comparison to scales with higher distance validities).

This result has two implications. First, it implies that the scale validity gains are related to the underlying construct as opposed to general participant response tendencies. Second, it suggests highly valid personality batteries can be constructed by including only highly valid scales in the assessment batteries and using sophisticated keying and scoring algorithms.

Table 8. *OMS Regressed on the Dot Product, Scatter and Elevation Metrics for Each Personality Scale*

Scale	Validation Sample #2								Validation Sample #1							
	Dis Val	Regression Model							Dis Val	Regression Model						
		Step 1: Dot	Step 2: Scatter and Elevation							Step1: Dot	Step 2: Scatter and Elevation					
	r	*R*	*R*	R^2	Adj R^2	ΔR^2	F-chng	Sig[a]	*r*	*R*	*R*	R^2	Adj R^2	ΔR^2	F-chag	Sig[a]
All 5-Point Scales																
Ach	.30	.35	.36	.126	.125	.003	3.29	.037	.27	.31	.32	.103	.102	.007	8.04	.001
AI	.03	.09	.13	.016	.015	.008	8.03	.001	.02	.08	.08	.006	.005	.000	0.26	.774
FM	.27	.30	.31	.098	.097	.008	8.25	.001	.27	.29	.31	.097	.095	.012	12.88	.001
GP	.09	.11	.15	.023	.021	.010	10.09	.001	.09	.13	.14	.018	.017	.002	2.15	.117
GSu	.11	.12	.12	.015	.013	.002	1.59	.204	.08	.09	.09	.009	.007	.002	1.56	.210
HAu	.08	.11	.15	.021	.020	.009	8.85	.001	.12	.16	.17	.028	.026	.002	1.60	.202
IT	.12	.12	.12	.014	.013	.000	0.16	.848	.10	.11	.11	.012	.011	.001	0.89	.410
PL	.18	.21	.22	.049	.047	.005	4.64	.010	.15	.19	.19	.035	.034	.001	1.23	.292
PW	.07	.11	.15	.021	.020	.008	8.03	.001	.07	.10	.11	.012	.010	.001	1.14	.319
SPu	.03	.20	.23	.053	.052	.015	15.84	.001	.03	.18	.21	.042	.040	.010	10.56	.001
ST	.15	.22	.26	.065	.063	.016	16.21	.001	.13	.19	.21	.044	.042	.007	7.51	.001
WC	.18	.26	.29	.084	.083	.017	18.58	.001	.19	.28	.31	.095	.093	.019	20.83	.001
All 9-Point Scales																
Ach	.27	.27	.28	.075	.074	.003	3.58	.028	.25	.24	.25	.061	.059	.005	4.72	.009
AI	.02	.11	.12	.014	.013	.002	2.15	.116	.03	.11	.12	.013	.012	.001	0.83	.435
CF	.05	.02[ns]	.10	.009	.008	.009	8.61	.001	.01	.06	.12	.014	.013	.011	10.35	.001
FM	.18	.19	.20	.041	.039	.005	4.68	.009	.16	.18	.18	.032	.031	.001	1.33	.263
HA	.01	.17	.20	.040	.038	.010	10.41	.001	.01	.15	.16	.025	.023	.003	2.80	.061
PL	.05	.20	.20	.039	.038	.001	0.96	.382	.04	.18	.19	.035	.034	.003	3.00	.050
SE	.06	.15	.17	.028	.026	.006	5.67	.004	.08	.15	.16	.025	.024	.004	4.14	.016
ST	.06	.18	.20	.038	.037	.006	5.85	.003	.06	.23	.25	.061	.059	.008	8.05	.001
TA	.08	.13	.16	.025	.024	.008	7.49	.001	.05	.08	.12	.013	.012	.006	6.12	.002
Mean	**.113**	**.172**	**.196**						**.097**	**.166**	**.181**					

Note: Sample size ranged from 1915 to 1955. All models sig at p < .001 at Step 1 unless noted. Significance values (final column) based on 2-tailed estimates. Distance validities have been reflected. The elevation metric was not included at Step 2 for three unidimensional scales because this metric is redundant with distance scores for unidirectional scales (i.e., $r = +1$ or -1)

[a] All *p*-values <.001 were rounded to .001.

Battery Analyses

We conducted a series of hierarchical regression analyses to provide battery-level validity estimates against the Cadet OMS criterion. The first set of analyses included all 21 scales in the battery. The second and third sets of analyses included either the 12 scales using the 5-point rating format, or the 9 scales using the 9-point rating format. The fourth and fifth sets of analyses limited the batteries to include only scales with high scale validities. These analyses were conducted using each of the cross-validation samples.

Battery Composed of 21 Scales Using 5-Point and 9-Point Response Formats

We computed validity estimates for four hierarchical analyses that primarily differed in the choice of predictor included in Step 1 of each model. These four focal predictors corresponded to:

1. Distance scores, which can be computed for all respondents.
2. Dot product scores, which can be computed for all respondents.
3. Shape scores based on the item criterion vectors, which can only be computed for respondents whose rating vectors contain scatter.
4. Shape scores based on the random vector keys, which can only be computed for respondents whose rating vectors contain scatter.

In Step 2, we entered the elevation and scatter metrics for each of 21 variables. We also specified a listwise deletion procedure. We note that the validity estimate for distance at Step 1 corresponds to the battery validity estimate computed using only distance scores and conventional scoring keys. This estimate provides a baseline against which to gauge validity estimates using the more sophisticated scoring algorithms and keying procedures.

Table 9 reports summary statistics for each of these regression models. These results indicate that the conventional distance estimate provides a respectable, albeit modest validity estimate: $\overline{R}_{\text{distance}} = .454$, which accounts for 21% of the OMS variance. In contrast, substantially higher validity estimates were obtained by using each of the three alternate focal predictors at Step 1, with the elevation and scatter metrics added at Step 2: $\overline{R}_{\text{dot}} = .609$; $\overline{R}_{\text{shape-ic}} = .610$; $\overline{R}_{\text{shape--rv}} = .614$. Each of these models account for approximately 37% of the criterion variance.

Table 9 reports the ratio of the coefficients of determination at Step 1 versus Step 2 for each model. These ratios indicate the variance associated with the scatter and elevation metrics account for approximately 35% of the composite variance for each model. These results demonstrate that variance associated with the elevation and scatter metrics is a secondary, but important source of predictive validity. These results are notable because the scatter and elevation metrics are computed as descriptive statistics (i.e., much scale validity can be captured by variance not associated with any keying vector).

We note that the sample sizes reported for analyses using shape scores were lower than those reported for the dot product and distance scores. This loss of sample primarily occurred because 3 of the 5-point CBEF scales were unidirectional and contained relatively few items. This resulted in respondent rating vectors often not containing variance so that shape scores could not be computed for approximately 18% of the respondents.

Nevertheless, the regression models demonstrated that substantial validity gains could be realized for personality scales scored and keyed using any of these procedures. However, the simplicity of computing dot product scores that are adjusted for scatter and elevation effects, combined with the avoidance of data associated with the shape scores, indicates that this is the most practical approach to optimize personality scale validity estimates for operational and many research applications.

Table 9. *OMS Regressed on the Full 21-Scale Personality Battery*

Validation Sample	Step 1: Primary Scale Metric				Step 2: Scatter and Elevation				Ratio of Coefficients of Determination[c]	*n*
	R	R^2	*F* stat[a]	Sig[b]	*R*	ΔR^2	*F* stat[a]	Sig[b]		
					Distance					
VS #2	.46	.212	24.19	.001	.59	.134	9.74	.001	.61	1914
VS #1	.45	.200	22.73	.001	.55	.104	7.12	.001	.66	1929
Mean	**.454**	**.206**			**.570**	**.119**			**.64**	
					Dot Product					
VS #2	.53	.282	35.33	.001	.63	.112	8.16	.001	.72	1914
VS #1	.50	.250	30.18	.001	.59	.101	6.87	.001	.71	1929
Mean	**.516**	**.266**			**.610**	**.106**			**.71**	
					Shape based on Item Criterion Vectors					
VS #2	.50	.253	25.17	.001	.63	.137	8.13	.001	.65	1581
VS #1	.46	.213	19.94	.001	.59	.137	7.58	.001	.61	1570
Mean	**.483**	**.233**			**.609**	**.137**			**.63**	
					Shape based on Random Vectors					
VS #2	.517	.267	27.09	.001	.628	.127	7.59	.001	.68	1581
VS #1	.483	.233	22.42	.001	.601	.128	7.16	.001	.65	1570
Mean	**.501**	**.251**			**.614**	**.127**			**.66**	

[a] Degrees of freedom lost at: Step 1, 21 *df*; Step 2, 42 *df*.

[b] All *p*-values <.001 were rounded to .001.

[c] Ratio of the Coefficient of Determination at Step 1 to the Coefficient of Determination at Step 2 (i.e., the ratio of R^2 at Step 1 to R^2 at Step 2) for each model.

Battery Composed of 12 Scales Using the 5-Point Response Format

We repeated these analyses using the CBEF battery that contained 12 scales and incorporated the 5-point response format. The hierarchical analyses followed the pattern described above for the battery containing the 12 scales. In summary, the four models primarily differed by the choice of predictor included in Step 1 of each model. These four focal predictors corresponded to:

1. Distance scores, which can be computed for all respondents.
2. Dot product scores, which can be computed for all respondents.
3. Shape scores based on the item criterion vectors, which can only be computed for respondents whose rating vectors contain scatter.
4. Shape scores based on the random vector keys, which can only be computed for respondents whose rating vectors contain scatter.

In Step 2, we entered the elevation and scatter metrics for each of 12 variables. As described for the first set of battery analyses, the Step 1 validity estimate computed for the distance model provides a baseline against which to gauge validity estimates for the sophisticated scoring and keying procedures.

Table 10 reports summary statistics for each of these analyses. These results indicate that the conventional distance estimate provides a respectable, but moderate validity estimate, $\overline{R}_{\text{distance}} = .421$, which accounts for 18% of the OMS variance In contrast, substantially higher validity estimates were obtained by using each alternate scoring procedure with the elevation and scatter metrics: $\overline{R}_{\text{dot}} = .567$; $\overline{R}_{\text{shape-ic}} = .567$; $\overline{R}_{\text{shape-rv}} = .578$. Each of these models account for approximately 32% of the OMS variance. Therefore, the regression models indicate that substantial validity gains could be realized using only the CBEF 12-scale battery, albeit with some loss of predictive validity.

Finally, the Determination Coefficient ratios indicate that variance associated with the elevation and scatter metrics is an important source of predictive validity. As in the earlier model, these metrics account for approximately 33% of the predictive variance in each model. Compare results in Tables 9 and 10.

Table 10. *OMS Regressed on the CBEF 12-Scale Personality Battery*

Validation Sample	Step 1: Primary Scale Metric				Step 2: Scatter and Elevation				Ratio of Coefficients of Determination[c]	*n*
	R	R^2	*F* stat[a]	Sig[b]	*R*	ΔR^2	*F* stat[a]	Sig[b]		
					Distance					
VS #2	.43	.185	36.84	.001	.53	.094	11.86	.001	.66	1955
VS #1	.41	.169	32.87	.001	.51	.088	10.88	.001	.66	1956
Mean	**.421**	**.177**			**.518**	**.091**			**.66**	
					Dot Product					
VS #2	.48	.229	47.98	.001	.59	.114	13.90	.001	.67	1955
VS #1	.46	.209	42.81	.001	.55	.090	10.27	.001	.70	1956
Mean	**.468**	**.219**			**.567**	**.102**			**.68**	
					Shape based on Item Criterion Vectors					
VS #2	.47	.221	38.49	.001	.59	.128	13.14	.001	.63	1640
VS #1	.42	.172	27.77	.001	.54	.122	11.34	.001	.59	1616
Mean	**.443**	**.197**			**.567**	**.125**			**.61**	
					Shape based on Random Vectors					
VS #2	.490	.240	42.75	.001	.597	.117	12.14	.001	.67	1640
VS #1	.446	.198	33.12	.001	.558	.113	10.79	.001	.64	1616
Mean	**.468**	**.219**			**.578**	**.115**			**.66**	

[a] Degrees of freedom lost at: Step 1, 12 *df;* Step 2, 24 *df.*

[b] All *p*-values <.001 were rounded to .001.

[c] Ratio of the Coefficient of Determination at Step 1 to the Coefficient of Determination at Step 2 (i.e., the ratio of R^2 at Step 1 to R^2 at Step 2) for each model.

Battery Composed of 9 Scales Using the 9-Point Response Format

We repeated these analyses using the CPM battery that contained the nine scales using the 9-point response format. The structure of the hierarchical analyses followed the pattern as described for the CBEF battery containing 12 scales. In summary, the four models primarily differed by the choice of focal predictor included in Step 1 of each model. These four focal predictors corresponded to:

1. Distance scores, which can be computed for all respondents.
2. Dot product scores, which can be computed for all respondents.
3. Shape scores based on the item criterion vectors, which can only be computed for respondents whose rating vectors contain scatter.
4. Shape scores based on the random vector keys, which can only be computed for respondents whose rating vectors contain scatter.

We entered the elevation and scatter metrics for each of nine scales at Step 2. As in the above models, the Step 1 validity estimate using the distance scores provides a baseline to gauge the validity estimates for the sophisticated scoring and keying procedures.

Table 11 provides results for each series of analyses. These results indicate that the conventional distance estimate provides a modest validity estimate, $\overline{R}_{\text{distance}} = .299$. This result is generally consistent with the view that scale validity is generally increased by using 5-point Likert rating formats as opposed to longer rating scales.

However, much more impressive validity estimates were obtained using each alternative scoring procedure at Step 1, and adding the elevation and scatter metrics at Step 2: $\overline{R}_{\text{dot}} = .472$; $\overline{R}_{\text{shape-ic}} = .460$; and $\overline{R}_{\text{shape-rv}} = .459$. While these validities are lower than the corresponding validity estimates for the Full and CBEF batteries, they still exceed the validities obtained for most personality batteries against high quality outcomes such as OMS. In fact, these estimates exceed the CBEF distance validity estimate reported in Table 10, $\overline{R}_{\text{distance}} = .421$.

As observed in the earlier models, the Determination Coefficient ratios indicate that variance associated with the elevation and scatter metrics is an important source of predictive validity – accounting for approximately 33% of the predictive variance in each model.

Table 11. *OMS Regressed on the CPM 9-Scale Personality Battery*

Validation Sample	Step 1: Primary Scale Metric				Step 2: Scatter and Elevation				Ratio of Coefficients of Determination[c]	*n*
	R	R^2	*F* stat[a]	Sig[b]	*R*	ΔR^2	*F* stat[a]	Sig[b]		
					Distance					
VS #2	.32	.100	23.37	.001	.44	.090	11.64	.001	.52	1914
VS #1	.28	.080	18.62	.001	.38	.067	8.35	.001	.54	1929
Mean	**.299**	**.090**			**.410**	**.079**			.53	
					Dot Product					
VS #2	.40	.163	41.19	.001	.50	.082	10.20	.001	.67	1914
VS #1	.37	.133	32.74	.001	.45	.067	8.86	.001	.66	1929
Mean	**.385**	**.148**			**.472**	**.074**			**.67**	
					Shape based on Item Criterion Vectors					
VS #2	.39	.155	38.05	.001	.47	.061	8.03	.001	.72	1877
VS #1	.37	.139	33.69	.001	.45	.068	8.83	.001	.67	1896
Mean	**.383**	**.147**			**.460**	**.064**			.69	
					Shape based on Random Vectors					
VS #2	.40	.161	39.74	.001	.47	.060	7.85	.001	.73	1877
VS #1	.37	.140	34.07	.001	.45	.061	7.91	.001	.70	1896
Mean	**.388**	**.150**			**.459**	**.060**			**.71**	

[a] Degrees of freedom lost at: Step 1, 9 *df;* Step 2, 18 *df.*

[b] All *p*-values <.001 were rounded to .001.

c Ratio of the Coefficient of Determination at Step 1 to the Coefficient of Determination at Step 2 (i.e., the ratio of R^2 at Step 1 to R^2 at Step 2) for each model.

Battery Composed of 11 High Value Scales

We repeated these analyses using those scales that are in Table 8 with validity coefficients near or exceeding the .20 standard for a moderately large validity correlation (Cohen, 1988). This strategy identified six scales from the CBEF and five scales from the CPM battery. These scales provided coverage for seven of the predictor constructs. The hierarchical analyses followed the same pattern as described for the previous battery analyses. The four models used the following focal predictors at Step 1:

1. Distance scores, which can be computed for all respondents.
2. Dot product scores, which can be computed for all respondents.
3. Shape scores based on the item criterion correlation vectors, which can only be computed for respondents whose rating vectors contain scatter.
4. Shape scores based on the random vector keys, which can only be computed for respondents whose rating vectors contain scatter.

We then entered the elevation and scatter metrics for each of 11 scales at Step 2. As in the above models, the Step 1 validity estimate using the distance scores provides a baseline to gauge the validity estimates for the sophisticated scoring and keying procedures.

Table 12 summarizes results for each series of analyses. These results indicate that the conventional distance estimate provides a modest validity estimate, $\overline{R}_{\text{distance}} = .437$ which accounts for 18% of the criterion variance. For this battery, substantially greater validity estimates were obtained for each of the alternate scoring procedures with the incorporation of the elevation and scatter metrics: $\overline{R}_{\text{dot}} = .575$; $\overline{R}_{\text{shape-ic}} = .576$; and $\overline{R}_{\text{shape-rv}} = .585$. Each of these models account for approximately 33% of the criterion variance.

While these validities are lower than the corresponding validity estimates for the Full 21-Scale Battery, they slightly exceed validities for the CBEF 12-Scale battery, despite having fewer scales. From an empirical perspective, the results for this battery is second to only the Full Battery, yet contains 10 fewer scales.

Table 12. *OMS Regressed on the Select 11-Scale Personality Battery*

Validation Sample	Step 1: Primary Scale Metric				Step 2: Scatter and Elevation				Ratio of Coefficients of Determination[c]	*n*
	R	R^2	*F* stat[a]	Sig[b]	*R*	ΔR^2	*F* stat[a]	Sig[b]		
					Distance					
VS #2	.43	.187	40.31	.001	.55	.112	14.52	.001	.63	1935
VS #1	.42	.176	37.41	.001	.52	.095	11.83	.001	.65	1938
Mean	**.427**	**.182**			**.535**	**.104**			**.64**	
					Dot Product					
VS #2	.49	.243	56.15	.001	.59	.103	13.64	.001	.70	1935
VS #1	.47	.217	48.62	.001	.56	.096	12.13	.001	.69	1938
Mean	**.480**	**.230**			**.575**	**.100**			**.70**	
					Shape based on Item Criterion Vectors					
VS #2	.48	.231	52.03	.001	.59	.113	14.83	.001	.67	1922
VS #1	.45	.201	43.91	.001	.57	.117	14.83	.001	.63	1928
Mean	**.465**	**.216**			**.576**	**.115**			**.65**	
					Shape based on Random Vectors					
VS #2	.50	.250	57.98	.001	.60	.106	14.07	.001	.70	1922
VS #1	.46	.211	46.58	.001	.57	.117	14.96	.001	.64	1928
Mean	**.480**	**.231**			**.585**	**.111**			**.67**	

[a] Degrees of freedom lost at: Step 1, 12 *df;* Step 2, 24 *df.*

[b] All *p*-values <.001 were rounded to .001.

[c] Ratio of the Coefficient of Determination at Step 1 to the Coefficient of Determination at Step 2 (i.e., the ratio of R^2 at Step 1 to R^2 at Step 2) for each model.

Battery Composed of 6 High Value CBEF Scales

Based on the concern that an 11-scale battery may be too large for some applications, we repeated these analyses using only CBEF 5-point response scales that are listed in Table 8 with validity coefficients near or exceeding the .20 standard for a moderately large validity correlation (Cohen, 1988). This strategy is also equivalent to using only the six CBEF scales that were included in the previous analysis. The hierarchical analyses followed the same pattern as described for the previous battery analyses. The four models used the following focal predictors at Step 1:

1. Distance scores, which can be computed for all respondents.
2. Dot product scores, which can be computed for all respondents.
3. Shape scores based on the item criterion correlation vectors, which can only be computed for respondents whose rating vectors contain scatter.
4. Shape scores based on the random vector keys, which can only be computed for respondents whose rating vectors contain scatter.

We then entered the elevation and scatter metrics for each of six scales at Step 2. As in the above models, the Step 1 validity estimate using the distance scores provides a baseline to gauge the validity estimates for the sophisticated scoring and keying procedures.

Table 13 provides the results for each series of analyses. These results indicate that the conventional distance estimate provides a modest validity estimate, $\overline{R}_{\text{distance}} = .395$, which accounts for 16% of the OMS variance. For this battery, substantially higher validity estimates were obtained for each of the alternate scoring procedures with the incorporation of the elevation and scatter metrics: $\overline{R}_{\text{dot}} = .533$; $\overline{R}_{\text{shape-ic}} = .538$; and $\overline{R}_{\text{shape-rv}} = .550$. Each of these models account for approximately 29% of the criterion variance.

While these validities are lower than the corresponding validity estimates for the Full and CBEF batteries, they still exceed the validities obtained for many personality batteries against high quality outcomes such as the OMS. Furthermore, they require the administration of personality scales corresponding to 6 constructs.

On a conceptual level, the result suggests that scales for a surprisingly small number of constructs can provide surprisingly high validity using sophisticated keying procedures and scoring algorithms – as opposed to attempting to better predict important outcomes by endlessly searching for new constructs that are scored using simple keying and scoring approaches.

Table 13. *OMS Regressed on Select the 6-Scale Personality Battery*

Validation Sample	Step 1: Primary Scale Metric				Step 2: Scatter and Elevation				Ratio of Coefficients of Determination[c]	*n*
	R	R^2	*F* stat[a]	Sig[b]	*R*	ΔR^2	*F* stat[a]	Sig[b]		
					Distance					
VS #2	.40	.161	62.22	.001	.49	.083	19.37	.001	.66	1955
VS #1	.39	.150	57.47	.001	.48	.083	19.00	.001	.65	1956
Mean	**.395**	**.156**			**.489**	**.083**			**.65**	
					Dot Product					
VS #2	.45	.199	80.74	.001	.55	.099	22.84	.001	.67	1955
VS #1	.43	.185	73.75	.001	.52	.085	18.80	.001	.69	1956
Mean	**.438**	**.192**			**.533**	**.092**			**.68**	
					Shape based on Item Criterion Vectors					
VS #2	.44	.192	76.93	.001	.55	.109	25.17	.001	.64	1948
VS #1	.41	.166	63.94	.001	.53	.111	24.42	.001	.60	1937
Mean	**.423**	**.179**			**.538**	**.110**			**.62**	
					Shape based on Random Vectors					
VS #2	.47	.219	89.82	.001	.56	.099	23.22	.001	.69	1934
VS #1	.43	.181	71.16	.001	.54	.107	23.89	.001	.63	1937
Mean	**.447**	**.200**			**.550**	**.103**			**.66**	

[a] Degrees of freedom lost at: Step 1, 12 *df;* Step 2, 24 *df.*

[b] All *p*-values <.001 were rounded to .001.

[c] Ratio of the Coefficient of Determination at Step 1 to the Coefficient of Determination at Step 2 (i.e., the ratio of R^2 at Step 1 to R^2 at Step 2) for each model.

[d] Ratios for the Adjusted R^2 values were computed using the Wherry formula (Wherry, 1931; Yin & Fan, 2001). Adj $R^2 = 1 - (1\text{-}R^2)\frac{n-1}{n-p-1}$, where *n* = sample size and *p* = number of predictors (i.e., independent variables).

Discussion

This research was based on several critical insights. First, we recognized from earlier PSM analyses that the shape metric can be leveraged to evaluate validity gains associated with the use of alternate keying procedures for rating-based scales. In contrast, validity estimates based on singular, global measures (e.g., distance, dot product and proportion correct scores) are confounded by variance associated with individual response rating tendencies that may limit scale validity. In addition, the simultaneous use of the shape, scatter and elevation metrics can be applied to elevate scale validity beyond the use of pure shape metrics (Legree et al., 2018).

Second, PSM and empirical analyses have demonstrated that conventional approaches to keying personality scales has limited the validity of these instruments. In fact, PSM analyses have mapped conventional dominance keys into the integer keying space and show that up to 30% of the vectors in the integer key space outperform conventional dominance keys (Legree et al., 2021). In addition, conventional dominance keying has been the primary method used to key personality scale keys for nearly 90 years (Likert, 1932; Goldberg, n.d.). However, this method often results in keying vectors being based on item writer opinion, as opposed to developing and utilizing more systematic keying methods. From this perspective, it appeared certain to us that much better keying procedures can be developed to improve the validity of many of these scales.

Third, we recognized that a keying space exists for all rating-based instruments that can be partially explored by searching the universe of integer based keys for each scale or by evaluating a large number random vectors for each scale. To be clear, these methods cannot identify the very best possible key because the best possible scoring vector would likely contain real numbers as opposed to integer values, and the Random Vector key space for any rating-based scale is infinite. Nonetheless, these methods are likely to provide much improved keys if only because conventional dominance keys can be highly limited.

Fourth, we recognized from the empirical keying literature that using the dot-product formula to compute scale scores often provides substantial improvements in personality scale validity over alternate empirical methods (Cucina et al., 2019). However, careful inspection and consideration of this formula indicates that the resulting dot product scores may confound scatter and elevation effects in ways that limit scale validity, much like distance and proportion correct scores. Therefore, we proposed using PSMs to model dot product score variance, and we proposed using the scatter and elevation metrics to enhance validity results for dot product scores.

Fundamental Findings

Using a cross-validation strategy designed to identify potential keying vectors, we demonstrated that by carefully keying personality scales and using methods that disentangle the effects of shape, scatter and elevation, extremely high validity estimates can be demonstrated for a large battery of personality scales, $R = .61$. This result matches and even exceeds the validity estimates of measures of general cognitive ability, which are often considered unique by meeting

traditional standards for "large" psychological effects[2] (Schmidt & Hunter, 1998; Cohen, 1988; Ree & Earles, 1991).

In addition, increases in scale validity were observed for each scale, albeit with larger increases being obtained for scales with higher validity – as opposed to documenting general validity increases that may duplicate validity gains at the scale level. This pattern of results indicates that scale content is critical to obtaining these validity improvements.

We also demonstrated that very high battery validity estimates can be obtained using 11 scales that were developed for seven predictor constructs, $R = .58$. This result strongly suggests that scale coverage for a limited number of personality constructs is required to predict outcome measures well, as opposed to endlessly searching for new constructs that may add incremental validity to established measures, as has been common in research settings.

Finally, our results broadly confirm previous demonstrations that shape, scatter and elevation metrics are critical to optimizing scale validity estimates, as opposed to focusing on only the shape metric (Legree et al., 2021). Our interpretation of this general result is that the elevation and shape metrics are useful because they quantify the intensity of behavioral tendencies related to the underlying personality scales.

Practical Guidance

Given the equivalence of validities across scoring procedures, we generally recommend computation of dot product, elevation, and metrics that would then be optimized through regression procedures for future research projects. However, other considerations may merit more careful consideration. For example, highly motivated researchers might consider modifying the item criterion vectors by using them to create variations on these vectors that may provide higher validity estimates.

On the other hand, researchers might consider defaulting to a far easier approach by essentially following the model used to optimize distance score validity. (Refer to the less impressive battery validities obtained at Step 2 for the hierarchical regressions that entered distance scores at Step 1 and the scatter and elevation metrics at Step 2. At least this approach is easy to implement.) This procedure could be legitimately recommended for relatively small applications for which the estimation of item criterion correlations is problematic (e.g., when only limited data are available).

To assist in the production of scales batteries to serve as marker variables for the development of alternate personality batteries, our analyses identified six specific personality constructs that should be included in research projects intended to create and validate personality scales for USACC cadet populations. Validity estimates for this short battery, $R = .53$, greatly exceed validity estimates for most personality measures. We also believe that scale coverage for these constructs should be broadly included in the development of personality scales for a broad range of research intended to predict officer performance.

[2] This assertion is not a direct comparison because our results used regression weightings. However, analyses associate nearly all the predictive validity of cognitive ability batteries with Psychometric g (Ree & Earles, 1991).

Future Research

Disparate Impact

These analyses were entirely focused on enhancing predictive validity and did not explore disparate impact issues. Therefore, this issue should be explored in the future, either through conducting additional analyses on this dataset, or by analyzing data as they become available for personality scales being used for operational personnel purposes.

We caution that rating-based personality scales carry minimal adverse impact (Hogan, 2005; Hough & Oswald, 2000; Ones & Anderson, 2002; Putka, 2009; Schmidt & Hunter, 1998). Therefore, improvements to personality scale validity may increase impact in a manner analogous to the general observation that less reliable measures of general cognitive ability are associated with reduced, yet lower predictive validity. However, results from the analysis of rating-based situational judgement tests indicate that the use of shape scores to compute scale scores can simultaneously improve scale validity and reduce adverse impact (McDaniel et al., 2011).

Response Format

Although higher validities were observed for the CBEF (5-point) scales over the CPM (9-point) scales, we should caution that the CBEF scales had been refined and improved over several decades. In contrast, the CPM scales had never been systematically refined. Therefore, validity advantages associated with the CBEF scales likely resulted from decades of refinement designed to improve the CBEF validity. From this perspective, the CPM format may represent a viable method to expand the range of psychological methods that can be leveraged to predict human performance.

REFERENCES

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

Cucina, J. M., Vasilopoulos, N. L., Su, C., Busciglio, H. H., Cozma, I., DeCostanza, A. H., Martin, N.R., Shaw, M.N. (2019). The effects of empirical keying of personality measures on faking and criterion-related validity. *Journal of Business and Psychology, 34,* 337–356.

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

Graves, C. R., Green, J. P., & Young, M. C. (Eds.) (2021). *Research on the Cadet Background and Experience Form (CBEF) to support Army ROTC personnel assessment (2019-2020)* (Technical Report 1398). U.S. Army Research Institute for the Behavioral and Social Sciences: Fort Belvoir, VA.

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.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org.

Legree, P.J., Kerner, B. S. Shewach, O.R. (2021). Identifying Optimal Keys to Enhance Personality Scale Validity (Research Note 2021-04). U.S. Army Research Institute for the Behavioral and Social Sciences: Fort Belvoir, VA.

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., Ness, A. M., Kilcullen, R. N., & Koch, A. J. (2018). *Enhancing the Validity of Rating-Based Tests* (Technical Report 1371). Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences.

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., Purl, J., Kilcullen, R. N., & Young, M. C. (2019). Cadet Training and Personality Metrics Longitudinally Predict Officer In-unit Performance: *R* = .37. (ARI Technical Report 1377). Fort Belvoir, VA: U.S. Army Research Institute.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5-55.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, *75*, 255-276.

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.

Putka, D. J. (2009). *Initial development and validation of assessments for predicting disenrollment of four-year scholarship recipients from the Reserve Officer Training Corps* (Study Report 2009-06). Arlington, VA: U.S. Army Research Institute for the Behavioral Sciences.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g. Personnel Psychology, 44,* 321-332.

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.

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

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *The Annals of Mathematical Statistics*, *2*, 440-457.

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, *69*, 203-224.