Cross Validated Temperament Scale Validities Computed Using Profile Similarity Metrics

Peter Legree, Robert N. Kilcullen and Mark C. Young

U.S. Army Research Institute, Fort Belvoir, VA

Peter.J.Legree.civ@mail.mil

Paper presented on 27 April 2017 at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
May 2017	Final	April 2016 – May 2017

4. TITLE AND SUBTITLE
Temperament Scale Validities Computed Using Profile Similarity Metrics

5a. CONTRACT NUMBER
W5J9CQ-12-C-0004

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S):
Peter J. Legree, Robert N. Kilcullen and Mark C. Young

5d. PROJECT NUMBER
311

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

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

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

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

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

13. SUPPLEMENTARY NOTES ARI Research POC: Dr. Peter J. Legree, Personnel Assessment Research Unit. Conference paper was delivered at the 2017 SIOP Conference in Orlando, Florida.

14. ABSTRACT

Personality and temperament scales are used in employment settings to predict performance because they are valid and have minimal adverse impact. This project investigated the use of profile similarity metrics (PSMs) in place of conventional distance-based indices to develop scale and composite scores for a battery of temperament scales. Using a sample of 5,191 ROTC cadets, we computed the following PSMs for six temperament scales: the shape of each respondent's rating profile relative to the key, $r_{x,k}$; the difference in elevation between each respondent rating profile and the key, $(X_{mean} - K_{mean})^2$; and profile rating scatter, sd_x^2. We then used regression procedures to develop optimally weighted PSM-based scores for each temperament scale and for the battery. Using a second sample of 5,720 ROTC cadets, we cross-validated the PSM scale and composite scores. Analyses documented that the cross-validated PSM scores maintained higher criterion validities for five of the six temperament scales. Furthermore, the cross validated battery composite based on the PSM scores had higher validity than the corresponding composite based on conventional scores (r = .41 vs. r = .32). These results demonstrated that PSMs can be used to increase scale validity of temperament scales against important performance criteria. Presented at the SIOP Conference held in Orlando, FL., April 27 2017 - April 29 2017.

15. SUBJECT TERMS
Personnel Selection, Profile Similarity Metrics, Temperament, Personality

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Tonia Heffner
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	21	19b. TELEPHONE NUMBER (703) 545-4408

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Cross Validated Temperament Scale Validities

Computed Using Profile Similarity Metrics

Abstract

Personality and temperament scales are used in employment settings to predict performance because they are valid and have minimal adverse impact. This project investigated the use of profile similarity metrics (PSMs) in place of conventional distance-based indices to develop scale and composite scores for a battery of temperament scales. Using a sample of 5,191 ROTC cadets, we computed the following PSMs for six temperament scales: the shape of each respondent's rating profile relative to the key, $r_{x,k}$; the difference in elevation between each respondent rating profile and the key, $(X_{mean} - K_{mean})^2$; and profile rating scatter, sd_x^2. We then used regression procedures to develop optimally weighted PSM-based scores for each temperament scale and for the battery. Using a second sample of 5,720 ROTC cadets, we cross-validated the PSM scale and composite scores. Analyses documented that the cross-validated PSM scores maintained higher criterion validities for five of the six temperament scales. Furthermore, the cross-validated battery composite based on the PSM scores had higher validity than the corresponding composite based on conventional scores (r = .41 vs. r = .32). These results demonstrated that PSMs can be used to increase scale validity of temperament scales against important performance criteria.

Cross Validated Temperament Scale Validities

Computed Using Profile Similarity Metrics

Temperament scales are widely used to predict performance because they have moderate validity and minimize adverse impact (Hogan, 2005; Hough & Oswald, 2000; Ones & Anderson, 2002). We suggest that one approach to enhance temperament scale validity is through the application of profile similarity metrics (PSMs). We begin by summarizing the rationale for our expectations that using PSMs as opposed to conventional algorithms to compute temperament scale scores will enhance the validity of these instruments.

We then describe a cross-validation design used to explore the efficacy of this approach. We utilized temperament data collected in 2013 as the developmental sample in order to optimize PSM-based scale scores for six temperament dimensions against important Army criteria. We used regression procedures to develop PSM weighting algorithms. We then used temperament data collected in 2015 to cross-validate the optimized PSM composite scales for the six temperament scales.

Finally, we used the cross validation sample to assess the validity gains achieved by optimally weighting either the conventional scale scores for the six temperament scales, or the PSM composite scores for the six temperament scales. Both the 2013 development and the 2015 cross-validation samples had temperament and performance data collected from approximately 5,500 U.S. Army ROTC cadets who later became U.S Army commissioned officers.

Conventional Scores Measure Distance

In previous papers, we have observed that most personality and temperament scales use a conventional scoring algorithm that computes a respondent's scale score as the mean item rating with some of the item ratings corrected for the direction of the item. For example, a "Fitness

Motivation (FM)" scale might contain items that indicate either a tendency to participate in physical activities (non-reversed), or a tendency to avoid physical activities (reversed). An example of a reversed Fitness Motivation item indicating a tendency to participate in physical activities might be written in this form: *"In school, I avoided sports: 1. Never; 2. Seldom; 3. Occasionally; 4. Frequently; 5. Often."* For reversed items, a respondent's score would be recoded so that higher values would indicate superior standing on Fitness Motivation. Therefore, conventional item scores near "5" will indicate higher standing on the underlying construct for both reversed and non-reversed items. Finally, a respondent's scale score is equal to the mean of the non-reversed and recoded-reversed items. Table 1 portrays the conventional scoring algorithm on the left, by showing raw and corrected item ratings and the scale score for a respondent across six items that use a five point scale, with half of the items reversed.

In previous papers, we have also observed that conventional temperament scores are more accurately described as "transformed distance metrics." Table 1 demonstrates this equivalence by presenting: (1) conventional scores that were computed using the conventional scoring algorithm, and (2) item distance scores. Unlike conventional scores, distance scores do not need to be corrected for reversal because they are computed as the absolute value of the difference between the respondent rating and the scoring key. The scoring key for distance scores is comprised of extreme values so that reversed items are keyed as "1," while non-reversed items are keyed as "5." Similarly to conventional scoring, a scale score is derived by averaging the item distance scores. However, superior standing on the construct is indicated by item and scale distance scores approaching "0."

The final column of Table 1 shows that the conventional scores and distance scores will always sum to the same value, "5." This statement is true at both the item and the scale level.

Moreover, the correlation between conventional scores and distance scores for these types of scales have a perfect negative correlation, r = -1.00. From this perspective, conventional and distance scores are completely redundant. Therefore, we argue that conventional temperament scores are properly viewed as "transformed distance metrics."

Profile Similarity Metrics (PSMs) and Research Expectations

Recognizing that conventional temperament scores constitute a distance metric is an important insight because distance metrics can be better understood and analyzed using a PSM perspective (Cronbach & Gleser, 1953; Legree et al., 2014). In order to compute PSMs, each set of respondent raw ratings must first be conceptualized as a "rating profile." The three primary PSMs correspond to:

1. Rating shape, which is computed as the correlation between each respondent's rating profile and the scale key: shape-scores = $r_{x,k}$.
2. Rating elevation difference, which is computed as the squared difference between the each respondent's mean scale rating and the mean value in the scale key: elevation difference scores = $(X_{mean} - K_{mean})^2$.
3. rating scatter, which is computed as the variance of each respondent's rating profile: scatter scores = sd_x^2.

As detailed below, we analyzed two datasets that were collected in 2013 (developmental sample) and 2015 (cross-validation sample). Regression analyses conducted with the development sample verified that:

1. Scale-level PSMs account for most of the variance in the conventional scale scores for each of the ten temperament scale that could be analyzed (i.e., all $R^2 > .91$). This

result is consistent with PSM formulaic derivations (cf., Cronbach & Gleser, 1953; Legree et al., 2014).

2. Scale-level PSMs add incrementally to the prediction of criterion variance beyond conventional scores for most temperament scales. This conclusion indicates that conventional scores may represent a sub-optimized composite of shape, elevation difference, and scatter effects.
3. Consensually derived scoring keys can be used to compute alternate shape scores for temperament scales that may add incrementally to the prediction of criterion variance beyond PSMs computed using conventional standards. This conclusion indicates that consensually derived keys may constitute a better scoring key to compute shape scores than the conventional scoring keys for some temperament scales.

It should be clarified that consensual keys are computed as the mean participant item rating for each item and have provided excellent standards to score rating-based situational judgement tests (Legree et al., 2014). Moreover, consensual keying is not equivalent to empirical keying because criterion information is not used to develop consensual keys (Legree, Psotka, Tremble & Bourne, 2005). Therefore, we expect that the cross-validation of temperament scale composites based on consensual keys will be less prone to shrinkage as is commonly observed using empirical keying procedures (Schwab & Oliver, 1974).

Cross-Validation Approach

To cross-validate and extend the previous findings, we identified seven temperament scales that had been administered to the "development sample" in 2013 and the "cross-validation sample" in 2015. We also obtained ROTC performance outcome data for cadets in both samples. Performance outcomes included: cadet Order of Merit List (OML) scores that reflect overall

performance in the ROTC program, Army Physical Fitness Test (APFT) scores, and college GPA.

Analyses conducted using the 2013 development sample indicated that: (1) PSMs add incremental validity to the conventional scores for the prediction of performance for six of the seven scales; and (2) shape-consensus scores add incremental validity to the PSMs for the prediction of performance for five of those six scales.

Before cross-validating these results, we used the development sample to create: (1) composite scale scores that optimally weighted the PSMs for the six scales for which PSMs incremented the conventional score validity estimates; and (2) consensual standards for the five scales for which shape-consensus scores added incrementally to the PSMs. To minimize multicollinearity issues during the cross-validation analyses, we used either shape scores based on the consensus standard, or shape scores based on the conventional key for each temperament scale. In addition to the shape scores, we used scatter scores and the elevation-difference scores to compute PSM composite scores for each temperament dimension using the development sample.

We then applied the regression weights and used the consensus standards obtained from the development sample to compute PSM composite scores for the cross-validation sample. Next, we compared the validities of the PSM composite and conventional scores for each scale. We tested the difference between the magnitude of corresponding validity estimates using the Fischer procedure to test the difference between correlation estimates obtained using a common sample (Steigler, 1980). We reasoned that the cross-validated composites would outperform the conventional scores for the six temperament scales.

Finally, we used regression procedures in the cross-validation sample to compare the validity of temperament composite scores that were derived from: (1) the conventional scores for the six temperament scales; and (2) the cross-validated, PSM composite scores for the six temperament scales.

Method

Participants

The development sample consists of 5,191 ROTC cadets who participated in the Leadership Development and Assessment Course (LDAC), during the summer of 2013 in Fort Lewis, WA. The cross-validation sample consists of 5,711 ROTC cadets who participated in the Leadership Development and Assessment Course (LDAC), during the summer of 2015 in Fort Knox, KY. Incomplete performance data limited the sample sizes for the cross-validity estimates to approximately 4,900 for each scale.

From a demographic perspective, the development and cross-validation samples were very similar. Both samples were primarily male, 78%. Individuals in the two samples identified as: Caucasian, 82%; African-American, 11%; Asian, 7%; American Indian or Alaskan Native, 2%; and Native Hawaiian or other Pacific Islander, 1%. In addition, 12% of the sample identified their ethnicity as Hispanic. Approximately 2% of the sample did not identify their race or ethnicity.

Measures

Cadet Background and Experiences Form (CBEF). The CBEF is a 139 item multiple-choice questionnaire assessing past behaviors and experiences, which is designed to predict officer performance and retention (Putka, 2009). We reviewed the CBEF and found seven scales with reversed and non-reversed items that were administered to the development and cross-

validation samples: Written Communication (WC), Fitness Motivation (FM), Army Identification (AI), Stress Tolerance (ST), Tolerance for Injury (TI), Past Withdrawal Propensity (PWP), and Oral Communication (OC). CBEF scale definitions and the ratio of the number of reversed items to the total number of items for each scale are summarized in Table 2. All items within the CBEF scales are measured using a 5-point Likert scale.

Criteria. Training and performance data were collected from the LDAC training cadre and represent a mix of training performance (e.g., Army Physical Fitness Test scores) and supervisor ratings. These data are then combined with other ROTC performance metrics to compute a cadet Order of Merit List (OML) that reflects the cadets' overall performance in the ROTC program. Therefore, the CBEF has traditionally been validated against OML. However, we also validated the temperament scales against performance on the Army Physical Fitness Test (APFT) and College GPA because these outcomes are important to the development of U.S. Army ROTC cadets (Putka, 2009).

Procedure

For both datasets, the Army ROTC cadets were administered the CBEF as a part of a battery of paper and pencil tests during their initial week at LDAC. The outcome criteria were then collected from the U.S. Army Cadet Command after the cadets had completed LDAC. Therefore, a longitudinal design was used to validate the CBEF scales against the OML, APFT, and GPA criteria. The same data collection procedure was followed for the development sample in 2013, and the cross-validation sample in 2015.

Results

We first confirmed that conventional scores for each of the seven temperament scales represent a PSM composite formed by regressing the scale scores onto the shape, scatter, and

elevation-difference metrics, all $R^2 > .92$. This result confirms that conventional scale scores can be viewed as a PSM composite for each of the temperament scales. Summary statistics are reported in Table 3 for the developmental sample.

We used the developmental sample to optimally weight the PSMs for each temperament scale and thereby create PSM composite scores for each scale. We subsequently applied those weights to score temperament data that had been collected two years later for the cross-validation sample. Correlation analyses among the outcome metrics, PSM composite scores, and the conventional scale scores indicated that the PSM composite scores continued to have greater validity against the outcome metrics even though the cross-validation sample was collected two years later.

Specifically, we regressed the OML performance scores onto the following scale metrics and examined the change in ΔR^2 for significance: conventional scale score in step 1; shape, scatter and elevation-difference metrics in step 2; shape consensus metric in step 3. The results documented that the PSMs resulted in a significant improvement in model fit for six of the seven scales at steps 2 or 3. The regression analyses also indicated that the shape-consensus scores added incremental variance against the OML performance outcome for five of the seven scales. These results were computed for the developmental sample and are detailed in Table 4, columns 1 through 6.

To limit multicollinearity effects and simplify PSM interpretations, we computed optimal PSM composite weights by regressing the OML performance outcome scores onto the "Best Three PSMs" for each scale. Functionally, this set included: shape-consensus, scatter, and elevation-difference for five scales; and shape-conventional, scatter, and elevation-difference for the sixth scale. The *Multiple Rs* and associated regression weights are provided Table 4,

columns 7 through 10. Comparison of the corresponding coefficients in Table 4, columns 5 and 7, documents that little predictive information was lost by limiting the PSM scale composite to the "Best three PSMs" (through regression weighting procedures) for each temperament scale.

We then applied the PSM regression weights from the developmental sample to compute composite scores for the cross-validation sample. We then correlated the conventional and PSM composite scores for each of the six temperament scales with the OML, GPA and APFT criteria. We used the Fischer procedure to test the difference between the corresponding conventional and PSM composite scale validities for each outcome. PSM composite validity gains against OML were maintained for 5 of the 6 temperament scales, and three of the validity gains were statistically significant and substantial. Gains were also observed for the GPA and APFT criteria. Results are detailed in Table 5.

Finally, we used regression procedures to assess the predictive validity of overall composites that were based on either the six conventional scale scores or the six PSM composite scale scores. We first regressed OML on the six conventional scales in step 1, and the six PSM scales in step 2 to determine whether the PSM scales add to the conventional scales. As detailed in Table 6, the change statistics were significant and increase in *R* was substantial (.32 vs. .43).

We also regressed OML on the six PSM scales in step 1, and the six conventional scales in step 2, to determine whether the conventional scales add to the PSM scales. As detailed in Table 7, the change statistics were significant, but the increase in *R* was minimal (.41 vs. .43). Therefore, the use of PSM composite scores to score the temperament scales was associated with higher levels of validity, $R = .41$, than the conventional scales, $R = .32$.

In summary, the cross-validation analyses indicated that the validity gains associated with the PSM-based scoring algorithm were maintained at both the individual scale level for six

temperament scales and at the composite level for the temperament battery. These results confirm that PSMs may be used to substantially increase validity at the scale level, and at the composite temperament scale level.

Discussion

In this paper we showed that conventional temperament scores are accurately conceptualized as a distance metric. We then demonstrated that nearly all the variance associated with conventional scores computed for seven temperament scales can be accounted for by a linear combination of PSMs: shape, elevation-difference, and scatter metrics.

These insights are important because they indicate that conventional temperament scales may represent an arbitrary and therefore suboptimal mix of these effects. We used the development sample to assess the extent to which the temperament scale validity may have been underestimated through the use of conventional/distance scores. Analyses conducted with the development sample indicated that this limitation is common and may be widespread. More importantly, the bivariate analyses indicated that the use of distance scores may have limited the validity estimates for six of the seven temperament scales

Cross-validation analyses, which were conducted using an independent sample, demonstrated that validity gains associated with the PSM composites were maintained two years later at the scale level. From an applied perspective, substantial gains associated with several of the temperament scales were documented. The most substantial validity gains were obtained for: the Written Communication scale, .30 versus .19; the Fitness Motivation scale, .31 versus .27; and the Stress Tolerance scale, .16 versus .14. It follows that the validity estimates computed for a broad array of personality and temperament have likely been underestimated by widespread practices favoring the use of distance-based metrics for temperament and personality scales.

In a direct comparison of the validity of personality scale composite that are based on either conventional scale scores or PSM scale scores, the regression analyses associated much higher levels of validity with composites based on the PSM scale scores, $R = .41$, than for composites based on the conventional scale scores, $R = .32$. This result extends the scale level conclusion to indicate that validity estimates associated with personality and temperament scales have been consistently and substantially underestimated through the widespread use of distance metrics to score personality and temperament scales.

In previous papers we have speculated that ensuring a mix of reversed and non-reversed items within each scale is important to enhancing the validity of temperament scales that are scored using PSMs. We also speculated that limitations with conventional keys may limit the validity of temperament scales. The current analyses support these contentions by demonstrating the inclusion of shape-consensus scores were critical to boosting the validity of the temperament scales. Finally, we note that the use of PSMs to score temperament scales greatly complicates attempts to improve scores through response distortion strategies because the effective use of these strategies would require advanced understandings of the statistical issues that underlie the rationale for PSMs to score personality and temperament scales.

References

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2010). *Applied multiple regression/correlation analysis for behavioral sciences.* New York: Taylor & Francis.

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

Grant, A., & Shwartz, B. (2011). Too Much of a Good Thing: The Challenge and Opportunity of the Inverted U. *Perspectives on Psychological Science, 6,* 61-76.

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

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

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

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

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

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

McDaniel, M., Psotka, J., Legree, P., 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.

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

Putka, D.J. (Ed.). (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.

Schwab, D.P., & Oliver, R.L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. Personnel Psychology, 27, 125-128.

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

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

Table 1

Conventional and Distance Scoring Algorithms

	Conventional Scoring		Distance Scoring			Sum
	Respondent Rating	Corrected Rating	Respondent Rating	Key	Distance Score	
Item 1 (Non-reversed)	4	4	4	5	1	5
Item 2 (Non-reversed)	5	5	5	5	0	5
Item 3 (Non-reversed)	3	3	3	5	2	5
Item 4 (Reversed)	2	4	2	1	1	5
Item 5 (Reversed)	1	5	1	1	0	5
Item 6 (Reversed)	2	4	2	1	1	5
Scale Score (Mean):		4.17			0.83	5.0

Table 2

CBEF Scales, Definitions, Ratio of Reversed Items to Total Number of Items[a]

Ratio: Reversed Items/Total Number of Items	Scale	Definition
3/8	Past Withdrawal Propensity (PWP)	Degree of commitment and continuance in groups
4/11	Oral Communication (OC)	Degree of comfort with oral communication
5/14	Fitness Motivation (FM)	Degree of enjoyment from physical exercise and willingness to stay in shape
2/7	Written Communication (WC)	Degree of comfort with written communication
1/5	Tolerance for Injury (TI)	Degree of enjoyment from risky and hazardous activities
1/11	Stress Tolerance (ST)	Degree of emotional control and composure under pressure
1/12	Army Identification (AI)	Degree of identification with, and interest in being, a U.S. Army Soldier

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

Table 3

Conventional Temperament Scale Scores Regressed on PSMs[a]

CBEF Scale (Item Ratio)	R^2	Shape		Elevation-Delta		Scatter	
		β	*(r)*	*β*	*(r)*	*β*	*(r)*
WC (.29)	.96	.80	.94	-.26	-.63	.13	.31
FM (.36)	.97	.65	.91	-.15	-.43	.39	.80
AI (.08)	.95	.41	.86	-.64	-.93	.05	.17
ST (.09)	.97	.36	.72	-.76	-.93	.08	.10
TI (.20)	.92	.67	.88	-.42	-.74	.1	.10
PWP (.38)	.97	.65	.85	-.19	-.17	-.51	-.76
OC (.36)	.93	.69	.89	-.02	-.21	.44	.74

[a]: Model Statistics: (df = 3, 5131-5187), (All F-statistics > 16760), all models and coefficients significant at $p < .001$.

Table 4

Developmental Sample: LDAC Criteria Regressed on Conventional, PSM, and Shape-consensus Scores by Temperament Scale[a]

Scale	Conventional Distance (Step 1)	PSM (Step2)		Shape-consensus (Step 3)		Best 3 PSMs					
	R	*R*	ΔR^2	*R*	ΔR^2	*R*	Shape		Elev-Delta		Scatter	
							B	(β)	B	(β)	B	(β)
Regression Models for OML												
WC	.169	.250	.034	.333	.048	.325	7.63	(.23)	-.69	(-.06)	2.17	(.20)
FM	.284	.304	.012	.325	.014	.312	4.18	(.16)	-4.76	(-.14)	1.19	(.11)
AI	.005^{ns}	.081	.007	.097	.003	.097	3.17	(.08)	.57	(.07)	.99	(.06)
ST	.097	.113	.003	.114	.000^{ns}	.113	2.71	(.09)	-.22	(.038)	-.41	(-.03)^{ns}
TI	.067	.073	.001^{ns}	.093	.003	.090	1.56	(.08)	-.23	(-.02)^{ns}	-.25	(-.02)^{ns}
PWP	.034	.054	.002^{ns}	.066	.001	.045	-1.05	(-.03)	.56	(.03)	-.08	(-.01)^{ns}
OC	.123	.126	.001^{ns}	.129	.001^{ns}							
Regression Models for APFT												
FM	.480	.500	.020	.519	.019	.510	7.424	(.309)	-6.16	(-.20)	1.50	(.152)

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

^{ns}: Not significant

Table 5

Cross-validated scale validities

Scale	n	PSM by	OML			APFT			GPA		
		Conv	Conv	PSM	ΔSig[a]	Conv	PSM	ΔSig	Conv	PSM	ΔSig
WC	3996	.602**	.187**	.297**	.001	.007	.055**	.001	.217**	.329**	.001
FM	3997	.937**	.267**	.312**	.001	.405**	.430**	.001	.062**	.083**	.001
AI	3997	.022	.027	-.009	.103	.045**	-.038*	.001	-.031	.068**	.001
ST	3984	.838**	.142**	.161**	.033	.086**	.094**	.373	.116**	.112**	.655
TI	3954	.691**	.110**	.116**	.074	.160**	.138**	.075	-.044**	-.014	.016
PWP	3994	.615**	.104**	.115**	.425	.060**	.057**	.829	.046**	.061**	.279

[a]: Followed the Steigler approach (1980) to test the difference between two dependent correlations with one variable in common. Program available at http://quantpsy.org/corrtest/corrtest2.htm
**: $p < .01$ level (2-tailed).
*: $p < .05$ level (2-tailed).

Table 6

Incremental Validity for PSM Scale Scores beyond Conventional Scores Against OML

Model Step	*R*	R^2	Adj R^2	Std. Error of the Estimate	Change Statistics				
					ΔR^2	F Change	df1	df2	Sig. F Change
1. Conventional	0.322	.104	.103	.0151598	.104	81.476	6	4212	.001
2. PSMs	0.428	.183	.181	.0144860	.079	67.821	6	4206	.001

Table 7

Incremental Validity for Conventional Scale Scores beyond PSM Scores Against OML

Model Step	*R*	R^2	Adj R^2	Std. Error of the Estimate	Change Statistics				
					ΔR^2	F Change	df1	df2	Sig. F Change
1. PSMs	0.412	.170	.169	.0145910	.170	143.751	6	4212	.001
2. Conventional	0.428	.183	.181	.0144860	.013	11.210	6	4206	.001