More on Testing for Validity Instead of Looking for It

John Antonakis
john.antonakis@unil.ch
Department of Organizational Behavior
Faculty of Business and Economics
University of Lausanne
Internef 618
1015 Lausanne, Switzerland
Tel.: +41 21 692 3438

Joerg Dietz
joerg.dietz@unil.ch
Department of Organizational Behavior
Faculty of Business and Economics
University of Lausanne
Internef 617
1015 Lausanne, Switzerland
Tel.: +41 21 692 3682

In press:

Personality and Individual Differences

More on Testing for Validity Instead of Looking for It

Using Monte Carlo simulations and reanalyzing the data of a validation study of the AEIM emotional intelligence test, we demonstrated that an atheoretical approach and the use of weak statistical procedures can result in biased validity estimates. These procedures included stepwise regression—and the general case of failing to include important theoretical controls—extreme scores analysis, and ignoring heteroscedasticity as well as measurement error. The authors of the AEIM test responded by offering more complete information about their analyses, allowing us to further examine the perils of ignoring theory and correct statistical procedures. In this paper we show with extended analyses that the AEIM test is invalid.

Keywords: incremental validity; control variables, emotional intelligence; general intelligence; heteroscedasticity; measurement error; validation; psychometrics; truncation.

Introduction

Using Monte Carlo simulations and an empirical example (Warwick, Nettelbeck, & Ward, 2010a) we recently demonstrated that an atheoretical approach and the use of poor statistical procedures can result in biased validity estimates (see Antonakis & Dietz, in press). In our study we showed that the Ability Emotional Intelligence Measure (AEIM) was invalid and should not be used for assessment purposes. Warwick et al. (in press) responded by offering more complete information about their analyses.

Their response was forthcoming and candid, explaining their choice of regression models and variables; we are also pleased to see in the meantime that the AEIM has been "withdrawn from use" (Warwick, Nettelbeck, & Ward, 2010b, p. 152). In their response, Warwick et al. (in press) (a) agreed that testing incremental validity is a "theory-driven endeavor" that should include established control variables, (b) stated that they "incorrectly used" standard statistical terminology (i.e., "stepwise" regression when they meant "hierarchical"), (c) mentioned that they had "not well understood" nor foreseen the problems that heteroscedasticity and measurement error can cause for estimating validity, and (d) acknowledged that they "overstated" their case (p. XXX).

Despite the evidence showing that the AEIM is invalid, the response by Warwick et al. (in press) unfortunately also attempted to justify the use of statistical analyses that produce uninterpretable findings. Given that we now know the precise models used in their attempts to validate the AEIM, we can add to our first article illustrating even more clearly how the lack of theory and flawed analyses undermine validation studies. Although we were glad to hear that Warwick et al. (in press) did not use stepwise regression, the simulations we published revealed the problems with this exploratory regression procedure and should serve as a stern warning to researchers who may consider using it.

In our rejoinder, we address two major aspects of Warwick et al.'s (in press) response that merit further consideration: (a) the lack of theory in the choice of control variables, and (b) flawed analyses due to the use of extreme scores.

On choosing control variables: Theory, theory, and more theory

We see validity testing as a theory-driven endeavor such that when testing the incremental validity of a new construct with respect to dependent outcomes, one would inquire about existing theories that also explain this outcome. For example, when predicting loneliness from emotional intelligence, one would review theories of loneliness. These theories include dispositional approaches; and, it appears obvious that among dispositional predictors, neuroticism and extraversion are the most likely candidates (cf. Lasgaard, 2007). For example, extraverts are motivated to seek-out the company of others and are gregarious, and usually have a larger network of friends. Thus, they should be less lonely than introverts (see review by Heinrich & Gullone, 2006). In light of these arguments, correlations larger than .50 reported in several studies are not surprising (Burger, 1995; Lasgaard, 2007; Lasgaard & Elklit, 2009).

What is surprising to us is that Warwick et al. (2010) state that incremental validity is a theory-driven exercise, yet they did not follow through when they examined the incremental validity of their emotional intelligence test; for example, they did not theorize extraversion as a predictor of loneliness. Instead, Warwick et al. claimed that "statistically significant personality variables [were] neuroticism and conscientiousness" (in press, p. XXX), although the latter variable correlated only r = -.12 with loneliness; however, the two strongest personality correlates of loneliness were, as theory would suggest, neuroticism (r = .55) and extraversion (r = -.49) (see Table 1 in Antonakis & Dietz, in press).

Estimating a model where only neuroticism and conscientiousness are used along with intelligence while ignoring measurement error and heteroscedasticity gives the allure of incremental

validity for the consensus score. However, this coefficient is confounded given that there are important omitted variables in the regression model (Antonakis, Bendahan, Jacquart, & Lalive, in press).

[Insert Table 1 here]

We initially estimated the model precisely as they did in their original paper and now as explicitly reported by Warwick, et al. (in press). We added the variables in three hierarchical blocks: Cognitive ability (block 1), neuroticism and conscientiousness (block 2), and emotional intelligence (EI) consensus and confidence scores (block 3). We then determined whether the *r*-square change in block 3 was significant, as an indication of the incremental validity of their measures. Warwick et al. (in press) reported a significant regression coefficient for the consensus score. However, they did not report that the *r*-square change in the last block—the simultaneous test that the two EI coefficients are significantly different from zero—which was marginally significant: $F(2, 266) = 2.96$, $p < .10$ (refer to Table 1, Model 1). This latter finding should already have raised red flags concerning the test's incremental validity. Note, using a heteroscedastic-robust variance estimator, which is required given the nonnormal distribution of the residuals, made this *F*-statistic even weaker: 2.63, $p < .10$; correcting both for measurement error and heteroscedasticity lowered the *F*-statistic further, 1.74, $p > .10$. Thus, even using the Warwick et al. (2010) specification, the *r*-square change was not significant, showing that the AEIM has no incremental validity.

In any case, Model 1 is not interpretable due to an omitted-variable bias. Theoretical arguments prescribe that extraversion should not have been omitted in the regression. Furthermore, as mentioned in Antonakis and Dietz (2010), when considering personality it is appropriate to include all personality controls because even if they are not significant individually, they may be jointly significant. Adding extraversion instead of conscientiousness to the regression model reduced the coefficient of the consensus score to -.10 (from -.13), making the coefficient marginally-significant, $p < .10$ (see Model 2); also, the *r*-square change for adding the two EI scores (i.e., adding block 3) was nonsignificant, $p >$

.10. Estimating a model where empathy—another theoretically relevant predictor of loneliness—was added reduced the coefficient of the consensus score to -.06, $p > .20$; again, the *r*-square change for the two EI coefficients was non-significant, $p > .40$ (Model 3). Finally, adding all the personality controls (Model 4) reduced the coefficient of the consensus score to -.04, $p > .50$ (and again, the *r*-square change for the two coefficients was non-significant). The correct model (Model 5, which we already reported in Antonakis and Dietz, in press) that uses an errors-in-variables regression model with bootstrapped standard errors shows that the EI scores remain nonsignificant.

Extreme-scores analyses: Dividing and conquering leaves the empire in shambles

Warwick et al. (in press) continue justifying the truncation of samples and the use of extreme scores for establishing validity although doing so makes it impossible to establish population estimates. In their original article they had referred, in *ipse dixit* fashion, to Petrides, Frederickson, and Furnham (2004), who had formed extreme groups in their study of emotional intelligence. Even in Petrides et al.'s article, however, it was indicated that for inferential statistics extreme-groups analyses must not be used; they noted expressly that these results were for informational purposes adding a *caveat emptor* that these results should "*not* be extrapolated" (p. 285). Furthermore, we cited Heckman (1979) and Tobin (1958)—Nobel prize winners in economics—who developed methods to recover population estimates in the presence of naturally-occurring truncation (suggesting that truncation is an undesirable property in samples, which should be avoided at all costs). We showed too that, in addition to reducing power, truncating a sample into small groups (i.e., from between $n = 22$ to $n = 41$) creates artificial variation, capitalizes on chance, and produces highly suspect estimates (Antonakis & Dietz, in press).

In their response to our first article, Warwick et al. (in press) now suggest that because their instrument was "demonstrably blunt" and the distribution of the AEIM scores was "non linear," it was thus "reasonable" to estimate regression models in groups of extreme scores (p. XXX). This argument is a *non sequitur*; non-linearity is not a justification for extreme scores analysis. Instead, if theoretical

reasons exist to assume curvilinearity, one would add a quadratic (and if needed cubic) term to model this curvature (Aiken & West, 1991). Although neither theory nor visual inspection of their data suggested a curvilinear relation, we estimated a model with a quadratic term, which proved to be nonsignificant when we used OLS regression (and the estimator did not converge when we employed errors-in-variables regression due to the relatively low reliability of the EI consensus score). Finally, supposing Warwick et al. (in press) were referring to problems with outliers instead of nonlinearities, we also estimated the full model using Huber's robust regression (Huber, 1964). Again, the EI coefficients were neither individually nor jointly significant when using the full specification. In summary, validation studies require that scientists arrive at interpretable population estimates; the truncation of data simply has no place in such scientific efforts.

Conclusion

We appreciate that Warwick et al. offered more information on their validation study of their emotional intelligence measure. Our analysis of their data leads us to conclude that it should not be used for diagnostic testing, whether for clinical, industrial, or training purposes. As far as ability-based measures of emotional intelligence are concerned, the invalidity of Warwick et al.'s measure seems representative (see also Amelang & Steinmayr, 2006; Fiori & Antonakis, in press; Føllesdal & Hagtvet, 2009; Schulte, Ree, & Carretta, 2004). We hope that more researchers will join our effort to call for stronger conceptual and methodological standards for developing such measures (Antonakis, Ashkanasy, & Dasborough, 2009; Antonakis & Dietz, 2010).

References

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications.

Amelang, M., & Steinmayr, R. (2006). Is there a validity increment for tests of emotional intelligence in explaining the variance of performance criteria? *Intelligence, 34*(5), 459-468.

Antonakis, J., Ashkanasy, N. M., & Dasborough, M. T. (2009). Does leadership need emotional intelligence? *The Leadership Quarterly, 20*(2), 247-261.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (in press). On making causal claims: A review and recommendations. *The Leadership Quarterly*.

Antonakis, J., & Dietz, J. (2010). Emotional intelligence: On definitions, neuroscience, and marshmallows. *Industrial and Organizational Psychology, 3*(2), 165-170.

Antonakis, J., & Dietz, J. (in press). Looking for Validity or Testing It? The Perils of Stepwise Regression, Extreme-Scores Analysis, Heteroscedasticity, and Measurement Error. *Personality and Individual Differences*.

Burger, J. M. (1995). Individual Differences in Preference for Solitude. *Journal of Research in Personality, 29*(1), 85-108.

Fiori, M., & Antonakis, J. (in press). The ability model of emotional intelligence: Searching for valid measures. *Personality and Individual Differences*.

Føllesdal, H., & Hagtvet, K. A. (2009). Emotional intelligence: The MSCEIT from the perspective of generalizability theory. *Intelligence, 37*, 94-105.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153-161.

Heinrich, L. M., & Gullone, E. (2006). The clinical significance of loneliness: A literature review. *Clinical Psychology Review, 26*(6), 695-718.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35,* 73–101.

Lasgaard, M. (2007). Reliability and validity of the Danish version of the UCLA Loneliness Scale. *Personality and Individual Differences, 42*(7), 1359-1366.

Lasgaard, M., & Elklit, A. (2009). Prototypic Features of Loneliness in a Stratified Sample of Adolescents. *Interpersona, 3*(Suppl.1), 85-110.

Petrides, K. V., Frederickson, N., & Furnham, A. (2004). The role of trait emotional intelligence in academic performance and deviant behavior at school. *Personality and Individual Differences, 36*(2), 277-293.

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional Intelligence: Not much more than g and personality. *Personality and Individual Differences, 37*(5), 1059-1068.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26,* 24-36.

Warwick, J., Nettelbeck, T., & Ward, L. (2010a). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences, 48*(1), 66-71.

Warwick, J., Nettelbeck, T., & Ward, L. (2010b). Erratum to "AEIM: A new measure and method of scoring abilities-based emotional intelligence" [Personality and Individual Differences 48(1) (2010) 66-71]. *Personality and Individual Differences, 49*(2), 152-152.

Warwick, J., Nettelbeck, T., & Ward, L. (in press). A response to Antonakis and Dietz: Looking for Validity of Testing it? *Personality and Individual Differences*.

Table 1: Errors-in-variables Regression Models Predicting Loneliness

	Model 1			Model 2			Model 3			Model 4			Model 5		
VARIABLES	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3
EI Consensus			-.13**			-.10*			-.06			-.04			.08
			(-2.16)			(-1.75)			(-1.08)			(-.68)			(.76)
EI Confidence			-.03			-.03			-.02			-.03			-.05
			(-.63)			(-.62)			(-.44)			(-.60)			(-.48)
Cog. Ability	-.14**	-.04	.04	-.14**	-.10**	-.04	-.14**	-.07	-.04	-.14**	-.09*	-.07	-0.17***	-0.14	-.18
	(-2.36)	(-.75)	(.68)	(-2.36)	(-2.08)	(-.62)	(-2.36)	(-1.52)	(-.62)	(-2.36)	(-1.95)	(-1.14)	(-2.72)	(-1.62)	(-1.37)
Extraversion					-.33***	-.32***		-.31***	-.31***		-.36***	-.35***		-0.48***	-.49***
					(-6.36)	(-6.13)		(-5.99)	(-5.88)		(-6.87)	(-6.75)		(-4.02)	(-3.55)
Neuroticism		.54***	.53***		.40***	.39***		.44***	.43***		.43***	.42***		0.52***	.52***
		(1.58)	(1.23)		(7.47)	(7.32)		(8.08)	(7.79)		(8.06)	(7.80)		(4.56)	(4.13)
Openness											.18***	.18***		0.43***	.45**
											(3.66)	(3.61)		(2.92)	(2.48)
Conscientiousness		-.12**	-.11**								-.05	-.05		0.00	.02
		(-2.26)	(-2.15)								(-1.06)	(-1.04)		(0.01)	(.14)
Agreeableness											-.11**	-.10**		-0.28*	-.31
											(-2.06)	(-1.97)		(-1.68)	(-1.53)
Empathy								-.14***	-.13**		-.15***	-.14**		-0.22**	-.23**
								(-2.97)	(-2.50)		(-2.85)	(-2.53)		(-2.15)	(-2.11)
R-square change	.02	.30	.01	.02	.37	.01	.02	.39	.01	.02	.43	.01	.02	.57	.00
F-test for Δ R-square	-	57.93***	2.96*	-	82.79***	2.04	-	59.75***	.79	-	34.82***	.51	-	20.44***	.70

R-squared	.02	.32	.33	.02	.39	.40	.02	.41	.42	.02	.45	.46	.02	.59	.59

Note: ***p < 0.01, **p < 0.05, *p < 0.10; parameter estimates are standardized; models are estimated in three nested blocks wherein we progressively added variables to the regression model and tested for the significance of the *r*-square change; Models 1-4 are estimated using a standard regression estimator and numbers in parentheses are *t* statistics using a normal variance estimator; Model 5 is the full errors-in-variables regression model with all controls, as reported in Antonakis and Dietz (in press) and (numbers in parentheses are *z* statistics from normal bootstrapped standard errors--findings regarding the AEIM were unchanged when using percentile or bias-corrected bootstraps; N=272.