The Validation Crisis in Psychology

Ulrich Schimmack

University of Toronto Mississauga

April 2019

Author Note

I am grateful for the financial support of this work by the Canadian Social Sciences & Humanities Research Council (SSHRC). Correspondence regarding this article should be addressed to ulrich.schimmack@utoronto.ca.

Abstract

Cronbach and Meehl (1955) introduced the concept of construct validity and described how researchers can demonstrate that their measures have construct validity. Although the term construct validity is widely used, few researchers follow Cronbach and Meehl's recommendation to quantify construct validity with the help of nomological networks. As a result, the construct validity of many popular measures in psychology is unknown. I call for rigorous tests of construct validity that follows Cronbach and Meehl's recommendations to improve psychology as a science. Without valid measures even replicable results are uninformative. I suggest that a proper program of validation research requires a multi-method approach and causal modeling of correlations with structural equation models. Construct validity should be quantified to enable cost benefit analyses and to replace existing measures with new measures with superior construct validity.

Keywords: Measurement, Construct Validity, Convergent Validity, Discriminant Validity, Structural Equation Modeling; Nomological Networks

The Validation Crisis in Psychology

Eight years ago, psychologists started to realize that they have a replication crisis. Many published results do not replicate in honest replication attempts that allow the data to decide whether a hypothesis is true or false (Open Science Collaboration, 2015). One key problem is that original studies often have low statistical power (Cohen, 1962; Schimmack, 2012). Another problem is that researchers use questionable research practices to increase power, which also increases the risk of false positive results (John et al., 2012). New initiatives that are called open science (e.g., preregistration, data sharing, a priori power analyses, registered reports) are likely to improve the replicability of psychological science in the future, although progress towards this goal is painfully slow.

Unfortunately, low replicability is not the only problem in psychological science. I argue that psychology not only has a replication crisis, but also a validation crisis. The need for valid measures seems obvious. To test theories that relate theoretical constructs to each other (e.g., construct A influences construct B for individuals drawn from population P under conditions C), it is necessary to have valid measures of constructs. For example, research on intelligence that uses hair length as a measure of intelligence would be highly misleading; highly replicable gender differences in hair length would be interpreted as evidence that women are more intelligent than men. This inference would be false because hair length is not a valid measure of intelligence, even though the relationship between gender and hair length is highly replicable. Thus, even successful and replicable tests of a theory may be false if measures lack construct validity; that is, they do not measure what researchers assume they are measuring.

The social sciences are notorious for imprecise use of terminology. The terms validity and validation are no exception. In educational testing, where the emphasis is on assessment of individuals, the term validation has a different meaning than in psychological science, where the emphasize is on testing psychological theories (Borsboom & Wijsen, 2016). In this article, I focus on construct validity. A measure possesses construct validity to the degree that quantitative variation in a measure reflects quantitative variation in the construct that the measure was design to measure. For example, a measure of anxiety is a valid measure of anxiety if scores on the measure reflect variation in anxiety.

Hundreds of measures are used in psychological science with the purpose of measuring variation in constructs such as learning, attention, emotions, attitudes, values, personality traits, abilities, or behavioral frequencies. Although measures of these constructs are used in thousands of articles, I argue that very little is known about the construct validity of these measures. That is, it is often claimed that psychological measures are valid, but evidence for this claim is often lacking or insufficient. I argue that psychologists could improve the quality of psychological science by following Cronbach and Meehl's (1955) recommendations for construct validation. Specifically, I argue that construct validation requires (a) a multi-method approach, (b) a causal model of the relationship between constructs and measures, and (c) quantitative information about the correlation between unobserved variation in constructs and observed scores on measures of constructs.

Construct Validity

The classic article on "Construct Validity" was written by two giants in the history of psychology; Cronbach and Meehl (1955). Every graduate student of psychology and surely every psychologist who published a psychological measure should be familiar with this article. The article was the result of an APA task force that tried to establish criteria, now called psychometric properties, that could be used to evaluate psychological measures. In this seminal article on construct validity Cronbach and Meehl note that construct validation is necessary “whenever a test is to be interpreted as a measure of some attribute or quality which is not “operationally defined” (p. 282). This definition makes it clear that there are other types of validity (e.g., criterion validity) and that not all measures require construct validity. However, studies of psychological theories that relate constructs to each other require valid measures of these constructs in order to test psychological theories.

In modern language, construct validity is the relationship between variation in observed scores on a measure (e.g., degree Celsius on a thermometer) and a latent variable that reflects corresponding variation in a theoretical construct (e.g., temperature; i.e., average kinetic energy of the particles in a sample of matter). The problem of construct validation can be illustrated with the development of IQ test. IQ scores can have predictive validity (e.g., performance in graduate school) without making any claims about the construct that is being measured (IQ tests measure whatever they measure and what they measure predicts important outcomes). However, IQ tests are often treated as measures of intelligence. For IQ tests to be valid measures of intelligence, it is necessary to define the construct of intelligence and to demonstrate that observed IQ scores are related to unobserved variation in intelligence. Thus, construct validation requires clear definitions of constructs that are independent of the measure that is being validated. Without clear definition of constructs, the meaning of a measure reverts essentially to “whatever the measure is measuring,” as in the old saying “Intelligence is whatever IQ tests are measuring.”

What are Constructs

Cronbach and Meehl (1955) define a construct as “some postulated attribute of people, assumed to be reflected in test performance (p. 283). The term “reflected” in Cronbach and Meehl’s definition makes it clear that they think about constructs as latent variables and the process of measurement as a reflective measurement model. This point is made even clearer when they write “It is clear that factors here function as constructs (p. 287). Individuals are assumed to have attributes; today we may say personality traits or states. These attributes are typically not directly observable (e.g., kindness rather than height), but systematic observation suggests that the attribute exists (some people are kinder than others across time and situations). The first step is to develop a measure of this attribute (e.g., a self-report measure “How kind are you?”). If the self-report measure is valid, variation in the ratings should reflect actual variation in kindness. This needs to be demonstrated in a program of validation research. For example, self-ratings should show convergent validity with informant ratings, and they should predict actual behavior in experience sampling studies or laboratory settings. Face validity is not sufficient; that is “I am kind” is not automatically a valid measure of kindness because the question directly maps on the construct.

Convergent Validity

To demonstrate construct validity, Cronbach and Meehl advocate a multi-method approach. The same construct has to be measured with several measures. If several measures are available, they can be analyzed with factor analysis. In this factor analysis, the factor represents the construct and factor loadings show how strongly scores in the observed measures are related to variation in the construct. For example, if multiple independent raters agree in their ratings of individuals’ kindness, the common factor in these ratings may correspond to the personality trait kindness, and the factor loadings provide evidence about the degree of construct validity of each measure (Schimmack, 2010).

It is important to distinguish factor analysis of items and factor analysis of multiple measures. Factor analysis of items is common and often used to claim validity of a measure.

However, correlations among self-report items are influenced by systematic measurement error (Anusic et al., 2009; Podsakoff, MacKenzie, & Podsakoff, 2012). The use of multiple independent methods (e.g., multiple raters) methods reduces the influences of shared method variance and makes it more likely that correlations among measures are caused by the influence of the common construct that the measures are intended to measure. In the section "Correlation matrices and factor analysis" Cronbach and Meehl (1955) clarify why factor analysis can reveal construct validity. "If two tests are presumed to measure the same construct, a correlation between them is predicted (p. 287).

The logic of this argument should be clear to any psychology student who was introduced to the third-variable problem in correlational research. Two measures may be related even if there is no causal relationship between them because they are both influenced by a common cause. For example, cities with more churches have a higher murder rate. Here the assumed common cause is population size. This makes it possible to measure population size with measures of the number of churches and murders. The shared variance between these measures reflects population size. Thus, we can think about constructs as third variables that produce shared variance among observed measures of the same construct.

This basic idea was refined by Campbell and Fiske (1959), who coined the term convergent validity. Two measures of the same construct possess convergent validity if they are positively correlated with each other. However, there is a catch. Two measures of the same construct could also be correlated for other reasons. For example, self-ratings of kindness and considerateness could be correlated due to socially desirable responding or evaluative biases in self-perceptions (Campbell & Fiske, 1959). Thus, Campbell and Fiske (1959) made clear that convergent validity is different from reliability. Reliability shows consistency in scores across measures without examining the source of the consistency in responses. Construct validity requires that consistency is produced by variation in the construct that a measure was designed to measure. For this reason, reliability is necessary, but not sufficient to demonstrate construct validity. An unreliable measure cannot be valid because there is no consistency, but a reliable measure can be invalid. For example, hair length can be measured reliably, but the reliable variance in the measure has no construct validity as a measure of intelligence.

One cause of the validation crisis in psychology is that validation studies ignore the distinction between same-method and cross-method correlations (Campbell & Fiske, 1959). Correlations among measures that share method variance (e.g., self-reports) cannot be used to examine convergent validity. It is unfortunate that psychologists have stopped using real behavior to validate self-report measures (Baumeister, Vohs, & Funder, 2007).

Discriminant Validity

The term discriminant validity was introduced by Campbell and Fiske (1959). However, Cronbach and Meehl already point out that high or low correlations can support construct validity. "Only if the underlying theory of the trait being measured calls for high item intercorrelations do the correlations support construct validity" (p. 288). Crucial for construct validity is that the correlations are consistent with theoretical expectations. For example, low correlations between intelligence and happiness do not undermine the validity of an intelligence measure because there is no theoretical expectation that intelligence is related to happiness. In contrast, low correlations between intelligence and job performance would be a problem if the jobs require problem solving skills and intelligence is an ability to solve problems faster or better.

It is often overlooked that discriminant validity also requires a multi-method approach (e.g., Greenwald, McGhee, & Schwartz, 1998). A multi-method approach is required because the upper limit for discriminant validity is the amount of convergent validity for different measures of the same construct, not a value of 1 or the reliability of a scale (Campbell & Fiske, 1959). For example, Martel, Schimmack, Nikolas, and Nigg (2015) examined multi-rater data of children's Attention Deficit and Hyperactivity (ADHD) symptoms. Table 1 shows the correlations for the items "listens" and "being organized." The cross-rater-same-item correlations (italics) show convergent validity of ratings of the same "symptom" by different raters. The cross-rater-different-item correlations (bold) show discriminant validity only if they are consistently lower than the convergent validity correlations. In this example, there is little evidence of discriminant validity because cross-construct correlations are nearly as high as same-construct correlations. An analysis with structural equation modeling of these data shows a latent correlation of r = .988 between a "listening" factor and an "organized" factor. This example illustrates why it is not possible to interpret items on an ADHD checklist as distinct symptoms (Martel et al., 2015). More important, the example shows that claims about discriminant validity require a multi-method assessment of convergent validity.

Quantifying Construct Validity

It is rare to see quantitative claims about construct validity in psychology, and sometimes information about reliability is falsely presented as evidence for high construct validity (Flake, Pek, & Hehman; 2017). Most method sections include a vague statement that measures have demonstrated construct validity as if a measure is either valid or invalid. Contrary to this current practice, Cronbach and Meehl made it clear that construct validity is a quantitative construct and that factor loadings can be used to quantify validity. "There is an understandable tendency to seek a “construct validity coefficient. A numerical statement of the degree of construct validity would be a statement of the proportion of the test score variance that is attributable to the construct variable. This numerical estimate can sometimes be arrived at by a factor analysis” (p. 289). And nobody today seems to remember Cronbach and Meehl’s (1955) warning that rejection of the null-hypothesis, the test has zero validity, is not the end goal of validation research. “It should be particularly noted that rejecting the null hypothesis does not finish the job of construct validation. The problem is not to conclude that the test “is valid” for measuring- the construct variable. The task is to state as definitely as possible the degree of validity the test is presumed to have (p. 290). Cronbach and Meehl are well-aware that it is difficult to quantify validity precisely, even if multiple measures of a construct are available because the factor may not be perfectly corresponding with the construct. “Rarely will it be possible to estimate definite “construct saturations,” because no factor corresponding closely to the construct will be available (p. 289). However, broad information about validity is better than no information about validity (Schimmack, 2010).

One reason why psychologists rarely quantify validity could be that estimates of construct validity for many tests are likely to be low. The limited evidence from some multi-method studies suggests that about 30% to 50% of the variance in rating scales is valid variance (Connelly & Ones, 2010; Zou, Schimmack, & Gere, 2013). Another reason is that it can be difficult or costly to measure the same construct with three independent methods, which is the minimum number of measures to quantify validity. Two methods are insufficient because it is not clear how much validity of each method contributes to the convergent validity correlation between them. For example, a correlation of r = .4 between self-ratings and informant ratings is open to very different interpretations. “If the obtained correlation departs from the expectation, however, there is no way to know whether the fault lies in test A, test B, or the formulation of the construct" (Cronbach & Meehl, 1955, p. 300).

I believe that the failure to treat construct validity as a quantitative construct is the root cause of the validation crisis in psychology. Every method is likely to have some validity (i.e., non-zero construct variance), but measures with less than 5% construct variance should not be used. Moreover, quantification of construct validity would provide an objective criterion to evaluate new measures and stimulate development of better measures.

One notable exception is the literature in industrial and organizational psychology, where construct validity has been quantified (Cote & Buckley, 1987). A meta-analysis of construct validation studies suggested that less than 50% of the variance was valid construct variance, and that a substantial portion of the variance is caused by systematic measurement error. The I/O literature shows that it is possible and meaningful to quantify construct validity. I suggest that other disciplines in psychology follow their example.

The Nomological Net

Some readers may be familiar with the term "nomological net" that was popularized by Cronbach and Meehl in their 1995 article. However, few readers will be able to explain what a nomological net is, despite the fact that Cronbach and Meehl considered nomological nets essential for construct validation. "To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist (p. 291).

Cronbach and Meehl state that "the laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These "laws" may be statistical or deterministic" (p. 290). I argue that Cronbach and Meehl would have used the term Structural

Equation Model, if structural equation modeling existed when they wrote their article. After all, structural equation modeling is simply an extension of factor analyses, and Cronbach and Meehl did equate constructs with factors, and structural equation modeling makes it possible to relate (a) observed indicators to each other, (b) observed indicators to latent variables, and (c) latent variables to each other. Thus, Cronbach and Meehl essentially proposed to examine construct validity by modeling multi-trait-multi-method data with structural equations.

Cronbach and Meehl also realize that constructs can change as more information becomes available. It may also occur that the data fail to provide evidence for a construct. In this sense, construct validation is an ongoing process of improved understanding of constructs and measures. Empirical data can suggest changes in measures or changes in concepts. For example, empirical data might show that intelligence is a general disposition that influences many different cognitive abilities or that it is better conceptualized as the sum of several distinct cognitive abilities.

Ideally this iterative process would start with a simple structural equation model that is fitted to some data. If the model does not fit, the model can be modified and tested with new data. Over time, the model would become more complex and more stable because core measures of constructs would establish the meaning of a construct, while peripheral relationships may be modified if new data suggest that theoretical assumptions need to be changed. “When observations will not fit into the network as it stands, the scientist has a certain freedom in selecting where to modify the network” (p. 290). The increasing complexity of a model is only an advantage if it is based on better understanding of a phenomenon. Weather models have become increasingly more complex and better able to forecast future weather changes. In the same way, better psychological models would be more complex and better able to predict behavior.

Structural equation modeling is sometimes called confirmatory factor analysis. In my opinion, the term confirmatory factor analysis has led to the idea that structural equation modeling can only be used to test whether a theoretical model fits the data or not. The consequences of this believe was that structural equation modeling was not used because it typically showed that simplistic models did not fit the data. For example, McCrae, Zonderman, Costa, Bond, and Paunonen (1996) dismissed structural equation modeling as a useful method to examine the construct validity of Big Five measures because it failed to support their conception of the Big Five as orthogonal dimensions with simple structure.

I argue that structural equation modeling is a statistical tool that can be used to test existing models and to explore new models. This flexible use of structural equation model would be in the spirit of Cronbach and Meehl's vision that construct validation is an iterative process that improves measurement and understanding of constructs as the nomological net is altered to accommodate new information.

This suggestion highlights a similarity between the validation crisis and the replication crisis. One cause of the replication crisis was the use of statistics as a tool that could only confirm theoretical predictions, $p < .05$. In the same way, confirmatory factor analysis was only used to confirm models. In both cases, confirmation bias impeded scientific progress and theory development. A better use of structural equation modeling is to use it as a general statistical framework that can be used fit nomological networks to data and to use the results in an iterative processes that leads to better understanding of constructs and better measures of these constructs.

Network Models are Not Nomological Nets

In the past decade, it has become popular to examine correlations among items with network models (Schmittmann et al., 2013). Network models are graphic representations of correlations or partial correlations among a set of variables. Importantly, network models do not have latent variables that could correspond to constructs. “Network modeling typically relies on the assumption that the covariance structure among a set of the items is not due to latent variables at all” (Epskamp et al., 2017, p. 923). Instead, “psychological attributes are conceptualized as networks of directly related observables” (Schmittmann et al., 2013, p. 43).

It is readily apparent that network models are not nomological nets because they avoid defining constructs independent of specific operationalizations. “Since there is no latent variable that requires causal relevance, no difficult questions concerning its reality arise” (Schmittmann et al., 2013, p. 49). Thus, network model return to operationalism at the level of the network components. Each component in the network is defined by a specific measure, which is typically a self-report item or scale. The difficulty of psychological measurement is no longer a problem because self-report items are treated as perfectly valid measures of network components. The example in Table 1 shows the problem with this approach. Rather than having six independent network components, the six items in Table 1 appear to be six indicators of a single construct that are measured with systematic and random measurement error. At least for these data, but probably for multi-method data in general, it makes little sense to postulate direct causal effects between observed scores. For example, it makes little sense to postulate that father’s ratings of forgetfulness causally influenced teachers’ ratings of attention.

It is noteworthy that recent trends in network modeling acknowledge the importance of latent variables and relegate the use of network modeling to modeling residual correlations

(Epskamp, Rhemtulla, & Borsboom, 2017). These network models with latent variables are functionally equivalent to structural equation models with correlated residuals. Thus, they are no longer conceptually distinct from structural equation models.

A detailed discussion of latent network models is beyond this article. The main point is that traditional network models without latent variables cannot be used to examine construct validity because constructs are by definition unobservable and can be studied only indirectly by examining their influence on observable measures. Any direct relationships between observables either operationalize constructs or avoid the problem of measurement and implicitly assume perfect measurement.

Recommendations for Users of Psychological Measures

The main recommendation for users of psychological measures is to be skeptical of claims that measures have construct validity. Many of these claims are not based on proper validation studies. At a minimum a measure should have demonstrated at least modest convergent validity with another measure that used a different method. Ideally, a multi-method approach was used to provide some quantitative information about construct validity. Researchers should be wary of measures that have low convergent validity. For example, it has been known for a long time that implicit measures of self-esteem have low convergent validity (Bosson et al., 2000), but this finding has not deterred researchers from claiming that the self-esteem IAT is a valid measure of implicit self-esteem (Greenwald & Farnham (2001). Proper evaluation of this claim with multi-method data shows no evidence of construct validity (Falk et al., 2015; Schimmack, 2019).

Consumers should also be wary of new constructs. It is very unlikely that all hunches by psychologists lead to the discovery of useful constructs and development of valid tests of these constructs. Given the lack of knowledge about the mind, it is rather more likely that many constructs turn out to be non-existent (i.e., like unicorns) or that measures have low construct validity. However, the history of psychological measurement has only seen development of more and more constructs and more and more measures to measure this increasing universe of constructs. Since the 1990s, constructs have doubled because every construct has been split into an explicit and an implicit version of the construct. Presumably, there is even implicit political orientation or gender identity.

The proliferation of constructs and measures is not a sign of a healthy science. Rather it shows the inability of empirical studies to demonstrate that a measure is not valid, a construct does not exist, or a construct is redundant with other constructs. This is mostly due to self-serving biases and motivated reasoning of test developers. The gains from a measure that is widely used are immense. Articles that introduced popular measures like the Implicit Association Test (Greenwald et al., 1998) have some of the highest citation rates. Thus, it is tempting to use weak evidence to make sweeping claims about validity because the rewards for publishing a widely used measure are immense. One task for meta-psychologists could be to critically evaluate claims of construct validity by original authors because original authors are unlikely to be unbiased in their evaluation of construct validity (Cronbach, 1989).

The Validation Crisis

Cronbach and Meehl make it clear that they were skeptical about the construct validity of many psychological measures. “For most tests intended to measure constructs, adequate criteria do not exist. This being the case, many such tests have been left unvalidated, or a fine-spun network of rationalizations has been offered as if it were validation. Rationalization is not construct validation. One who claims that his test reflects a construct cannot maintain his claim in the face of recurrent negative results because these results show that his construct is too loosely defined to yield verifiable inferences (p. 291). In my opinion, nothing much has changed in the world of psychological measurement. Flake et al. (2017) reviewed current practices and found that reliability is often the only criterion that is used to claim construct validity. However, reliability of a single measure cannot be used to demonstrate construct validity because reliability is only necessary, but not sufficient for validity (e.g., hair length example in the beginning). Thus, many articles provide no evidence for construct validity and even if the evidence were sufficient to claim that a measure is valid, it remains unclear how valid a measure is.

Another sign that psychology has a validity crisis is that psychologists today still use measures that were developed decades ago (cf. Schimmack, 2010). Although these measures could be highly valid, it is also likely that they have not been replaced by better measures because quantitative evaluations of measures are lacking. For example, Rosenberg's (1965) 10-item self-esteem scale is still the most widely used measure of self-esteem (Bosson et al., 2000). However, the construct validity of this measure has never been quantified and it is not clear whether it is more valid than other measures of self-esteem. The same is true for various implicit measures of self-esteem (Schimmack, 2019).

What is the Alternative?

While there is general agreement that current practices have serious limitations (Kane, 2017; Maul, 2017), there is no general agreement about the best way to address the validation crisis. Some comments suggest that psychology might fare better without quantitative measurement (Maul, 2017). If we look to the natural sciences, this does not appear to be an attractive alternative. In the natural sciences progress has been made by increasingly more sophisticated measurements of basic units such as time and length (nanotechnology). Meehl was an early proponent of more rather than less rigorous methods in psychology. If psychologists had followed his advice to quantify validity, psychological science would have made more progress. Thus, I do not think that abandoning quantitative psychology is an attractive alternative.

Others believe that Cronbach and Meehl's agenda is too ambitious (Kane, 2016, 2017). "Where the theory is strong enough to support such efforts, I would be in favor of using them, but in most areas of research, the required theory is lacking" (Kane, 2017, p. 81). This may be true for some areas of psychology, such as educational testing, but it is not true for basic psychological science where the sole purpose of measures is to test psychological theories. In this context, construct validation is crucial for testing of causal theories. For example, theories of implicit social cognition require valid measure of implicit cognitive processes (Greenwald et al., 1998; Schimmack, 2019). Thus, I am more optimistic than Kane that psychologists have causal theories of important constructs such as attitudes, personality traits, and well-being that can inform a program develop and test causal theories of measurement processes and use these theories to quantify the validity of psychological measures. The industrial literature shows that it is possible to estimate construct validity even with rudimentary causal theories (Cote & Buckley, 1987), and there are some examples in social and personality psychology where I used SEM to specify measurement models of attitudes (Schimmack, 2019), personality traits (Schimmack, 2010), or well-being (Zou et al., 2013).

Conclusion

Just like psychologist have started to appreciate replication failures in the past years, they need to embrace validation failures. Some of the measures that are currently used in psychology are likely to have insufficient construct validity. If the 2010s were the decade of replication, the 2020s may become the decade of validation. It is time to examine how valid the most widely used psychological measures actually are. Cronbach and Meehl (1955) outlined a program of construct validation research. Ample citations show that they were successful in introducing the term, but psychologists failed in adopting the rigorous practices they were recommending. It is time to change this and establish clear standards of construct validation that psychological measures should meet. Most important, validity has to be expressed in quantitative terms to encourage competition for developing new measures of existing constructs with higher validity.

References

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The nature and structure of correlations among Big Five ratings: The halo-alpha-beta model. *Journal of Personality and Social Psychology, 97*(6), 1142-1156. http://dx.doi.org/10.1037/a0017159

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? Perspectives on Psychological Science, 2(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Borsboom, D. & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice, 23*, 281-283. DOI: 10.1080/0969594X.2016.1141750

Bosson, J. K., Swann, W. B., Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? Journal of Personality and Social Psychology, 79(4), 631-643. http://dx.doi.org/10.1037/0022-3514.79.4.631

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105. http://dx.doi.org/10.1037/h0046016

Cohen, J. (1962). Statistical power of abnormal–social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145–153. doi:10.1037/h0045186

Connelly, B. S., & Ones, D. S. (2010). An other-perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136(6), 1092-1122.http://dx.doi.org/10.1037/a0021212

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing, 24*, 315-318.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement, theory, and public policy (pp. 147-171). Chicago: University of Illinois Press

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302. http://dx.doi.org/10.1037/h0040957

Falk, C., Heine, S. J., Takemura, K., Zhang, C., & Hsu, C. W. (2015). Are implicit self-esteem measures valid for assessing individual and cultural differences? *Journal of Personality, 83*, 56-68. DOI:10.1111/jopy.12082

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. Social Psychological and Personality Science, 8(4), 370–378. https://doi.org/10.1177/1948550617693063

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social Psychology, 79(6), 1022-1038. http://dx.doi.org/10.1037/0022-3514.79.6.1022

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23, 524–532. doi:10.1177/0956797611430953

Kane, M. T. (2016) Explicating validity, Assessment in Education: Principles, Policy & Practice, 23:2, 198-211, DOI: 10.1080/0969594X.2015.1060192

Kane, M. T. (2017) Causal Interpretations of Psychological Attributes. *Measurement: Interdisciplinary Research and Perspectives, 15*, 79-82, DOI: 10.1080/15366367.2017.1369771

Martel, M. M., Schimmack, U., Nikolas, M., & Nigg, J. T. (2015). Integration of symptom ratings from multiple informants in ADHD diagnosis: a psychometric model with clinical utility. Psychological Assessment, 27(3), 1060-71.

Maul. A. (2017). Moving beyond traditional methods of survey validation. Measurement: Interdisciplinary Research and Perspectives, 15, 103-109. https://doi.org/10.1080/15366367.2017.1369786

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. Journal of Personality and Social Psychology, 70(3), 552-566. http://dx.doi.org/10.1037/0022-3514.70.3.552

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, (6251), 943-950.

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P. (2012). Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annual Review of Psychology, 63*, 539-569.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.

Schimmack, U. (2010). What multi-method data tell us about construct validity. European *Journal of Personality, 24*, 241–257. DOI: 10.1002/per.771

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551-566. http://dx.doi.org/10.1037/a0029487

Schimmack (2019). *The Implicit Association Test at Age 21: No Evidence for Construct Validity*. https://replicationindex.com/2019/02/15/iat21/

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*, 43-53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Zou, C., Schimmack, U., & Gere, J. (2013). The validity of well-being measures: A multiple-indicator–multiple-rater model. Psychological Assessment, 25(4), 1247-1254. http://dx.doi.org/10.1037/a0033902

Table 1. Correlation among ratings of ADHD symptoms

	M-Listen	F-Listen	T-Listen	M-Organized	F-Organized	T-Organized
M-Listen	-					
F-Listen	*0.558*	-				
T-Listen	*0.450*	*0.436*	-			
M-Organized	0.664	**0.494**	**0.392**	-		
F-Organized	**0.432**	0.561	**0.324**	*0.437*	-	
T-Organized	**0.376**	**0.407**	0.698	*0.350*	*0.304*	-

Note. M = Mother, F = Father, T = Teacher, Ratings of child listens and child is organized. Data from Martel et al. (2015)