

What Multi-Method Data Tell Us About Construct Validity

ULRICH SCHIMMACK*

University of Toronto Mississauga, Canada

Abstract

Structural equation modelling of multi-method data has become a popular method to examine construct validity and to control for random and systematic measurement error in personality measures. I review the essential assumptions underlying causal models of multi-method data and their implications for estimating the validity of personality measures. The main conclusions are that causal models of multi-method data can be used to obtain quantitative estimates of the amount of valid variance in measures of personality dispositions, but that it is more difficult to determine the validity of personality measures of act frequencies and situation-specific dispositions. Copyright © 2010 John Wiley & Sons, Ltd.

Key words: statistical methods; personality scales and inventories; regression methods; history of psychology; construct validity; causal modelling; multi-method; measurement

INTRODUCTION

Fifty years ago, Campbell and Fiske (1959) published the groundbreaking article *Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix*. With close to 5000 citations (Web of Science, February 1, 2010), it is the most cited article in *Psychological Bulletin*. The major contribution of this article was to outline an empirical procedure for testing the validity of personality measures. It is difficult to overestimate the importance of this contribution because it is impossible to test personality theories empirically without valid measures of personality.

Despite its high citation count, Campbell and Fiske's work is often neglected in introductory textbooks, presumably because validation is considered to be an obscure and complicated process (Borsboom, 2006). Undergraduate students of personality psychology learn little more than the definition of a valid measure as a measure that measures what it is supposed to measure. However, they are not taught how personality psychologists validate their measures. One might hope that aspiring personality researchers learn about Campbell

*Correspondence to: Ulrich Schimmack, University of Toronto Mississauga, Canada.
E-mail: uli.schimmack@utoronto.ca

Received 5 January 2010
Revised 17 February 2010
Accepted 17 February 2010

and Fiske's multi-method approach during graduate school. Unfortunately, even handbooks dedicated to research methods in personality psychology pay relatively little attention to Campbell and Fiske's (1959) seminal contribution (John & Soto, 2007; Simms & Watson, 2007). More importantly, construct validity is often introduced in qualitative terms. In contrast, when Cronbach and Meehl (1955) introduced the concept of construct validity, they proposed a quantitative definition of construct validity as the proportion of construct-related variance in the observed variance of a personality measure. Although the authors noted that it would be difficult to obtain precise estimates of construct validity coefficients (CVCs), they stressed the importance of estimating 'as definitely as possible the degree of validity the test is presumed to have' (p. 290). Campbell and Fiske's (1959) multi-method approach paved the way to do so. Although Campbell and Fiske's article examined construct validity qualitatively, subsequent developments in psychometrics allowed researchers to obtain quantitative estimates of construct validity based on causal models of multi-method data (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Kenny & Kashy, 1992). Research articles in leading personality journals routinely report these estimates (Biesanz & West, 2004; DeYoung, 2006; Diener, Smith, & Fujita, 1995), but a systematic and accessible introduction to causal models of multi-method data is lacking. The main purpose of this paper is to explain how causal models of multi-method data can be used to obtain quantitative estimates of construct validity and which assumptions these models make to yield accurate estimates.

I prefer the term causal model to the more commonly used term structural equation model because I interpret latent variables in these models as unobserved, yet real causal forces that produce variation in observed measures (Borsboom, Mellenbergh, & van Heerden, 2003). I make the case below that this realistic interpretation of latent factors is necessary to use multi-method data for construct validation research because the assumption of causality is crucial for the identification of latent variables with construct variance (CV).

Campbell and Fiske (1959) distinguished absolute and relative (construct) validity. To examine relative construct validity it is necessary to measure multiple traits and to look for evidence of convergent and discriminant validity in a multi-trait-multi-method matrix (Simms & Watson, 2007). However, to examine construct validity in an absolute sense, it is only necessary to measure one construct with multiple methods. In this paper, I focus on convergent validity across multiple measures of a single construct because causal models of multi-method data rely on convergent validity alone to examine construct validity.

As discussed in more detail below, causal models of multi-method data estimate construct validity quantitatively with the factor loadings of observed personality measures on a latent factor (i.e. an unobserved variable) that represents the valid variance of a construct. The amount of valid variance in a personality measure can be obtained by squaring its factor loading on this latent factor. In this paper, I use the terms *construct validity coefficient (CVC)* to refer to the factor loading and the term *construct variance (CV)* for the amount of valid variance in a personality measure.

Validity

A measure is valid if it measures what it was designed to measure. For example, a thermometer is a valid measure of temperature in part because the recorded values covary with humans' sensory perceptions of temperature (Cronbach & Meehl, 1955). A modern thermometer is a more valid measure of temperature than humans' sensory perceptions, but

the correlation between scores on a thermometer and humans' sensory perceptions is necessary to demonstrate that a thermometer measures temperature. It would be odd to claim that highly reliable scores recorded by an expensive and complicated instrument measure temperature if these scores were unrelated to humans' everyday perceptions of temperature.

The definition of validity as a property of a measure has important implications for empirical tests of validity. Namely, researchers first need a clearly defined construct before they can validate a potential measure of the construct. For example, to evaluate a measure of anxiety researchers first need to define anxiety and then examine the validity of a measure as a measure of anxiety. Although the importance of clear definitions for construct validation research may seem obvious, validation research often seems to work in the opposite direction; that is, after a measure has been created psychologists examine what it measures. For example, the widely used Positive Affect and Negative Affect Schedule (PANAS) has two scales named Positive Affect (PA) and Negative Affect (NA). These scales are based on exploratory factor analyses of mood ratings (Watson, Clark, & Tellegen, 1988). As a result, Positive Affect and Negative Affect are merely labels for the first two VARIMAX rotated principal components that emerged in these analyses. Thus, it is meaningless to examine whether the PANAS scales are valid measures of PA and NA. They are valid measures of PA and NA by definition because PA and NA are mere labels of the two VARIMAX rotated principal components that emerge in factor analyses of mood ratings. A construct validation study would have to start with an *a priori* definition of Positive Affect and Negative Affect that does not refer to the specific measurement procedure that was used to create the PANAS scales. For example, some researchers have defined Positive Affect and Negative Affect as the valence of affective experiences and have pointed out problems of the PANAS scales as measures of pleasant and unpleasant affective experiences (see Schimmack, 2007, for a review). However, the authors of the PANAS do not view their measure as a measure of hedonic valence. To clarify their position, they proposed to change the labels of their scales from Positive Affect and Negative Affect to Positive Activation and Negative Activation (Watson, Wiese, Vaidya, & Tellegen, 1999). The willingness to change labels indicates that PANAS scales do not measure *a priori* defined constructs and as a result there is no criterion to evaluate the construct validity of the PANAS scales.

The previous example illustrates how personality measures assume a life of their own and implicitly become the construct; that is, a construct is operationally defined by the method that is used to measure it (Borsboom, 2006). A main contribution of Campbell and Fiske's (1959) article was to argue forcefully against operationalism and for a separation of constructs and methods. This separation is essential for validation research because validation research has to allow for the possibility that some of the observed variance is invalid.

Other sciences clearly follow this approach. For example, physics has clearly defined concepts such as time or temperature. Over the past centuries, physicists have developed increasingly precise ways of measuring these concepts, but the concepts have remained the same. Modern physics would be impossible without these advances in measurement. However, psychologists do not follow this model of more advanced sciences. Typically, a measure becomes popular and after it becomes popular it is equated with the construct. As a result, researchers continue to use old measure and rarely attempt to create better measures of the same construct. Indeed, it is hard to find an example, in which one measure of a construct has replaced another measure of the same construct based on an empirical

comparison of the construct validity of competing measures of the same construct (Grucza & Goldberg, 2007).

One reason for the lack of progress in the measurement of personality constructs could be the belief that it is impossible to quantify the validity of a measure. If it were impossible to quantify the validity of a measure, then it also would be impossible to say which of two measures is more valid. However, causal models of multi-method data produce quantitative estimates of validity that allow comparisons of the validity of different measures.

One potential obstacle for construct validation research is the need to define psychological constructs *a priori* without reference to empirical data. This can be difficult for constructs that make reference to cognitive processes (e.g. working memory capacity) or unconscious motives (implicit need for power). However, the need for *a priori* definitions is not a major problem in personality psychology. The reason is that everyday language provides thousands of relatively well-defined personality constructs (Allport & Odbert, 1936). In fact, all measures in personality psychology that are based on the lexical hypothesis assume that everyday concepts such as helpful or sociable are meaningful personality constructs. At least with regard to these relatively simple constructs, it is possible to test the construct validity of personality measures. For example, it is possible to examine whether a sociability scale really measures sociability and whether a measure of helpfulness really measures helpfulness.

Convergent validity

I start with a simple example to illustrate how psychologists can evaluate the validity of a personality measure. The concept is people's weight. Weight can be defined as 'the vertical force exerted by a mass as a result of gravity' (wordnet.princeton.edu). In the present case, only the mass of human adults is of interest. The main question, which has real practical significance in health psychology (Kroh, 2005), is to examine the validity of self-report measures of weight because it is more economical to use self-reports than to weigh people with scales.

To examine the validity of self-reported weight as a measure of actual weight, it is possible to obtain self-reports of weight and an objective measure of weight from the same individuals. If self-reports of weight are valid, they should be highly correlated with the objective measure of weight. In one study, participants first reported their weight before their weight was objectively measured with a scale several weeks later (Rowland, 1990). The correlation in this study was $r(N = 11284) = .98$. The implications of this finding for the validity of self-reports of weight depend on the causal processes that underlie this correlation, which can be examined by means of causal modelling of correlational data.

It is well known that a simple correlation does not reveal the underlying causal process, but that some causal process must explain why a correlation was observed (Chaplin, 2007). Broadly speaking, a correlation is determined by the strength of four causal effects, namely, the effect of observed variable A on observed variable B, the effect of observed variable B on observed variable A, and the effects of an unobserved variable C on observed variable A and on observed variable B. In the present example, the observed variables are the self-reported weights and those recorded by a scale. To make inferences about the validity of self-reports of weight it is necessary to make assumptions about the causal processes that produce a correlation between these two methods. Fortunately, it is relatively easy to do so in this example. First, it is fairly certain that the values recorded by a scale are not influenced by individuals' self-reports. No matter how much individuals insist that the scale

is wrong, it will not change its score. Thus, it is clear that the causal effect of self-reports on the objective measure is zero. It is also clear that self-reports of weight in this study were not influenced by the objective measurement of weight in this study because self-reports were obtained weeks before the actual weight was measured. Thus, the causal effect of the objectively recorded scores on self-rating is also zero. It follows that the correlation of $r = .98$ must have been produced by a causal effect of an unobserved third variable. A plausible third variable is individuals' actual mass. It is their actual mass that causes the scale to record a higher or lower value and their actual mass also caused them to report a specific weight. The latter causal effect is probably mediated by prior objective measurements with other scales, and the validity of these scales would influence the validity of self-reports among other factors (e.g. socially desirable responding). In combination, the causal effects of actual mass on self-reports and on the scale produce the observed correlation of $r = .98$. This correlation is not sufficient to determine how strong the effects of weight on the two measures are. It is possible that the scale was a perfect measure of weight. In this case, the correlation between weight and the values recorded by the scale is 1. It follows, that the size of the effect of weight on self-reports of weight (or the factor loading of self-reported weight on the weight factor) has to be $r = .98$ to produce the observed correlation of $r = .98$ ($1 \times .98 = .98$). In this case, the CVC of the self-report measure of weight would be .98. However, it is also possible that the scale is a slightly imperfect measure of weight. For example, participants may not have removed their shoes before stepping on the scale and differences in the weight of shoes (e.g. boots versus sandals) could have produced measurement error in the objective measure of individuals' true weight. It is also possible that changes in weight over time reduce the validity of objective scores as a validation criterion for self-ratings several weeks earlier. In this case, the estimate underestimates the validity of self-ratings.

In the present context, the reasons for the lack of perfect convergent validity are irrelevant. The main point of this example was to illustrate how the correlation between two independent measures of the same construct can be used to obtain quantitative estimates of the validity of a personality measure. In this example, a conservative estimate of the CVC of self-reported weight as a measure of weight is .98 and the estimated amount of CV in the self-report measure is 96% ($.98^2 = .96$).

The example of self-reported weight was used to establish four important points about construct validity. First, the example shows that convergent validity is sufficient to examine construct validity. The question of how self-reports of weight are related to measures of other constructs (e.g. height, social desirable responding) can be useful to examine sources of measurement error, but correlations with measures of other constructs are not needed to estimate CVCs. Second, empirical tests of construct validity do not have to be an endless process without clear results (Borsboom, 2006). At least for some self-report measures it is possible to provide a meaningful answer to the question of their validity. Third, validity is a quantitative construct. Qualitative conclusions that a measure is valid because validity is not zero ($CVC > 0$, $p < .05$) or that a measure is invalid because validity is not perfect ($CVC < 1.0$, $p < .05$) are not very helpful because most measures are valid and invalid ($0 < CVC < 1$). As a result, qualitative reviews of validity studies are often the source of fruitless controversies (Schimmack & Oishi, 2005). The validity of personality measures should be estimated quantitatively like other psychometric properties such as reliability coefficients, which are routinely reported in research articles (Schmidt & Hunter, 1996). Validity is more important than reliability because reliable and invalid measures are potentially more dangerous than unreliable measures (Blanton & Jaccard, 2006).

Moreover, it is possible that a less reliable measure is more valid than a more reliable measure if the latter measure is more strongly contaminated by systematic measurement error (John & Soto, 2007). A likely explanation for the emphasis on reliability is the common tendency to equate constructs with measures. If a construct is equated with a measure, only random error can undermine the validity of a measure. The main contribution of Campbell and Fiske (1959) was to point out that systematic measurement error can also threaten the validity of personality measures. As a result, high reliability is insufficient evidence for the validity of a personality measure (Borsboom & Mellenbergh, 2002).

The third point illustrated in this example is that tests of convergent validity require independent measures. Campbell and Fiske (1959) emphasized the importance of independent measures when they defined convergent validity as the correlation between 'maximally different methods' (p. 83). In a causal model of multi-method data the independence assumption implies that the only causal effects that produce a correlation between two measures of the same construct are the causal effect of the construct on the two measures. This assumption implies that all the other potential causal effects that can produce correlations among observed measures have an effect size of zero. If this assumption is correct, the shared variance across independent methods represents CV. It is then possible to estimate the proportion of the shared variance relative to the total observed variance of a personality measure as an estimate of the amount of CV in this measure.

For example, in the previous example I assumed that actual mass was the only causal force that contributed to the correlation between self-reports of weight and objective scale scores. This assumption would be violated if self-ratings were based on previous measurements with objective scales (which is likely) and objective scales share method variance that does not reflect actual weight (which is unlikely). Thus, even validation studies with objective measures implicitly make assumptions about the causal model underlying these correlations.

In sum, the weight example illustrated how a causal model of the convergent validity between two measures of the same construct can be used to obtain quantitative estimates of the construct validity of a self-report measure of a personality characteristic. The following example shows how the same approach can be used to examine the construct validity of measures that aim to assess personality traits without the help of an objective measure that relies on well-established measurement procedures for physical characteristics like weight.

CONVERGENT VALIDITY OF PERSONALITY MEASURES

A Hypothetical Example

I use helpfulness as an example. Helpfulness is relatively easy to define as 'providing assistance or serving a useful function' (wordnetweb.princeton.edu/perl/webwn). Helpful can be used to describe a single act or an individual. If helpful is used to describe a single act, helpful is not only a characteristic of a person because helping behaviour is also influenced by situational factors and interactions between personality and situational factors. Thus, it is still necessary to provide a clearer definition of helpfulness as a personality characteristic before it is possible to examine the validity of a personality measure of helpfulness.

Personality psychologists use trait concepts like helpful in two different ways. The most common approach is to define helpful as an internal disposition. This definition implies causality. There are some causal factors within an individual that make it more likely for this individual to act in a helpful manner than other individuals. The alternative approach is to define helpfulness as the frequency with which individuals act in a helpful manner. An individual is helpful if he or she acted in a helpful manner more often than other people. This approach is known as the act frequency approach. The broader theoretical differences between these two approaches are well known and have been discussed elsewhere (Block, 1989; Funder, 1991; McCrae & Costa, 1995). However, the implications of these two definitions of personality traits for the interpretation of multi-method data have not been discussed. Ironically, it is easier to examine the validity of personality measures that aim to assess internal dispositions that are not directly observable than to do so for personality measures that aim to assess frequencies of observable acts. This is ironic because intuitively it seems to be easier to count the frequency of observable acts than to measure unobservable internal dispositions. In fact, not too long ago some psychologists doubted that internal dispositions even exist (cf. Goldberg, 1992).

The measurement problem of the act frequency approach is that it is quite difficult to observe individuals' actual behaviours in the real world. For example, it is no trivial task to establish how often John was helpful in the past month. In comparison it is relatively easy to use correlations among multiple imperfect measures of observable behaviours to make inferences about the influence of unobserved internal dispositions on behaviour. Figure 1 illustrates how a causal model of multi-method data can be used for this purpose. In Figure 1, an unobserved general disposition to be helpful influences three observed measures of helpfulness. In this example, the three observed measures are informant ratings of helpfulness by a friend, a co-worker and a spouse. Unlike actual informant

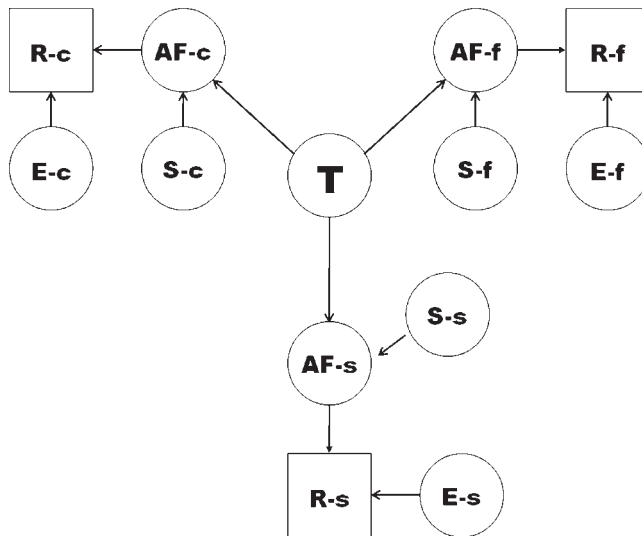


Figure 1. Theoretical model of multi-method data. *Note.* T = trait (general disposition); AF-c, AF-f, AF-s = act frequencies with colleague, friend and spouse; S-c, S-f, S-s = situational and person \times situation interaction effects on act frequencies; R-c, R-f, R-s = reports by colleague, friend and spouse; E-c, E-f, E-s = errors in reports by colleague, friend and spouse.

ratings in personality research, informants in this hypothetical example are only asked to report how often the target helped them in the past month. According to Figure 1, each informant report is influenced by two independent factors, namely, the actual frequency of helpful acts towards the informant and (systematic and random) measurement error in the reported frequencies of helpful acts towards the informant. The actual frequency of helpful acts is also influenced by two independent factors. One factor represents the general disposition to be helpful that influences helpful behaviours across situations. The other factor represents situational factors and person-situation interaction effects. To fully estimate all coefficients in this model (i.e. effect sizes of the postulated causal effects), it would be necessary to separate measurement error and valid variance in act frequencies. This is impossible if, as in Figure 1, each act frequency is measured with a single method, namely, one informant report. In contrast, the influence of the general disposition is reflected in all three informant reports. As a result, it is possible to separate the variance due to the general disposition from all other variance components such as random error, systematic rating biases, situation effects and person \times situation interaction effects. It is then possible to determine the validity of informant ratings as measures of the general disposition, but it is impossible to (precisely) estimate the validity of informant ratings as measures of act frequencies because the model cannot distinguish reporting errors from situational influences on helping behaviour.

The causal model in Figure 1 makes numerous independence assumptions that specify Campbell and Fiske's (1959) requirement that traits should be assessed with independent methods. First, the model assumes that biases in ratings by one rater are independent of biases in ratings by other raters. Second, it assumes that situational factors and person \times situation interaction effects that influence helping one informant are independent of the situational and person \times situation factors that influence helping other informants. Third, it assumes that rating biases are independent of situation and person \times situation interaction effects for the same rater and across raters. Finally, it assumes that rating biases and situation effects are independent of the global disposition. In total, this amounts to 21 independence assumptions (i.e. Figure 1 includes seven exogeneous variables, that is, variables that do not have an arrow pointing at them, which implies 21 ($7 \times 6/2$) relationships that the model assumes to be zero). If these independence assumptions are correct, the correlations among the three informant ratings can be used to determine the variation in the unobserved personality disposition to be helpful with perfect validity. This variance can then be used like the objective measure of weight in the previous example as the validation criterion for personality measures of the general disposition to be helpful (e.g. self-ratings of general helpfulness). In sum, Figure 1 illustrates that a specific pattern of correlations among independent measures of the same construct can be used to obtain precise estimates of the amount of valid variance in a single measure.

The main challenge for actual empirical studies is to ensure that the methods in a multi-method model fulfil the independence assumptions. The following examples demonstrate the importance of the neglected independence assumption for the correct interpretation of causal models of multi-method data. I also show how researchers can partially test the independence assumption if sufficient methods are available and how researchers can estimate the validity of personality measures that aggregate scores from independent methods. Before I proceed, I should clarify that strict independence of methods is unlikely, just like other null-hypotheses are likely to be false. However, small violations of the independence assumption will only introduce small biases in estimates of CVCs.

Example 1: Multiple response formats

The first example is a widely cited study of the relation between Positive Affect and Negative Affect (Green, Goldman, & Salovey, 1993). I chose this paper because the authors emphasized the importance of a multi-method approach for the measurement of affect, while neglecting Campbell and Fiske's requirement that the methods should be maximally different. A major problem for any empirical multi-method study is to find multiple independent measures of the same construct. The authors used four self-report measures with different response formats for this purpose. However, the variation of response formats can only be considered a multi-method study, if one assumes that responses on one response format are independent of responses on the other response formats so that correlations across response formats can only be explained by a common causal effect of actual momentary affective experiences on each response format. However, the validity of all self-report measures depends on the ability and willingness of respondents to report their experiences accurately. Violations of this basic assumption introduce shared method variance among self-ratings on different response formats. For example, socially desirable responding can inflate ratings of positive experiences across response formats. Thus, Green et al.'s (1993) study assumed rather than tested the validity of self-ratings of momentary affective experiences. At best, their study was able to examine the contribution of stylistic tendencies in the use of specific response formats to variance in mood ratings, but these effects are known to be small (Schimmack, Bockenholt, & Reizenzein, 2002). In sum, Green et al.'s (1993) article illustrates the importance of critically examining the similarity of methods in a multi-method study. Studies that use multiple self-report measures that vary response formats, scales, or measurement occasions should not be considered multi-method studies that can be used to examine construct validity.

Example 2: Three different measures

The second example of a multi-method study also examined the relation between Positive Affect and Negative Affect (Diener et al., 1995). However, it differs from the previous example in two important ways. First, the authors used more dissimilar methods that are less likely to violate the independence assumption, namely, self-report of affect in the past month, averaged daily affect ratings over a 6 week period and averaged ratings of general affect by multiple informants. Although these are different methods, it is possible that these methods are not strictly independent. For example, Diener et al. (1995) acknowledge that all three measures could be influenced by impression management. That is, retrospective and daily self-ratings could be influenced by social desirable responding, and informant ratings could be influenced by targets' motivation to hide negative emotions from others. A common influence of impression management on all three methods would inflate validity estimates of all three methods.

For this paper, I used Diener et al.'s (1995) multi-method data to estimate CVCs for the three methods as measures of general dispositions that influence people's positive and negative affective experiences. I used the data from Diener et al.'s (1995) Table 15 that are reproduced in Table 1. I used MPLUS5.1 for these analyses and all subsequent analyses (Muthén & Muthén, 2008). I fitted a simple model with a single latent variable that represents a general disposition that has causal effects on the three measures. Model fit was perfect because a model with three variables and three parameters has zero degrees of freedom and can perfectly reproduce the observed pattern of correlations. The perfect fit

Table 1. Cross-method correlations in Diener et al. (1995)

| | S | D | I |
|---------------|-----|-----|-----|
| Self (S) | — | .67 | .39 |
| Daily (D) | .68 | — | .34 |
| Informant (I) | .53 | .55 | — |

Note. PA below diagonal, NA above diagonal.

Table 2. Construct validity coefficients for Diener et al. (1995) multi-method data

| | CVC positive affect | | CVC negative affect | |
|-----------|---------------------|-----|---------------------|-----|
| Self | .81 [.75 .89] | 66% | .88 [.75 1.00] | 77% |
| Daily | .84 [.76 .92] | 71% | .76 [.65 .88] | 58% |
| Informant | .65 [.56 .75] | 42% | .45 [.32 .57] | 20% |

Note. CVC = construct validity coefficient [95% confidence interval], % = percentage of valid variance.

implies that CVC estimates are unbiased if the model assumptions are correct, but it also implies that the data are unable to test model assumptions.

These results suggest impressive validity of self-ratings of affect (Table 2). In contrast, CVC estimates of informant ratings are considerably lower, despite the fact that informant ratings are based on averages of several informants. The non-overlapping confidence intervals for self-ratings and informant ratings indicate that this difference is statistically significant. There are two interpretations of this pattern. On the one hand, it is possible that informants are less knowledgeable about targets' affective experiences. After all, they do not have access to information that is only available introspectively. However, this privileged information does not guarantee that self-ratings are more valid because individuals only have privileged information about their momentary feelings in specific situations rather than the internal dispositions that influence these feelings. On the other hand, it is possible that retrospective and daily self-ratings share method variance and do not fulfil the independence assumption. In this case, the causal model would provide inflated estimates of the validity of self-ratings because it assumes that stronger correlations between retrospective and daily self-ratings reveal higher validity of these methods, when in reality the higher correlation is caused by shared method effects. A study with three methods is unable to test these alternative explanations.

Example 3: Informants as multiple methods

One limitation of Diener et al.'s (1995) study was the aggregation of informant ratings. Although aggregated informant ratings provide more valid information than ratings by a single informant, the aggregation of informant ratings destroys valuable information about the correlations among informant ratings. The example in Figure 1 illustrated that ratings by multiple informants provide one of the easiest ways to measure dispositions with multiple methods because informants are more likely to base their ratings on different situations, which is necessary to reveal the influence of internal dispositions. Example 3 shows how ratings by multiple informants can be used in construct validation research. The data for this example are based on multi-method data from the Riverside Accuracy Project (Funder, 1995; Schimmack, Oishi, Furr, & Funder, 2004). To make the CVC estimates

Table 3. Cross-rater correlations in Funder's Riverside Accuracy Project

| | S | P | C | H |
|--------------|-----|-----|-----|-----|
| Self (S) | — | .41 | .43 | .27 |
| Parent (P) | .27 | — | .20 | .36 |
| College (C) | .39 | .13 | — | .27 |
| Hometown (H) | .35 | .42 | .19 | — |

Note. Cheerfulness below diagonal, depression above diagonal; self = self-ratings, parent = ratings by parents, college = ratings by college friends, hometown = ratings by hometown friends.

comparable to those based on the previous example, I used scores on the depression and cheerfulness facets of the NEO-PI-R (Costa & McCrae, 1992). These facets are designed to measure affective dispositions. The multi-method model used self-ratings and informant ratings by parents, college friends and hometown friends as different methods. Table 3 shows the correlation matrices for cheerfulness and depression.

I first fitted a causal model that assumed independence of all methods to the data. The model also included sum scores of observed measures to examine the validity of aggregated informant ratings and an aggregated measure of all four raters (Figure 2). Model fit was evaluated using standard criteria of model fit, namely, comparative fit index (CFI) > .95, root mean square error of approximation (RMSEA) < .06 and standardized root mean residuals (SRMR) < .08. Neither cheerfulness, χ^2 (df = 2, N = 222) = 11.30, p < .01, CFI = .860, RMSEA = .182, SRMR = .066, nor depression, χ^2 (df = 2, N = 222) = 8.31, p = .02, CFI = .915, RMSEA = .150, SRMR = .052, had acceptable CFI and RMSEA values. One possible explanation for this finding is that self-ratings are not independent of informant ratings because self-ratings and informant ratings could be partially based on overlapping situations. For example, self-ratings of cheerfulness could be heavily influenced by the same situations that are also used by college friends to rate cheerfulness (e.g. parties). In this case, some of the agreement between self-ratings and informant ratings by college friends would reflect the specific situational factors of overlapping situations, which leads to shared variance between these ratings that does not reflect the general disposition. In contrast, it is more likely that informant ratings are

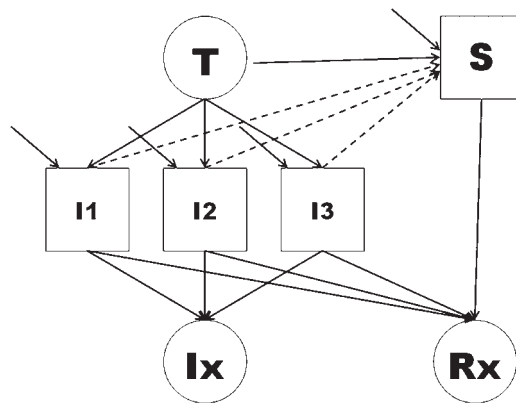


Figure 2. Causal model of multi-method data. Note. T = trait (general disposition); I1-I3 = informant reports; S = self-report; Ix = aggregated informant reports; Rx = aggregated informant and self-reports, arrows without origin represent error variance due to rating biases, situation specific variance and random error.

independent of each other because informants are less likely to rely on the same situations (Funder, 1995). For example, college friends may rely on different situations than parents. To examine this possibility, I fitted a model that included additional relations between self-ratings and informant ratings (dotted lines in Figure 2). For cheerfulness, an additional relation between self-ratings and ratings by college friends was sufficient to achieve acceptable model fit, χ^2 ($df = 1$, $N = 222$) = 0.08, $p = .78$, CFI = 1.00, RMSEA = .000, SRMR = .005. For depression, additional relations of self-ratings to ratings by college friends and parents were necessary to achieve acceptable model fit. Model fit of this model was perfect because it has zero degrees of freedom. In these models, CVC can no longer be estimated by factor loadings alone because some of the valid variance in self-ratings is also shared with informant ratings. In this case, CVC estimates represent the combined total effect of the direct effect of the latent disposition factor on self-ratings and the indirect effects that are mediated by informant ratings. I used the model indirect option of MPLUS5.1 to estimate the total effects. The model also included sum scores with equal weights for the three informant ratings and all four ratings. I used the model indirect option to estimate the total effect of the latent factor on both sum scores to estimate the CVCs of aggregated ratings. Table 4 lists the CVC estimates for the four ratings and the two measures based on aggregated ratings.

The CVC estimates of self-ratings are considerably lower than those based on Diener et al.'s (1995) data. Moreover, the results suggest that in this study aggregated informant ratings are more valid than self-ratings, although the confidence intervals overlap. The results for the aggregated measure of all four raters show that adding self-ratings to informant ratings did not increase validity above and beyond the validity obtained by aggregating informant ratings.

These results should not be taken too seriously because they are based on a single, relatively small sample. Moreover, it is important to emphasize that these CVC estimates depend on the assumption that informant ratings do not share method variance. Violation of this assumption would lead to an underestimation of the validity of self-ratings. For example, an alternative assumption would be that personality changes. As a result, parent ratings and ratings by hometown friends may share variance because they are based in part on situations before personality changed, whereas college friends' ratings are based on more recent situations. This model fits the data equally well and leads to much higher estimates of CV in self-ratings. To test these competing models it would be necessary to include additional measures. For example, standardized laboratory tasks and biological measures could be added to the design to separate valid variance from shared rating biases by informants.

Table 4. Construct validity coefficients for self and informant ratings of cheerfulness and depression

| | CVC cheerfulness | | CVD depression | |
|----------------|------------------|-----|----------------|-----|
| Self (S) | .48 [.30 .65] | 23% | .39 [.17 .61] | 15% |
| Parent (P) | .58 [.38 .77] | 34% | .51 [.26 .76] | 26% |
| College (C) | .25 [.05 .45] | 6% | .38 [.17 .60] | 14% |
| Hometown (H) | .73 [.51 .95] | 53% | .70 [.39 1.00] | 48% |
| All Informants | .75 [.59 .92] | 56% | .74 [.57 .90] | 54% |
| All Ratings | .75 [.59 .91] | 56% | .70 [.52 .88] | 49% |

Note. Self = self-ratings, parent = ratings by parents, college = ratings by college friends, hometown = ratings by hometown friends.

These inconsistent findings might suggest that it is futile to obtain wildly divergent quantitative estimates of construct validity. However, the same problem arises in other research areas and it can be addressed by designing better studies that test assumptions that cannot be tested in existing data sets. In fact, I believe that publication of conflicting validity estimates will stimulate research on construct validity, whereas the view of construct validation research as an obscure process without clear results has obscured the lack of knowledge about the validity of personality measures.

IMPLICATIONS

I used two multi-method datasets to illustrate how causal models of multi-method data can be used to estimate the validity of personality measures. The studies produced different results. It is not the purpose of this paper to examine the sources of disagreement. The results merely show that it is difficult to make general claims about the validity of commonly used personality measures. Until more precise information becomes available, the results suggest that about 30–70% of the variance in self-ratings and single informant ratings is CV. Until more precise estimates become available I suggest an estimate of $50 \pm 20\%$ as a rough estimate of construct validity of personality ratings. I suggest the verbal labels low validity for measures with less than 30% CV (e.g. implicit measures of well-being, Walker & Schimmack, 2008), moderate validity for measures with 30–70% CV (most self-report measures of personality traits) and high validity for measures with more than 70% CV (self-ratings of height and weight). Subsequently, I briefly discuss the practical implications of using self-report measures with moderate validity to study the causes and consequences of personality dispositions.

Correction for invalidity

Measurement error is nearly unavoidable, especially in the measurement of complex constructs such as personality dispositions. Schmidt and Hunter (1996) provided 26 examples of how the failure to correct for measurement error can bias substantive conclusions. One limitation of their important article was the focus on random measurement error. The main reason is probably that information about random measurement error is readily available. However, invalid variance due to systematic measurement error is another factor that can distort research findings. Moreover, given the moderate amount of valid variance in personality measures, corrections for invalidity are likely to have more dramatic practical implications than corrections for unreliability. The following examples illustrate this point.

Heritability of personality dispositions

Hundreds of twin studies have examined the similarity between MZ and DZ twins to examine the heritability of personality characteristics. A common finding in these studies are moderate to large MZ correlations ($r = .3-.5$) and small to moderate ($r = .1-.3$) DZ correlations. This finding has led to the conclusion that approximately 40% of the variance is heritable and 60% of the variance is caused by environmental factors. However, this interpretation of twin data fails to take measurement error into account. As it turns out, MZ correlations approach, if not exceed, the amount of validity variance in personality measures as estimated by multi-method data. In other words, ratings by two different individuals of two different individuals (self-ratings by MZ twins) tend to correlate as

highly with each other as those of a single individual (self ratings and informant ratings of a single target). This finding suggests that heritability estimates based on mono-method studies severely underestimate heritability of personality dispositions (Riemann, Angleitner, & Strelau, 1997). A correction for invalidity would suggest that most of the valid variance is heritable (Lykken & Tellegen, 1996). However, it is problematic to apply a direct correction for invalidity to twin data because this correction assumes that the independence assumption is valid. It is better to combine a multi-method assessment with a twin design (Riemann et al., 1997). It is also important to realize that multi-method models focus on internal dispositions rather than act frequencies. It makes sense that heritability estimates of internal dispositions are higher than heritability estimates of act frequencies because act frequencies are also influenced by situational factors.

Stability of personality dispositions

The study of stability of personality has a long history in personality psychology (Conley, 1984). However, empirical conclusions about the actual stability of personality are hampered by the lack of good data. Most studies have relied on self-report data to examine this question. Given the moderate validity of self-ratings, it is likely that studies based on self-ratings underestimate true stability of personality. Even corrections for unreliability alone are sufficient to achieve impressive stability estimates of $r = .98$ over a 1-year interval (Conley, 1984). The evidence for stability of personality from multi-method studies is even more impressive. For example, one study reported a retest correlation of $r = .46$ over a 26-year interval for a self-report measure of neuroticism (Conley, 1985). It seems possible that personality could change considerably over such a long time period. However, the study also included informant ratings of personality. Self-informant agreement on the same occasion was also $r = .46$. Under the assumption that self-ratings and informant ratings are independent methods and that there is no stability in method variance, this pattern of correlations would imply that variation in neuroticism did not change at all over this 26-year period ($.46/.46 = 1.00$). However, this conclusion rests on the validity of the assumption that method variance is not stable. Given the availability of longitudinal multi-method data it is possible to test this assumption. The relevant information is contained in the cross-informant, cross-occasion correlations. If method variance was unstable, these correlations should also be $r = .46$. In contrast, the actual correlations are lower, $r = .32$. This finding indicates that (a) personality dispositions changed and (b) there is some stability in the method variance. However, the actual stability of personality dispositions is still considerably higher ($r = .32/.46 = .70$) than one would have inferred from the observed retest correlation $r = .46$ of self-ratings alone. A retest correlation of $r = .70$ over a 26-year interval is consistent with other estimates that the stability of personality dispositions is about $r = .90$ over a 10-year period and $r = .98$ over a 1-year period (Conley, 1984; Terracciano, Costa, & McCrae, 2006). The failure to realize that observed retest correlations underestimate stability of personality dispositions can be costly because it gives personality researchers a false impression about the likelihood of finding empirical evidence for personality change. Given the true stability of personality it is necessary to wait a long time or to use large sample sizes and probably best to do both (Mroczek, 2007).

Prediction of behaviour and life outcomes

During the person-situation debate, it was proposed that a single personality trait predicts less than 10% of the variance in actual behaviours. However, most of these studies relied on

self-ratings of personality to measure personality. Given the moderate validity of self-ratings, the observed correlation severely underestimates the actual effect of personality traits on behaviour. For example, a recent meta-analysis reported an effect size of conscientiousness on GPA of $r = .24$ (Noftle & Robins, 2007). Ozer (2007) points out that strictly speaking the correlation between self-reported conscientiousness and GPA does not represent the magnitude of a causal effect. That is, unfortunately for undergraduate students, simply increasing one's rating of conscientiousness does not increase GPA (Robins & Beer, 2001). Thus, a different causal model must explain the correlation between self-ratings of conscientiousness and GPA. One plausible model postulates that conscientiousness is a personality disposition that influences self-ratings of conscientiousness and performance on academic tests. Assuming 40% valid variance in self-report measures of conscientiousness (DeYoung, 2006), the true effect size of a conscientious disposition on GPA is $r = .38$ ($.24/.40^{1/2}$). As a result, the amount of explained variance in GPA increases from 6 to 14%. Once more, failure to correct for invalidity in personality measures can be costly. For example, a personality researcher might identify seven causal factors that independently produce observed effect size estimates of $r = .24$, which suggests that these seven factors explain less than 50% of the variance in GPA ($7 \times .24^2 = 42\%$). However, decades of future research are unable to uncover additional predictors of GPA. The reason could be that the true amount of explained variance is nearly 100% and that the unexplained variance is due to invalid variance in personality measures ($7 \times .38^2 = 100\%$).

CONCLUSION

This paper provided an introduction to the logic of a multi-method study of construct validity. I showed how causal models of multi-method data can be used to obtain quantitative estimates of the construct validity of personality measures. I showed that accurate estimates of construct validity depend on the validity of the assumptions underlying a causal model of multi-method data such as the assumption that methods are independent. I also showed that multi-method studies of construct validity require postulating a causal construct that can influence and produce covariances among independent methods. Multi-method studies for other constructs such as actual behaviours or act frequencies are more problematic because act frequencies do not predict a specific pattern of correlations across methods. Finally, I presented some preliminary evidence that commonly used self-ratings of personality are likely to have a moderate amount of valid variance that falls broadly in a range from 30 to 70% of the total variance. This estimate is consistent with meta-analyses of self-informant agreement (Connolly, Kavanagh, & Viswesvaran, 2007; Schneider & Schimmack, 2009). However, the existing evidence is limited and more rigorous tests of construct validity are needed. Moreover studies with large, representative samples are needed to obtain more precise estimates of construct validity. Hopefully, this paper will stimulate more research in this fundamental area of personality psychology by challenging the description of construct validity research as a Kafkaesque pursuit of an elusive goal that can never be reached (cf. Borsboom, 2006). Instead empirical studies of construct validity are a viable and important scientific enterprise that faces the same challenges as other studies in personality psychology that try to make sense of correlational data.

REFERENCES

- Allport, G. W., & Odbert, H. S. (1936). Trait-names a psycho-lexical study. *Psychological Monographs*, 47(1), 1–171.
- Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, 72(4), 845–876.
- Blanton, H., & Jaccard, J. (2006). Arbitrary metrics redux. *American Psychologist*, 61(1), 62–71.
- Block, J. (1989). Critique of the act frequency approach to personality. *Journal of Personality and Social Psychology*, 56(2), 234–245.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
- Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, 30(6), 505–514.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203–219.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Chaplin, W. F. (2007). Moderator and mediator models in personality research: A basic introduction. In R. W. Robins, C. R. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (602–632). New York, NY: Guilford Press.
- Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, 5(1), 11–25.
- Conley, J. J. (1985). Longitudinal stability of personality traits: A multitrait-multimethod-multi-occasion analysis. *Journal of Personality and Social Psychology*, 49(5), 1266–1282.
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment*, 15(1), 110–117.
- Costa, J. P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEOPI-R) and Five Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, 91(6), 1138–1151.
- Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, 69(1), 130–141.
- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, 8(1), 38–60.
- Funder, D. C. (1991). Global traits—a Neo-Allportian approach to personality. *Psychological Science*, 2(1), 31–39.
- Funder, D. C. (1995). On the accuracy of personality judgment—a realistic approach. *Psychological Review*, 102(4), 652–670.
- Goldberg, L. R. (1992). The social psychology of personality. *Psychological Inquiry*, 3, 89–94.
- Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology*, 64(6), 1029–1041.
- Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment*, 89(2), 167–187.
- John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, C. R. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (461–494). New York, NY: Guilford Press.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112(1), 165–172.
- Kroh, M. (2005). Effects of interviews during body weight checks in general population surveys. *Gesundheitswesen*, 67(8–9), 646–655.

- Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science*, 7(3), 186–189.
- McCrae, R. R., & Costa, P. T. (1995). Trait explanations in personality psychology. *European Journal of Personality*, 9(4), 231–252.
- Mroczek, D. K. (2007). The analysis of longitudinal data in personality research. In R. W. Robins, C. R. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 543–556). New York, NY, US: Guilford Press.
- Muthén, L. K., & Muthén, B. O. (2008). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93(1), 116–130.
- Ozer, D. J. (2007). *Evaluating effect size in personality research*. In R. W. Robins, C. R. Fraley, & R. F. Krueger (Eds.). New York, NY, US: Guilford Press.
- Riemann, R., Angleitner, A., & Strelau, J. (1997). Genetic and environmental influences on personality: A study of twins reared together using the self- and peer-report NEO-FFI scales. *Journal of Personality*, 65(3), 449–475.
- Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology*, 80(2), 340–352.
- Rowland, M. L. (1990). Self-reported weight and height. *American Journal of Clinical Nutrition*, 52(6), 1125–1133.
- Schimmack, U. (2007). The structure of subjective well-being. In M. Eid, & R. J. Larsen (Eds.), *The science of subjective well-being* (pp. 97–123). New York: Guilford.
- Schimmack, U., Bockenholt, U., & Reisenzein, R. (2002). Response styles in affect ratings: Making a mountain out of a molehill. *Journal of Personality Assessment*, 78(3), 461–483.
- Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, 89(3), 395–406.
- Schimmack, U., Oishi, S., Furr, R. M., & Funder, D. C. (2004). Personality and life satisfaction: A facet-level analysis. *Personality and Social Psychology Bulletin*, 30(8), 1062–1075.
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1(2), 199–223.
- Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: A meta-analysis. *Social Indicators Research*, 94, 363–376.
- Simms, L. J., & Watson, D. (2007). The construct validation approach to personality scale construction. In R. W. Robins, C. R. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (240–258). New York, NY: Guilford Press.
- Terracciano, A., Costa, J. P. T., & McCrae, R. R. (2006). Personality plasticity after age 30. *Personality and Social Psychology Bulletin*, 32, 999–1009.
- Walker, S. S., & Schimmack, U. (2008). Validity of a happiness Implicit Association Test as a measure of subjective well-being. *Journal of Research in Personality*, 42(2), 490–497.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070.
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, 76(5), 820–838.