

JYX



This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Rönkkö, Mikko; Cho, Eunseong

Title: An Updated Guideline for Assessing Discriminant Validity

Year: 2022

Version: Published version

Copyright: © The Author(s) 2020

Rights: CC BY-NC 4.0

Rights url: <https://creativecommons.org/licenses/by-nc/4.0/>

Please cite the original version:

Rönkkö, M., & Cho, E. (2022). An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods*, 25(1), 6-14. <https://doi.org/10.1177/1094428120968614>

An Updated Guideline for Assessing Discriminant Validity

Mikko Rönkkö¹  and Eunseong Cho² 

Organizational Research Methods
1-42

© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/1094428120968614

journals.sagepub.com/home/orm



Abstract

Discriminant validity was originally presented as a set of empirical criteria that can be assessed from multitrait-multimethod (MTMM) matrices. Because datasets used by applied researchers rarely lend themselves to MTMM analysis, the need to assess discriminant validity in empirical research has led to the introduction of numerous techniques, some of which have been introduced in an ad hoc manner and without rigorous methodological support. We review various definitions of and techniques for assessing discriminant validity and provide a generalized definition of discriminant validity based on the correlation between two measures after measurement error has been considered. We then review techniques that have been proposed for discriminant validity assessment, demonstrating some problems and equivalencies of these techniques that have gone unnoticed by prior research. After conducting Monte Carlo simulations that compare the techniques, we present techniques called $CI_{CFA}(sys)$ and $\chi^2(sys)$ that applied researchers can use to assess discriminant validity.

Keywords

discriminant validity, Monte Carlo simulation, measurement, confirmatory factor analysis, validation, average variance extracted, heterotrait-monotrait ratio, cross-loadings

Among various types of validity evidence, organizational researchers are often required to assess the discriminant validity of their measurements (e.g., J. P. Green et al., 2016). However, there are two problems. First, the current applied literature appears to use several different definitions for discriminant validity, making it difficult to determine which procedures are ideal for its assessment. Second, existing guidelines are far from the practices of organizational researchers. As originally presented, “*more than one method* must be employed in the [discriminant] validation process” (Campbell & Fiske, 1959, p. 81), and consequently, literature on discriminant validation has focused on techniques that require that multiple distinct measurement methods be used (Le et al., 2009;

¹Jyväskylä University School of Business and Economics, University of Jyväskylä, Finland

²College of Business Administration, Kwangwoon University, Seoul, Republic of Korea

Corresponding Author:

Eunseong Cho, College of Business Administration, Kwangwoon University, 20 Kwangwoonro, Nowon-gu, Seoul 01897, Republic of Korea.

Email: bene@kw.ac.kr

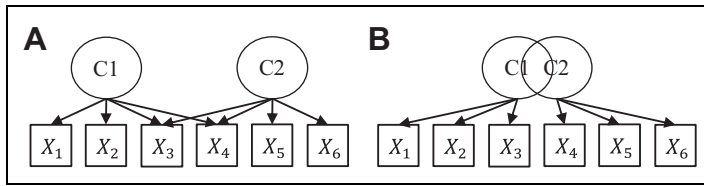


Figure 1. Two definitions of discriminant validity, as shown in the *AMJ* and *JAP* articles. (A) Items measure more than one construct (i.e., cross-loadings). (B) Constructs are not empirically distinct (i.e., high correlation).

Woehr et al., 2012), which is rare in applied research. To fill this gap, various less-demanding techniques have been proposed, but few of these techniques have been thoroughly scrutinized.

We present a comprehensive analysis of discriminant validity assessment that focuses on the typical case of single-method and one-time measurements, updating or challenging some of the recent recommendations on discriminant validity (Henseler et al., 2015; J. A. Shaffer et al., 2016; Voorhees et al., 2016). We start by reviewing articles in leading organizational research journals and demonstrating that the concept of discriminant validity is understood in at least two different ways; consequently, empirical procedures vary widely. We then assess what researchers mean by discriminant validity and synthesize this meaning as a definition. After defining what discriminant validity means, we provide a detailed discussion of each of the techniques identified in our review. Finally, we compare the techniques in a comprehensive Monte Carlo simulation. The conclusion section is structured as a set of guidelines for applied researchers and presents two techniques. One is $CI_{CFA}(sys)$, which is based on the confidence intervals (CIs) in confirmatory factor analysis (CFA), and the other is $\chi^2(sys)$, a technique based on model comparisons in CFA.

Current Practices of Discriminant Validity Assessment in Organizational Research

To understand what organizational researchers try to accomplish by assessing discriminant validity, we reviewed all articles published between 2013 and 2016 by the *Academy of Management Journal* (*AMJ*), the *Journal of Applied Psychology* (*JAP*), and *Organizational Research Methods* (*ORM*). We included only studies that directly collected data from respondents through multiple-item scales. A total of 97 out of 308 papers in *AMJ*, 291 out of 369 papers in *JAP*, and 5 out of 93 articles in *ORM* were included.

The term “discriminant validity” was typically used without a definition or a citation, giving the impression that there is a well-known and widely accepted definition of the term. However, the few empirical studies that defined the term revealed that it can be understood in two different ways: One group of researchers used discriminant validity as a property of a measure and considered a measure to have discriminant validity if it measured the construct that it was supposed to measure but not any other construct of interest (A in Figure 1). For these researchers, discriminant validity means that “two measures are tapping separate constructs” (R. Krause et al., 2014, p. 102) or that the measured “scores are not (or only weakly) associated with potential confounding factors” (De Vries et al., 2014, p. 1343). Another group of researchers used discriminant validity to refer to whether two constructs were empirically distinguishable (B in Figure 1). For this group of researchers, the term referred to “whether the two variables . . . are distinct from each other” (Hu & Liden, 2015, p. 1110).

These definitions were also implicitly present in other studies through the use of various statistical techniques summarized in Table 1. In the studies that assessed whether measures of two constructs were empirically distinguishable, comparison of CFA models was the most

Table 1. Techniques Used to Assess Discriminant Validity in *AMJ*, *JAP*, and *ORM*.

	AMJ (n = 27)		JAP (n = 73)		ORM (n = 5)	
Techniques using correlation estimates						
Scale score correlation (ρ_{SS})	7	25.9%	8	11.0%	3	60.0%
Factor correlation (ρ_{CFA})	0	0.0%	2	2.7%	1	20.0%
Disattenuated correlation (ρ_{DTR})	0	0.0%	1	1.4%	1	20.0%
Techniques to compare AVE to a certain value						
AVE _{CFA} vs. Square of ρ_{CFA} (AVE/SV _{CFA})	2	7.4%	4	5.5%	1	20.0%
AVE _{CFA} vs. Square of ρ_{SS} (AVE/SV _{SS})	1	3.7%	1	1.4%	0	0.0%
AVE _{CFA} vs. .5	2	7.4%	1	1.4%	0	0.0%
AVE _{PLS} vs. Square of ρ_{SS}	2	7.4%	0	0.0%	0	0.0%
Techniques to show low cross-loadings						
CFA (structure coefficients)	2	7.4%	0	0.0%	0	0.0%
Exploratory factor analysis	1	3.7%	0	0.0%	0	0.0%
Techniques using fit indices of CFA models						
No comparison (only the proposed model)	3	11.1%	1	1.4%	0	0.0%
Compared with nested models with fewer factors (χ^2 (merge))	8	29.6%	43	58.9%	1	20.0%
Compared with model with fixed correlation of 1 (χ^2 (1))	4	14.8%	1	1.4%	2	40.0%
Compared with model with fixed correlation of 1 (CFI(1))	0	0.0%	0	0.0%	1	20.0%
Techniques requiring multiple measurement methods						
GCES approach	0	0.0%	0	0.0%	1	20.0%
MTMM approach	0	0.0%	1	1.4%	0	0.0%
Generalizability theory approach	0	0.0%	1	1.4%	0	0.0%
Techniques that are difficult to classify						
CFA results not presented in detail	1	3.7%	4	5.5%	0	0.0%
No clear evidence provided	0	0.0%	3	4.1%	0	0.0%
Comparison with existing research results	0	0.0%	2	2.7%	0	0.0%
Experimental results as expected	0	0.0%	2	2.7%	0	0.0%

Note: The sum exceeds 100% because some studies use multiple techniques. *AMJ* = *Academy of Management Journal*; *JAP* = *Journal of Applied Psychology*; *ORM* = *Organizational Research Methods*; CFA = confirmatory factor analysis; AVE = average variance extracted; ρ_{CFA} = factor correlation obtained from CFA; ρ_{DTR} = disattenuated correlation using tau-equivalent reliability; AVE_{CFA} = AVE obtained from CFA; AVE_{PLS} = AVE obtained from partial least squares; GCES = generalized coefficient of equivalence and stability; MTMM = multitrait-multimethod. For a detailed description of the symbols, see Table 4.

common technique, followed by calculating a correlation that was compared against a cutoff. The CFA comparison was most commonly used to assess whether two factors could be merged, but a range of other comparisons were also presented. In the correlation-based techniques, correlations were often calculated using scale scores; sometimes, correction for attenuation was used, whereas other times, estimated factor correlations were used. These correlations were evaluated by comparing the correlations with the square root of average variance extracted (AVE) or comparing their CIs against cutoffs. Some studies demonstrated that correlations were not significantly different from zero, whereas others showed that correlations were significantly different from one. In the studies that considered discriminant validity as the degree to which each item measured one construct only and not something else, various factor analysis techniques were the most commonly used, typically either evaluating the fit of the model where cross-loadings were constrained to be zero or estimating the cross-loadings and comparing their values against various cutoffs.

Our review also revealed two findings that go beyond cataloging the discriminant validation techniques. First, no study evaluated discriminant validity as something that can exist to a degree but instead used the statistics to answer a yes/no type of question. Second, many techniques were used

differently than originally presented. Given the diversity of how discriminant validity is conceptualized, the statistics used in its assessment, and how these statistics are interpreted, there is a clear need for a standard understanding of which technique(s) should be used and how the discriminant validity evidence produced by these techniques should be evaluated. However, to begin the evaluation of the various techniques, we must first establish a definition of discriminant validity.

Defining Discriminant Validity

Most methodological work defines discriminant validity by using a correlation but differs in what specific correlation is used, as shown in Table 2. For example, defining discriminant validity in terms of a (true) correlation between constructs implies that a discriminant validity problem cannot be addressed with better measures. In contrast, defining discriminant validity in terms of measures or estimated correlation ties it directly to particular measurement procedures. Some studies (i.e., categories 3 and 4 in Table 2) used definitions involving both constructs and measures stating that a measure should not correlate with or be affected by an unrelated construct. Moreover, there is no consensus on what kind of attribute discriminant validity is; some studies (Bagozzi & Phillips, 1982; Hamann et al., 2013; Reichardt & Coleman, 1995; J. A. Shaffer et al., 2016) define discriminant validity as a matter of degree, while others (Schmitt & Stults, 1986; Werts & Linn, 1970) define discriminant validity as a dichotomous attribute. These findings raise two important questions: (a) Why is there such diversity in the definitions? and (b) How exactly should discriminant validity be defined?

Origin of the Concept of Discriminant Validity

The term “discriminant validity” was coined by Campbell and Fiske (1959), who presented a validation technique based on the long-standing idea that tests can be invalidated by too high correlations with unrelated tests (Campbell, 1960; Thorndike, 1920). However, the term was introduced without a clear definition of the concept (Reichardt & Coleman, 1995); instead, the article focuses on *discriminant validation* or how discriminant validity can be shown empirically using multitrait-multimethod (MTMM) matrices. The original criteria, illustrated in Table 3, were as follows: (a) two variables that measure the same trait (T1) with two different methods (M1, M2) should correlate more highly than any two variables that measure two different traits (T1, T2) with different methods (M1, M2); (b) two variables that measure the same trait (T1) with two different methods (M1, M2) should correlate more highly than any two variables that measure two different traits (T1, T2) but use the same method (M1); and (c) the pattern of correlations between variables that measure different traits (T1, T2) should be very similar across different methods (M1, M2) (Campbell & Fiske, 1959).

Generalizing the concept of discriminant validity outside MTMM matrices is not straightforward. Indeed, the definitions shown in Table 2 show little connections to the original MTMM matrices. A notable exception is Reichardt and Coleman (1995), who criticized the MTMM-based criteria for being dichotomous and declared that a natural (dichotomous) definition of discriminant validity outside the MTMM context would be that two measures x_1 and x_2 have discriminant validity if and only if x_1 measures construct T1 but not T2, x_2 measures T2 but not T1, and the two constructs are not perfectly correlated. They then concluded that a preferable continuous definition would be “*the degree to which the absolute value of the correlation between the two constructs differs from one*” (Reichardt & Coleman, 1995, p. 516). A similar interpretation was reached by McDonald (1985), who noted that two tests have discriminant validity if “the common factors are correlated, but the correlations are low enough for the factors to be regarded as distinct ‘constructs’” (p. 220).

Table 2. Definitions of Discriminant Validity in Existing Studies.

Category	Definition/Description of Technique
1: True or estimated correlation between constructs ^a	<p>“[T]he degree to which the absolute value of the correlation between the two constructs differ from one.” (Reichardt & Coleman, 1995, p. 516)</p> <p>“Evidence of discriminant validity exists if other constructs do not correlate strongly enough with the construct of interest to suggest that they measure the same construct.” (McKenny et al., 2013, p. 156)</p> <p>“Discriminant validation implies that correlation between traits is low. If both traits were identical, the correlation between the trait factors would be near one.” (Kenny, 1976, p. 251)</p> <p>“[D]iscriminant validity exists when estimates of the trait correlations were two or more standard errors below 1.0.” (Schmitt & Stults, 1986, p. 18)</p> <p>“[D]iscriminant validity consists of demonstrating that the true correlation of [two traits] is meaningfully less than unity.” (Werts & Linn, 1970, p. 208)</p>
2: Correlation between measures	<p>“[A] test [should] not correlate too highly with measures from which it is supposed to differ.” (Campbell, 1960, p. 548)</p> <p>“[A test] correlates less well or not all with tests with which theory implies it should not correlate well.” (McDonald, 1985, p. 220)</p> <p>“[T]he extent to which measures of theoretically distinct constructs are unrelated empirically to one another.” (J. A. Shaffer et al., 2016, p. 82)</p> <p>“[I]f two or more concepts are unique, then valid measures of each should not correlate too highly.” (Bagozzi et al., 1991, p. 425)</p> <p>“[T]he degree of divergence among indicators that are designed to measure different constructs.” (Hamann et al., 2013, p. 72)</p> <p>“[T]he degree to which measures of distinct concepts differ.” (Bagozzi & Phillips, 1982, p. 469)</p> <p>“Measures of different attributes should . . . not correlated to an extremely high degree.” (Nunnally & Bernstein, 1994, p. 93)</p> <p>“[A] measure of a construct is unrelated to indicators of theoretically irrelevant constructs in the same domain.” (Strauss & Smith, 2009, p. 1)</p>
3: Correlation between measure and other construct	<p>“[D]iscriminant validity is shown when each measurement item correlates weakly with all other constructs except for the one to which it is theoretically associated.” (Gefen & Straub, 2005, p. 92)</p> <p>“Discriminant validity is inferred when scores from measures of different constructs do not converge. It thus provides information about whether scores from a measure of a construct are unique rather than contaminated by other constructs.” (Schwab, 2013, p. 33)</p>
4: Combination of categories 1 and 3.	<p>“[T]he item’s . . . loading on constructs other than the intended one is relevant to discriminant validity At the level of the constructs, this correlation tells us about discriminant validity.” (John & Benet-Martínez, 2000, p. 359)</p> <p>Voorhees et al. (2016) classified discriminant validity into the construct-level (i.e., low correlation) and the item-level (i.e., absence of cross-loading).</p>

a. Many articles in this category are ambiguous on whether discriminant validity is a property of a construct or a property of a scale from which construct correlation is estimated.

Constructs and Measures

Discriminant validity is sometimes presented as the property of a construct (Reichardt & Coleman, 1995) and other times as the property of its measures or empirical representations constructed from those measures (McDonald, 1985). This ambiguity may stem from the broader confusion over common factors and constructs: The term “construct” refers to the concept or trait being measured,

Table 3. Multitrait-Multimethod Correlation Matrix and Original Criteria for Discriminant Validity.

Traits		Method M1			Method M2			
		T1	T2	T3	T1	T2	T3	
M1	T1	1						Discriminant Validity: All MTHM > HTHM All MTHM > HTMM
	T2	HTMM ₁₁	1					
	T3	HTMM ₁₂	HTMM ₁₃	1				
M2	T1	MTHM ₂₁	HTHM ₂₄	HTHM ₂₇	1			HTMM ₁₁ ≈ HTMM ₃₁
	T2	HTHM ₂₂	MTHM ₂₅	HTHM ₂₈	HTMM ₃₁	1		HTMM ₁₂ ≈ HTMM ₃₂
	T3	HTHM ₂₃	HTHM ₂₆	MTHM ₂₉	HTMM ₃₂	HTMM ₃₃	1	HTMM ₁₃ ≈ HTMM ₃₃

Note: HTMM = same method and different traits (heterotrait-monomethod); MTHM = different methods and same trait (monotrait-heteromethod); HTHM = different methods and different traits (heterotrait-heteromethod).

whereas a common factor is part of a statistical model estimated from data (Maraun & Gabriel, 2013). Indeed, Campbell and Fiske (1959) define validity as a feature of a test or measure, not as a property of the trait or construct being measured. In fact, if one takes the realist perspective that constructs exist independently of measurement and can be measured in multiple different ways (Chang & Cartwright, 2008),¹ it becomes clear that we cannot use an empirical procedure to define a property of a construct.

The Role of the Factor Model

Factor analysis has played a central role in articles on discriminant validation (e.g., McDonald, 1985), but it cannot serve as a basis for a definition of discriminant validity for two reasons. First, validity is a feature of a test or a measure or its interpretation (Campbell & Fiske, 1959), not of any particular statistical analysis. Moreover, discriminant validity is often presented as a property of “an item” (Table 2), implying that the concept should also be applicable in the single-item case, where factor analysis would not be applicable. Second, a factor model where each item loads on only one factor may be too constraining for applied research (Asparouhov et al., 2015; Marsh et al., 2014; Morin et al., 2017; Rodriguez et al., 2016). For example, Marsh et al. (2014) note that in psychology research, the symptoms or characteristics of different disorders commonly overlap, producing non-negligible cross-loadings in the population. Constraining these cross-loadings to be zero can inflate the estimated factor correlations, which is problematic, particularly for discriminant validity assessment (Marsh et al., 2014). Moreover, a linear model where factors, error terms, and observed variables are all continuous (Bartholomew, 2007) is not always realistic. Indeed, a number of item response theory models have been introduced (Foster et al., 2017; Reise & Revicki, 2014) to address these scenarios and have been applied to assess discriminant validity using MTMM data (Jeon & Rijmen, 2014).

Generalized Definition of Discriminant Validity

We present a definition that does not depend on a particular model and makes it explicit that discriminant validity is a feature of a measure instead of a construct:² *Two measures intended to measure distinct constructs have discriminant validity if the absolute value of the correlation between the measures after correcting for measurement error is low enough for the measures to be regarded as measuring distinct constructs.*

This definition encompasses the early idea that even moderately high correlations between distinct measures can invalidate those measures if measurement error is present (Thorndike,

1920), which serves as the basis of discriminant validity (Campbell & Fiske, 1959). The definition can also be applied on both the scale level and the scale-item level. Consider the proposed definition in the context of the common factor model:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \quad (1)$$

where Σ is the interitem correlation matrix; Φ is the factor correlation matrix, where all correlations are assumed to be positive for simplicity; Λ is a factor pattern (loading) matrix; and Θ is the item error covariance matrix. Within this context, our definition can be understood in two equivalent ways:

$$\Sigma_{i,j} \ll \Lambda J \Lambda'_{i,j} \quad (2)$$

$$(\Lambda' \Lambda)^{-1} \Lambda' (\Sigma - \Theta) \Lambda (\Lambda' \Lambda)^{-1}_{k,l} \ll 1 \quad (3)$$

where J is a unit matrix (a matrix of ones) and \ll denotes much less than. Equation 2 is an item-level comparison (category 2 in Table 2), where the correlation between items i and j , which are designed to measure different constructs, is compared against the implied correlation when the items depend on perfectly correlated factors but are not perfectly correlated because of measurement error. Equation 3 shows an equivalent scale-level comparison (part of category 1 in Table 2) focusing on two distinct scales k and l . The factor correlations are solved from the interitem correlations by multiplying with left and right inverses of the factor pattern matrix to correct for measurement error and are then compared against a perfect correlation. Generalizing beyond the linear common factor model, Equation 3 can be understood to mean that *two scales intended to measure distinct constructs have discriminant validity if the absolute value of the correlation between two latent variables estimated from the scales is low enough for the latent variables to be regarded as representing distinct constructs*.

Our definition has several advantages over previous definitions shown in Table 2. First, it clearly states that discriminant validity is a feature of measures and not constructs and that it is not tied to any particular statistical test or cutoff (Schmitt, 1978; Schmitt & Stults, 1986). Second, the definition is compatible with both continuous and dichotomous interpretations, as it suggests the existence of a threshold, that is, a correlation below a certain level has no problem with discriminant validity but does not dictate a specific cutoff, thus also allowing the value of the correlation to be interpreted instead of simply tested. Third, the definition is not tied to any particular measurement process (e.g., single administration) but considers measurement error generally, thus supporting rater, transient, and other errors (Le et al., 2009; Schmidt et al., 2003). Fourth, the definition is not tied to either the individual item level or the multiple item scale level but works across both, thus unifying the category 1 and category 2 definitions of Table 2. Fifth, the definition does not confound the conceptually different questions of whether two measures measure different things (discriminant validity) and whether the items measure what they are supposed to measure and not something else (i.e., lack of cross-loadings in Λ , factorial validity),³ which some of the earlier definitions (categories 3 and 4 in Table 2) do.

This definition also supports a broad range of empirical practice: If considered on the scale level, the definition is compatible with the current tests, including the original MTMM approach (Campbell & Fiske, 1959). However, it is not limited to simple linear common factor models where each indicator loads on just one factor but rather supports any statistical technique including more complex factor structures (Asparouhov et al., 2015; Marsh et al., 2014; Morin et al., 2017; Rodriguez et al., 2016) and nonlinear models (Foster et al., 2017; Reise & Revicki, 2014) as long as these techniques can estimate correlations that are properly corrected for measurement error and supports scale-item level evaluations.

Table 4. Techniques Included in the Simulation of This Study.

Statistic/Technique	Symbol	Description/Criteria
Correlation estimation methods discussed in this study		
Scale score correlation	ρ_{SS}	Correlation between unit-weighted sums of scale scores (i.e., most commonly reported)
Factor correlation	ρ_{CFA}	Correlation obtained through confirmatory factor analysis (CFA) (i.e., Figure 3)
Disattenuated correlation using parallel reliability	ρ_{DPR}	ρ_{SS} is converted to an error-adjusted correlation using parallel reliability (i.e., the standardized alpha), also called the heterotrait-monotrait (HTMT) ratio
Disattenuated correlation using tau-equivalent reliability	ρ_{DTR}	ρ_{SS} is converted to an error-adjusted correlation using tau-equivalent reliability (i.e., the coefficient alpha)
Disattenuated correlation using congeneric reliability	ρ_{DCR}	ρ_{SS} is converted to an error-adjusted correlation using congeneric reliability (i.e., the composite reliability or omega)
Common notation for expressing correlation-related techniques		
The point estimate is less than a cutoff	$\rho_{XX}(\text{cut})$	For every factor pair, the point estimate of ρ_{XX} is less than a cutoff
The confidence interval (CI) is less than 1	$CI_{XX}(1)$	For every factor pair, the CI of ρ_{XX} is strictly less than one
The CI is less than a cutoff	$CI_{XX}(\text{cut})$	For every factor pair, the CI of ρ_{XX} is strictly less than a cutoff
Other correlation-related techniques		
Average variance extracted (AVE) compared with the SV obtained from factor correlations	AVE/SV_{CFA}	For every factor pair, the AVEs of the two factors are greater than the square of ρ_{CFA} (i.e., used as originally proposed by Fornell & Larcker, 1981a)
AVE compared with the SV obtained from scale score correlations	AVE/SV_{SS}	For every factor pair, the AVEs of the two factors are greater than the square of ρ_{SS} (i.e., common misuse)
Techniques that focus on model fit		
A chi-square comparison with a model with a fixed correlation of 1	$\chi^2(1)$	For every factor pair, the chi-square difference between the unconstrained model and the constrained model in which ρ_{CFA} is fixed at one is statistically significant
A chi-square comparison with a model that merges two factors into one	$\chi^2(\text{merge})$	For every factor pair, the chi-square difference between the unconstrained model and the constrained model in which the two factors are merged into one is statistically significant
A comparative fit index (CFI) comparison with a model with a fixed correlation of 1	$CFI(1)$	For every factor pair, the CFI difference between the unconstrained model and the constrained model in which ρ_{CFA} is fixed at one is greater than .002
A chi-square comparison with a model with a fixed correlation of a cutoff	$\chi^2(\text{cut})$	For every factor pair, the chi-square difference between the unconstrained model and the constrained model in which ρ_{CFA} is fixed at a cutoff is statistically significant
A CFI comparison with a model with a fixed correlation of a cutoff	$CFI(\text{cut})$	For every factor pair, the CFI difference between the unconstrained model and the constrained model in which ρ_{CFA} is fixed at a cutoff is greater than .002

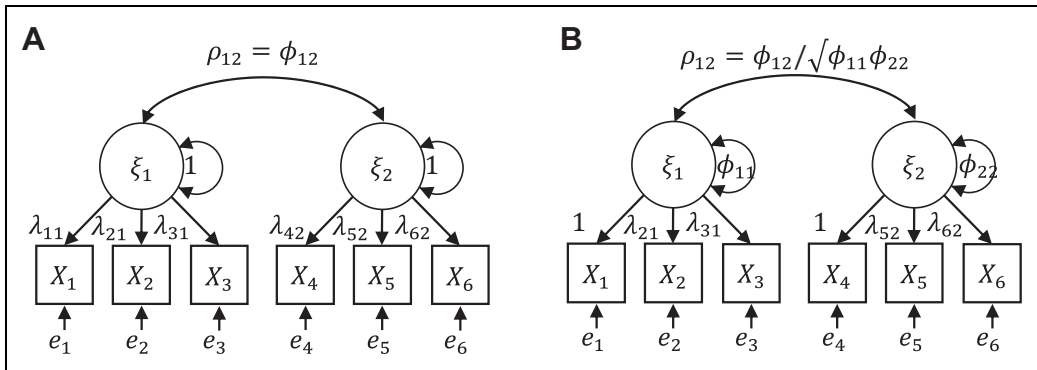


Figure 2. Factor correlation estimation. (A) Fixing the variances of factors to unity (i.e., not using the default option). (B) Fixing one of the loadings to unity (i.e., using the default option).

Overview of the Techniques for Assessing Discriminant Validity

The techniques for assessing discriminant validity identified in our review can be categorized into (a) techniques that assess correlations and (b) techniques that focus on model fit assessment. The techniques and the symbols that we use for them are summarized in Table 4.

Techniques That Assess Correlations

There are three main ways to calculate a correlation for discriminant validity assessment: a factor analysis, a scale score correlation, and the disattenuated version of the scale score correlation. While item-level correlations or their disattenuated versions could also be applied in principle, we have seen this practice neither recommended nor used. Thus, in practice, the correlation techniques always correspond to the empirical test shown as Equation 3. Regardless of how the correlations are calculated, they are used in just two ways: either by comparing the point estimates against a cutoff (i.e., the square root of the AVE) or by checking whether the absolute value of the CI contains either zero or one. Using the cutoff of zero is clearly inappropriate as requiring that two factors be uncorrelated is not implied by the definition of discriminant validity and would limit discriminant validity assessment to the extremely rare scenario where two constructs are assumed to be (linearly) independent. We will next address the various techniques in more detail.

Factor Analysis Techniques. Factor correlations can be estimated directly either by exploratory factor analysis (EFA) or CFA, but because none of the reviewed guidelines or empirical applications reported EFA correlations, we focus on CFA. The estimation of factor correlations in a CFA is complicated by the fact by default latent variables are scaled by fixing the first indicator loadings, which produces covariances that are not correlations. Correlations (denoted ρ_{CFA}) can be estimated by either freeing the factor loadings and scaling the factors by fixing their variances to 1 (i.e., A in Figure 2) or standardizing the factor covariance matrix (i.e., B), for example, by requesting standardized estimates that all SEM software provides. Both techniques produce the same estimate, although the standard errors (and CIs) can be different (Gonzalez & Griffin, 2001).

Of the correlation estimation techniques, CFA is the most flexible because it is not tied to a particular model but requires only that the model be correctly specified. Thus, cross-loadings, nonlinear factor loadings or nonnormal error terms can be included because a CFA model can also be used in the context of item response models (Foster et al., 2017), bifactor models (Rodriguez et al., 2016), exploratory SEMs (Marsh et al., 2014) or other more advanced techniques. However,

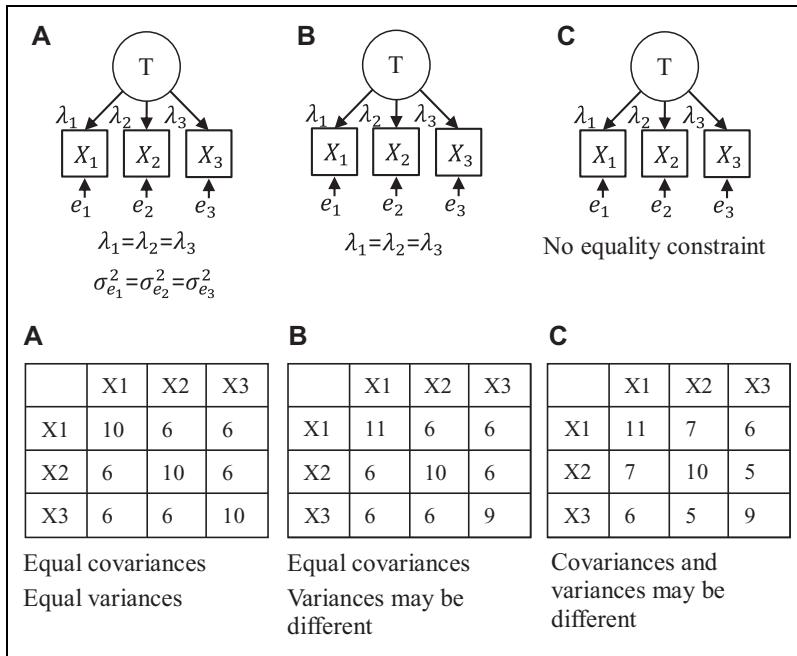


Figure 3. The assumptions of parallel, tau-equivalent, and congeneric reliability. (A) parallel, (B) tau-equivalent, (C) congeneric.

these techniques tend to require larger sample sizes and advanced software and are consequently less commonly used.

Scale Score Correlations and Disattenuated Correlations. The simplest and most common way to estimate a correlation between two scales is by summing or averaging the scale items as scale scores and then taking the correlation (denoted ρ_{SS}).⁴ The problem with this approach is that the scores contain measurement errors, which attenuate the correlation and may cause discriminant validity issues to go undetected.⁵ To address this issue, the use of *disattenuated* or error-corrected correlations where the effect of unreliability is removed is often recommended (Edwards, 2003; J. A. Shaffer et al., 2016):

$$\rho_{12} = \frac{\rho_{XY}}{\sqrt{\rho_X \rho_Y}} \tag{4}$$

where the disattenuated correlation (ρ_{12}) is a function of the scale score correlation (ρ_{XY}) and the scale score reliabilities (ρ_X , ρ_Y). That is, a disattenuated correlation is the scale score correlation from which the effect of unreliability is removed.⁶

Reliability can be estimated in different ways, including test-retest reliability, interrater reliability and single-administration reliability,⁷ which each provide information on different sources of measurement error (Le et al., 2009; Schmidt et al., 2003). Given our focus on single-method and one-time measurements, we address only single-administration reliability, where measurement errors are operationalized by uniqueness estimates, ignoring time and rater effects that are incalculable in these designs. The two most commonly used single-administration reliability coefficients are tau-equivalent reliability,⁸ often referred to as Cronbach’s alpha, and congeneric reliability, usually called composite reliability by organizational researchers and McDonald’s omega or ω by methodologists.⁹ As the names indicate, the key difference is whether we assume that the items share the

same true score (tau-equivalent, B in Figure 3) or make the less constraining assumption that the items simply depend on the same latent variable but may do so to different extents (congeneric, C in Figure 3). Recent research suggests that the congeneric reliability coefficient is a safer choice because of the less stringent assumption (Cho, 2016; Cho & Kim, 2015; McNeish, 2017).

The reliability coefficients presented above make a unidimensionality assumption, which may not be realistic in all empirical research. While the disattenuation formula (Equation 4) is often claimed to assume that the only source of measurement error is random noise or unreliability, the assumption is in fact more general: All variance components in the scale scores that are not due to the construct of interest are independent of the construct and measurement errors of other scale scores. This more general formulation seems to open the option of using hierarchical omega (Cho, 2016; Zinbarg et al., 2005), which assumes that the scale measures one main construct (main factor) but may also contain a number of minor factors that are assumed to be uncorrelated with the main factor. However, using hierarchical omega for disattenuation is problematic because it introduces an additional assumption that the minor factors (e.g., disturbances in the second-order factor model and group factors in the bifactor model) are also uncorrelated between two scales, which is neither applied nor tested when reliability estimates are calculated separately for both scales, as is typically the case. While the basic disattenuation formula has been extended to cases where its assumptions are violated in known ways (Wetecher-Hendricks, 2006; Zimmerman, 2007), the complexities of modeling the same set of violations in both the reliability estimates and the disattenuation equation do not seem appealing given that the factor correlation can be estimated more straightforwardly with a CFA instead.

While simple to use, the disattenuation correction is not without problems. A common criticism is that the correction can produce inadmissible correlations (i.e., greater than 1 or less than -1) (Charles, 2005; Nimon et al., 2012), but this issue is by no means a unique problem because the same can occur with a CFA. However, a CFA has three advantages over the disattenuation equation. First, CFA correlations estimated with maximum likelihood (ML) can be expected to be more efficient than multistep techniques that rely on corrections for attenuation (Charles, 2005; Muchinsky, 1996). Second, the disattenuation equation assumes that the scales are unidimensional and that all measurement errors are uncorrelated, whereas a CFA simply assumes that the model is correctly specified and identified. Third, calculating the CIs for a disattenuated correlation is complicated (Oberski & Satorra, 2013). However, this final concern can be alleviated to some extent through the use of bootstrap CIs (Henseler et al., 2015); in particular, the bias-corrected and accelerated (BCa) technique has been shown to work well for this particular problem (Padilla & Veprinsky, 2012, 2014).

AVE/SV or Fornell-Larcker Criterion. While commonly used, the AVE statistic has been rarely discussed by methodological research and, consequently, is poorly understood.¹⁰ One source of confusion is the similarity between the formula for AVE and that of congeneric reliability (Fornell & Larcker, 1981a):

$$AVE = \frac{\sum_{i=1}^k \lambda_i^2}{\sum_{i=1}^k \lambda_i^2 + \sum_{i=1}^k \sigma_{e_i}^2} \quad (5)$$

$$CR = \frac{(\sum_{i=1}^k \lambda_i)^2}{(\sum_{i=1}^k \lambda_i)^2 + \sum_{i=1}^k \sigma_{e_i}^2} \quad (6)$$

The meaning of AVE becomes more apparent if we rewrite the original equation as:

$$AVE = \frac{\sum_{i=1}^k \sigma_{x_i}^2 \rho_i}{\sum_{i=1}^k \sigma_{x_i}^2} \quad (7)$$

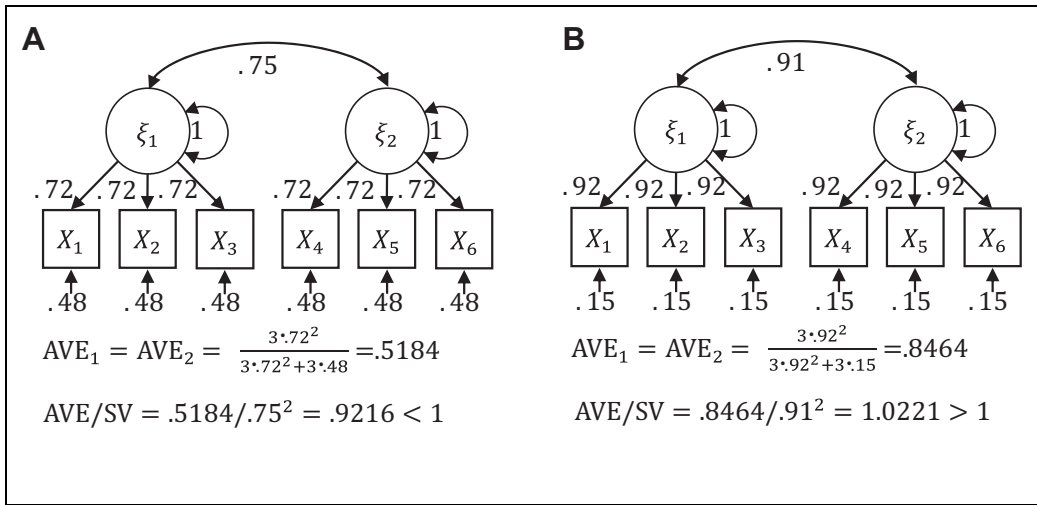


Figure 4. The inappropriateness of the AVE as an index of discriminant validity. (A) despite high discriminant validity, the AVE/SV criterion fails, (B) despite low discriminant validity, the AVE/SV criterion passes.

where $\rho_i = \frac{\lambda_i^2}{\sigma_{x_i}^2}$ is the reliability and $\sigma_{x_i}^2 = \lambda_i^2 + \sigma_{e_i}^2$ is the variance of item i . This alternative form shows that AVE is actually an item-variance *weighted* average of item reliabilities. With standardized estimates, AVE reduces to an average of item reliabilities. Thus, the term “average indicator reliability” might be more informative than “average variance extracted.”

Fornell and Larcker (1981a) presented a decision rule that discriminant validity holds for two scales if the AVEs for both are higher than the squared factor correlation between the scales. We refer to this rule as AVE/SV because the squared correlation quantifies shared variance (SV; Henseler et al., 2015). Because a factor correlation corrects for measurement error, the AVE/SV comparison is similar to comparing the left-hand side of Equation 3 against the right-hand side of Equation 2. Therefore, AVE/SV has a high false positive rate, indicating a discriminant validity problem under conditions where most researchers would not consider one to exist, as indicated by A in Figure 4. This tendency has been taken as evidence that AVE/SV is “a very conservative test” (Voorhees et al., 2016, p. 124), whereas the test is simply severely biased.

In applied research, the AVE/SV criterion rarely shows a discriminant validity problem because it is commonly misapplied. The most common misapplication is to compare the AVE values with the square of the scale score correlation, not the square of the factor correlation (Voorhees et al., 2016). Another misuse is to compare AVEs against the .5 rule-of-thumb cutoff, which Fornell and Larcker (1981a) presented as a convergent validity criterion. Other misuses are that only one of the two AVE values or the average of the two AVE values should be greater than the SV (Farrell, 2010; Henseler et al., 2015). The original criterion is that *both* AVE values must be greater than the SV. Finally, the AVE statistic is sometimes calculated from a partial least squares analysis (AVE_{PLS}), which overestimates indicator reliabilities and thus cannot detect even the most serious problems (Rönkkö & Evermann, 2013).

Heterotrait-Monotrait Ratio (HTMT). The heterotrait-monotrait (HTMT) ratio was recently introduced in marketing (Henseler et al., 2015) and is being adopted in other disciplines as well (Kuppelwieser et al., 2019). While Henseler et al. (2015) motivate HTMT based on the original MTMM approach (Campbell & Fiske, 1959), this index is actually neither new nor directly based on the MTMM

approach. The original version of the HTMT equation is fairly complex, but to make its meaning more apparent, it can be simplified as follows:

$$\text{HTMT}_{ij} = \frac{\overline{\sigma}_{ij}}{\sqrt{\overline{\sigma}_i \overline{\sigma}_j}} \quad (8)$$

where $\overline{\sigma}_i$ and $\overline{\sigma}_j$ denote the average within scale item correlation and $\overline{\sigma}_{ij}$ denotes the average between scale item correlation for two scales i and j . This simpler form makes it clear that HTMT is related to the disattenuation formula (Equation 4). In fact, HTMT is equivalent to a disattenuated correlation of unit-weighted composites using parallel reliability (i.e., the standardized alpha, proof in the appendix). Thus, while marketed as a new technique, the HTMT index has actually been used for decades; parallel reliability is the oldest reliability coefficient (Brown, 1910), and disattenuated correlations have been used to assess discriminant validity for decades (Schmitt, 1996). Thus, the term “HTMT” is misleading, giving the false impression that HTMT is related to MTMM and obscuring the fact that it is simply a variant of disattenuated correlation. Thus, we favor a more transparent term “disattenuated correlation using parallel reliability” (denoted ρ_{DPR}) because this more systematic name tells what the coefficient is (i.e., correlation), how it is obtained (i.e., disattenuated), and under what conditions it can be used (i.e., parallel reliability).

Because the ρ_{DPR} statistic is a disattenuated correlation, it shares all the interpretations, assumptions, and limitations of the techniques that were explained earlier. Compared to the tau-equivalence assumption, this technique makes an even more constraining parallel measurement assumption that the error variances between items are the same (A in Figure 3). Most real-world data deviate from these assumptions, in which case ρ_{DPR} yields inaccurate estimates (Cho, 2016; McNeish, 2017), making this technique an inferior choice.

Model Fit Comparison Techniques

Model comparison techniques involve comparing the original model against a model where a factor correlation is fixed to a value high enough to be considered a discriminant validity problem. Their general idea is that if the two models fit equally well, the model with a discriminant validity problem is plausible, and thus, there is a problem. The most common constraints are that (a) two factors are fixed to be correlated at 1 (i.e., A in Figure 5) or (b) two factors are merged into one factor (i.e., C in Figure 5), thus reducing their number by one.

The key advantage of these techniques is that they provide a test statistic and a p-value. However, this also has the disadvantage that it steers a researcher toward making yes/no decisions instead of assessing the degree to which discriminant validity holds in the data. Compared to correlation-based techniques, where a single CFA model provides all the estimates required for discriminant validity assessment, model comparison techniques require more work because a potentially large number of comparisons must be managed.¹¹ We next assess the various model comparison techniques presented and used in the literature.

$\chi^2(1)$. In the $\chi^2(1)$ test, the constrained model has the *correlation* between two factors fixed to be 1, after which the model is compared against the original one with a nested model χ^2 test. While the nested model χ^2 test is a standard tool in SEM, there are four issues that require attention when $\chi^2(1)$ is applied for discriminant validity assessment. First, it is easy to specify the constrained model incorrectly. As discussed earlier, SEM models estimate factor covariances, and implementing $\chi^2(1)$ involves constraining one of these covariances to 1.¹² However, methodological articles commonly fail to explain that the 1 constraint *must* be accompanied by setting the variances of the latent variables to 1 instead of scaling the latent variables by fixing the first item loadings (J. A. Shaffer et al., 2016; Voorhees et al., 2016). Indeed, our review provided evidence that incorrect application

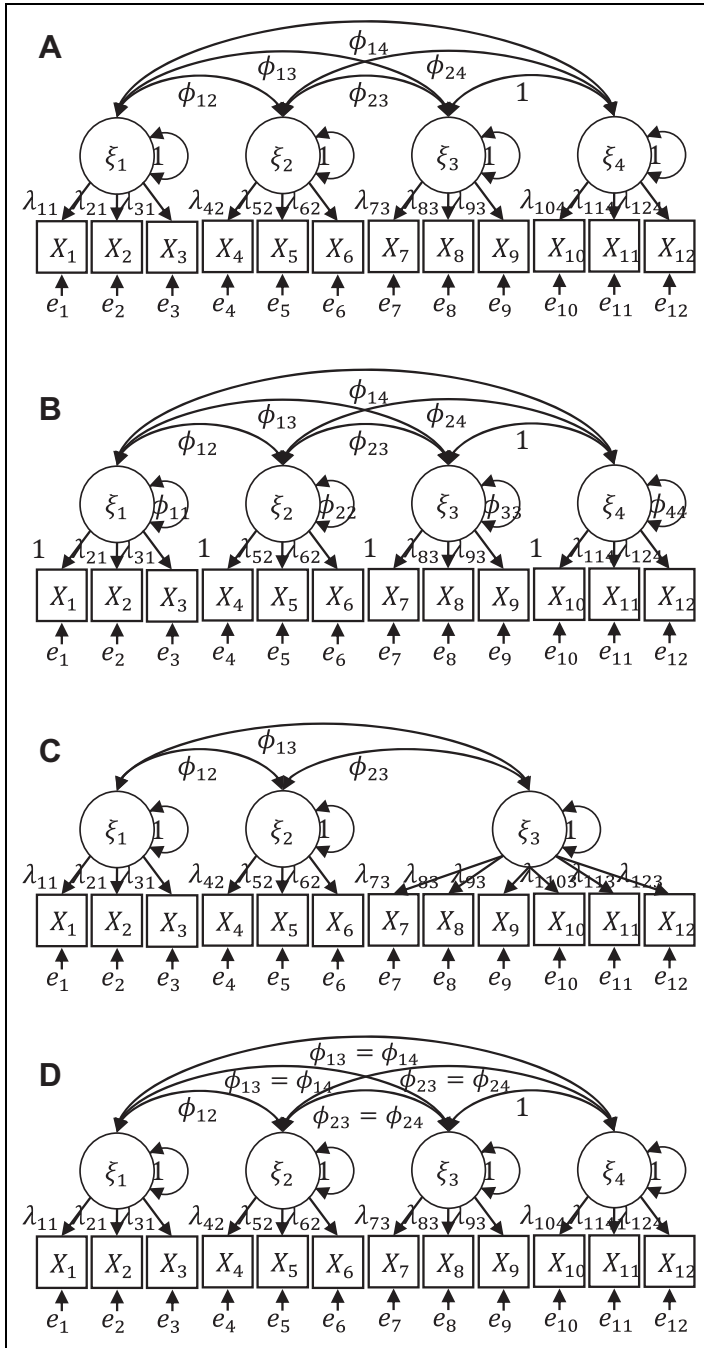


Figure 5. (A) constrained model for $\chi^2(1)$, (B) common misuse of $\chi^2(1)$, (C) constrained model for $\chi^2(\text{merge})$, (D) model equivalent to C.

of this test may be fairly common.¹³ An incorrect scaling of the latent variable (i.e., B in Figure 5) can produce either an inflated false positive or false negative rate, depending on whether the estimated factor variances are greater than 1 or less than 1. We demonstrate this problem in Online Supplement 1.

The second issue is that some articles suggest that the significance level of the χ^2 difference test should be adjusted for multiple comparisons (Anderson & Gerbing, 1988; Voorhees et al., 2016). Their logic is that if the Type I error of an individual test equals α , the probability that a Type I error occurs at least once (familywise Type I error) is much greater than α when multiple tests are done. Multiple comparison corrections address this issue by adjusting the individual test α level to keep the familywise Type I error at the intended level. To address this issue, Anderson and Gerbing (1988, n. 2) recommend applying the Šidák correction. Voorhees et al. (2016) further claim that the common omission of the correction is “the most troublesome issue with the [$\chi^2(1)$] approach” (p. 123). These concerns are ill-founded. Šidák and the related Bonferroni corrections make the universal null hypothesis *that all individual null hypotheses are true* (Hancock & Klockars, 1996; Perneger, 1998; J. P. Shaffer, 1995). Thus, in the context of $\chi^2(1)$, the universal null hypothesis is that all factors are perfectly correlated. This hypothesis is almost certainly always false, rendering tests that rely on it meaningless. Additionally, if the hypothesis is of interest, fitting a single-factor model to the data (i.e., merging all factors into one) provides a more straightforward test. While multiple comparison techniques can be useful (Castañeda et al., 1993), there are scenarios where they should not be applied (Perneger, 1998), and the literature on discriminant validity has failed to provide compelling reasons to do so.

The third issue is that the $\chi^2(1)$ technique omits constraints that the perfect correlation implies: If the correlation between two factors equals 1, their correlations with all other factors should be equal as well. However, estimated correlations are unlikely to be exactly the same, often producing an inadmissible solution with a nonpositive definite latent variable covariance matrix. While this is not a problem for the χ^2 test itself, it produces a warning in the software and may cause unnecessary confusion.¹⁴ This can be addressed by adding the implied equality constraints, but none of the reviewed works did this. Moreover, it is easier to estimate an equivalent model by simply merging the two factors as one ($\chi^2(\text{merge})$).

The fourth and final issue is that the $\chi^2(1)$ technique is a very powerful test for detecting whether the factor correlation is exactly 1. Paradoxically, this power to reject the null hypothesis has been interpreted as a lack of power to detect discriminant validity (Voorhees et al., 2016). Even if the latent variable correlation is only slightly different from 1 (e.g., .98), such small differences will be detected as statistically significant if the sample size is sufficiently large. However, in this case, it is difficult to interpret the latent variables as representing distinct concepts. This is a genuine problem with the $\chi^2(1)$ test, and two proposals for addressing it have been presented in the literature.

CFI(1). To address the issue that the $\chi^2(1)$ test can flag correlations that differ from 1 by trivial amounts as significant, some recent articles (Le et al., 2010; J. A. Shaffer et al., 2016) have suggested comparing models by calculating the difference between the comparative fit indices (CFIs) of two models (ΔCFI), which is compared against the .002 cutoff (CFI(1)). This idea is adapted from Cheung and Rensvold’s (2002) proposal in the measurement invariance literature, and the .002 cutoff is based on the simulation by Meade et al. (2008).

The idea behind using ΔCFI in measurement invariance assessment is that the degrees of freedom of the invariance hypothesis depend on the model complexity, and the CFI index and consequently ΔCFI are less affected by this than the χ^2 (Meade et al., 2008). This idea is reasonable in the original context, but it does not apply in the context of CFI(1) comparison where the difference in degrees of freedom is always one, leaving this test without its main justification. Indeed, CFI(1) can be proved (see the appendix) to be equivalent to calculating the $\Delta\chi^2$ and comparing this statistic against a

cutoff defined based on the fit of the null model (χ_B^2) and its degrees of freedom (df_B),

$$\Delta\chi^2 > 1 + .002(\chi_B^2 - df_B) \quad (9)$$

instead of using the .95 percentile from the $\chi_{(1)}^2$ distribution, or 3.84, as a cutoff. Because the expected value of χ_B^2 increases with sample size in a way similar to $\Delta\chi^2$, this comparison can be regarded as less sensitive to sample size. However, it is unclear whether this alternative cutoff has more or less power (i.e., whether $1 + .002(\chi_B^2 - df_B)$ is greater or less than 3.84) because the effectiveness of CFI(1) has not been studied.

Changing the test statistic—or equivalently the cutoff value—is an ultimately illogical solution because the problem with the $\chi^2(1)$ test is not that its power increases with sample size but that a researcher is ultimately not interested in whether the correlation between two variables differs from exactly 1; rather, a researcher is interested in whether the correlation is sufficiently different from 1. Thus, a more logical approach is to change the null hypothesis instead of adjusting the tests to be less powerful.

$\chi^2(\text{cut})$ and $CFI(\text{cut})$. The lack of a perfect correlation between two latent variables is ultimately rarely of interest, and thus, it is more logical to use a null hypothesis that covers an interval (e.g., $\phi_{12} > .9$). This test can be implemented in any SEM software by first fitting a model where ϕ_{12} is freely estimated. If the estimate falls outside the interval (e.g., less than .9), then the correlation is constrained to be at the endpoint of the interval, and the model is re-estimated. A significant result from a nested model comparison means that the original interval hypothesis can be rejected. We use the labels $\chi^2(\text{cut})$ and $CFI(\text{cut})$ to denote tests that depend on whether the comparison is performed based on the χ^2 test or by comparing ΔCFI against the .002 cutoff. While the idea that a number less than 1 can be used as a cutoff was briefly mentioned in John and Benet-Martínez (2000) and J. A. Shaffer et al. (2016), we are unaware of any studies that have applied interval hypothesis tests or tested their effectiveness.

$\chi^2(\text{merge})$ and Other Comparisons Against Models with Fewer Factors. Many studies assess discriminant validity by comparing the hypothesized model with a model with fewer factors. This is typically done by merging two factors (i.e., C in Figure 5), and we refer to the associated nested model comparison as $\chi^2(\text{merge})$. As mentioned above, this test is a more constrained version of $\chi^2(1)$ where all pairs of correlations with either of the factors and a third factor are constrained to be the same (i.e., D in Figure 5). Because this test imposes more constraints than $\chi^2(1)$ does, it has more statistical power. It also avoids the inadmissible solution issue of $\chi^2(1)$. Merging two factors will always produce the same χ^2 regardless of how the latent variables are scaled, and thus, this test is less likely to be incorrectly applied. However, the test also has a major weakness that, in contrast to $\chi^2(1)$, this test cannot be extended to other cutoffs (i.e., $\chi^2(\text{cut})$). Thus, while the test can be applied for testing perfect overlap between two latent variables, it cannot answer the question of whether the latent variables are sufficiently distinct.

As with the other techniques, various misconceptions and misuses are found among empirical studies. The most common misuse is to include unnecessary comparisons, for example, by testing alternative models with two or more factors less than the hypothesized model. Another common misuse is to omit some necessary comparisons, for example, by comparing only some alternative models, instead of comparing all possible alternative models with one factor less than the original model.

Table 5. Factor Pattern Coefficients, Correlations, and Structure Coefficients.

Items	Definitions				Numerical Example			
	Factor Pattern Coefficients		Factor Structure Coefficients		Factor Pattern Coefficients		Factor Structure Coefficients	
	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2
x ₁	λ_{11}	0	λ_{11}	$\lambda_{11}\rho$.70	0	.70	.35
x ₂	λ_{21}	0	λ_{21}	$\lambda_{21}\rho$.60	0	.60	.30
x ₃	λ_{31}	0	λ_{31}	$\lambda_{31}\rho$.80	0	.80	.40
x ₄	0	λ_{42}	$\lambda_{42}\rho$	λ_{42}	0	.90	.45	.90
x ₅	0	λ_{52}	$\lambda_{52}\rho$	λ_{52}	0	.70	.35	.70
x ₆	0	λ_{62}	$\lambda_{62}\rho$	λ_{62}	0	.80	.40	.80

Note: Factor correlation $\rho = .5$. The factor structure coefficients are the matrix multiplication of the factor pattern coefficients and factor correlations.

Single Model Fit Techniques

Discriminant validity has also been assessed by inspecting the fit of a single model without comparing against another model. These techniques fall into two classes: those that inspect the factor loadings and those that assess the overall model fit.

Cross-Loadings. Cross-loadings indicate a relationship between an indicator and a factor other than the main factor on which the indicator loads. Beyond this definition, the term can refer to two distinct concepts, factor pattern coefficients or factor structure coefficients (see Table 5), and has been confusingly used with both meanings in the discriminant validity literature.¹⁵ Structure coefficients are correlations between items and factors, so their values are constrained to be between -1 and 1 . Pattern coefficients, on the other hand, are analogous to (standardized) coefficients in regression analysis and are directional (Thompson & Daniel, 1996). Structure coefficients are not directly estimated as part of a factor analysis; instead, they are calculated based on the pattern coefficients and factor correlations. If the factors are rotated orthogonally (e.g., Varimax) or are otherwise constrained to be uncorrelated, the pattern coefficients and structure coefficients are identical (Henson & Roberts, 2006). However, the use of uncorrelated factors can rarely be justified (Fabrigar et al., 1999), which means that, in most cases, pattern and structure coefficients are not equal.

While both the empirical criteria shown in Equation 2 and Equation 3 contain pattern coefficients, assessing discriminant validity based on loadings is problematic. First, these comparisons involve assessing a single item or scale at a time, which is incompatible with the idea that discriminant validity is a feature of a measure pair. Second, pattern coefficients do not provide any information on the correlation between two scales, and structure coefficients are an indirect measure of the correlation at best.¹⁶ Third, while the various guidelines differ in how loadings should be interpreted (Henseler et al., 2015; Straub et al., 2004; Thompson, 1997), they all share the features of relying mostly on authors' intuition instead of theoretical reasoning or empirical evidence. In summary, these techniques fall into the rules of thumb category and cannot be recommended.

Single-Model fit. The final set of techniques is those that assess the single-model fit of a CFA model. A CFA model can fit the data poorly if there are unmodeled loadings (pattern coefficients), omitted factors, or error correlations in the data, none of which are directly related to discriminant validity.

While some of the model fit indices do depend on factor correlations, they do so only weakly and indirectly (Kline, 2011, chap. 8). Thus, while a well-fitting factor model is an important assumption (either implicitly—e.g., the various ρ_D —or explicitly—e.g., CFA techniques), model fit itself will not provide any information on whether either of the two empirical criteria shown in Equation 2 and Equation 3 holds.

Summary: Under What Conditions Is Each Technique Useful?

Our review of the literature provides several conclusions. Overall, $\chi^2(\text{cut})$ and $CI_{\text{CFA}}(\text{cut})$ can be recommended as general solutions because they meet the definition of discriminant validity, have the flexibility to adapt to various levels of cutoffs, and can be extended to more complex scenarios such as nonlinear measurement models (Foster et al., 2017), scales with minor dimensions (Rodriguez et al., 2016), or cases in which factorial validity is violated because of cross-loadings. Some of the other techniques can be useful for specific purposes. $\chi^2(\text{merge})$, $\chi^2(1)$, and $CI_{\text{CFA}}(1)$ can be used if theory suggests nearly perfect but not absolutely perfect correlations. They can also be useful as a first step in discriminant validity assessment; if any of them indicates a problem, then so will any variant of the techniques that use a cutoff of less than 1. Disattenuated correlations are useful in single-item scenarios, where reliability estimates could come from test-retest or interrater reliability checks or from prior studies. These techniques could also be used in multiple item scenarios, if a researcher does not have access to SEM software, or in some small sample scenarios (Rosseel, 2020). If the effect of measurement error can be assumed to be negligible, even using scale score correlations can be useful as a rough check. The techniques that assess the lack of cross-loadings (pattern coefficients) and model fit provide (factorial) validity information, which is important in establishing the assumptions of the other techniques, but these techniques are of limited use in providing actual discriminant validity evidence.

Monte Carlo Simulations

We will next compare the various discriminant validity assessment techniques in a Monte Carlo simulation with regard to their effectiveness in two common tasks: (a) quantifying the degree to which discriminant validity can be a problem and (b) making a dichotomous decision on whether discriminant validity is a problem in the population. For statistics with meaningful interpretations, we assessed the bias and variance of the statistic and the validity of the CIs. For the dichotomous decision, we set .85, .9, .95, and 1 as cutoffs and estimated the Type I error (i.e., false positive) and Type II error (i.e., false negative) rates for the conclusion that the factor correlation was at least at the cutoff value in the population.

Simulation Design

For simplicity, we followed the design used by Voorhees et al. (2016) and generated data from a three-factor model. We assessed the discriminant validity of the first two factors, varying their correlation as an experimental condition. Voorhees et al. (2016) used only two factor correlation levels, .75 and .9. We wanted a broader range from low levels where discriminant validity is unlikely to be a problem up to perfect correlation, so we used six levels: .5, .6, .7, .8, .9, and 1. The third factor was always correlated at .5 with the first two factors. The number of items was varied at 3, 6, and 9, and factor loadings were set to [0.9,0.8,0.8], [0.8,0.7,0.7], [0.5,0.4,0.4], [0.9,0.6,0.3], and [0.8, 0.8, 0.8]. For the six- and nine-item scenarios, each factor loading value was used multiple times. All factors had unit variances in the population, and we scaled the error variances so that the population variances of the items were one. Cross-loadings (in the pattern coefficients) were either 0, 1, or 2.

This condition was implemented following the approach by Voorhees et al. (2016); in the cross-loading condition, either the first or first two indicators of the second latent variable also loaded on the first latent variable. The same value was used for both loadings, and the values were scaled down from their original values so that the factors always explained the same amount of variance in the indicators. In the six- and nine-item conditions, the number of cross-loaded items was scaled up accordingly. Sample size was the final design factor and varied at 50, 100, 250, and 1,000. The full factorial ($6 \times 3 \times 5 \times 3 \times 4$) simulation was implemented with the R statistical programming environment using 1,000 replications for each cell. We estimated the factor models with the lavaan package (Rosseel, 2012) and used semTools to calculate the reliability indices (Jorgensen et al., 2020). The full simulation code is available in Online Supplement 2, and the full set of simulation results at the design level can be found in Online Supplement 3.

Correlation Estimates and Their Confidence Intervals

We first focus on the scenarios where the factor model was correctly specified (i.e., there were no cross-loadings). Because ρ_{DPR} and HTMT were proven equivalent and always produced identical results, we report only the former. We also omit the two low correlation conditions (i.e., .5, .6) because the false positive rates are already clear in the .7 condition.

Because the pattern of results is very similar for correlations and their confidence intervals, we present both sets of results together. Table 6 shows the correlation estimates by sample size, number of items, and factor loading conditions. Table 7 shows the coverage and the balance of the CIs by sample size and selected values of loading condition, omitting ρ_{SS} because of its generally poor performance in the correlation results. Ideally, the coverage of a 95% CI should be .95, and the balance should be close to zero. The CIs for ρ_{CFA} were obtained from the CFAs, and for ρ_{DPR} , we used bootstrap percentile CIs, following Henseler et al. (2015). For comparison, we also calculated the bootstrap percentile CIs for ρ_{DTR} and ρ_{DCR} . All bootstrap analyses were calculated with 1,000 replications.

The first set of rows in Table 6 shows the effects of sample size. Scale score correlation ρ_{SS} was always negatively biased due to the well-known attenuation effect. All disattenuation techniques and CFA performed better, and in large samples (250, 1,000), their performance was indistinguishable. This result was expected because all these approaches are consistent and their assumptions hold in this set of conditions. In the smallest sample size (50), CFA was slightly biased to be less efficient than the disattenuation-based techniques, but the differences were in the third digit and thus were inconsequential. The number of indicators, shown in the second set of rows in Table 6, affects the bias of the scale score correlation ρ_{SS} because increasing the number of indicators increases reliability and, consequently, reduces the attenuation effect. The effect for other techniques was an increase in precision, which was expected because more indicators provide more information from which to estimate the correlation.

The third set of rows in Table 6 demonstrates the effects of varying the factor loadings. When the factor loadings were equal (all at .8), the performance of CFA and all disattenuation techniques was identical, which was expected, as explained above. Table 7 shows that in this condition, the confidence intervals of all techniques performed reasonably well. The balance statistics were all negative, indicating that when the population value was outside the CI, it was generally more frequently below the lower limit of the interval than above the upper limit. In other words, all CIs were slightly positively biased. This effect and the general undercoverage of the CIs were most pronounced in small samples. CI_{CFA} performed slightly worse than disattenuated correlations when the sample size was very small and the correlation between the factors was weaker. However, outside the smallest sample sizes, the differences were negligible in the third decimals.

Table 6. Correlation Estimates by Sample Size, Number of Items, and Loadings.

Sample Size, Indicators, Loadings ^a	Technique	Factor Correlation							
		.7		.8		.9		1.0	
		M	SD	M	SD	M	SD	M	SD
50	PCFA	.694	.111	.801	.088	.897	.070	1.000	.049
	PDPR	.697	.108	.802	.086	.901	.066	1.003	.047
	PDTR	.698	.108	.804	.086	.903	.066	1.005	.048
	PDCR	.694	.108	.800	.087	.898	.066	1.000	.047
	PSS	.583	.095	.672	.080	.753	.065	.838	.045
100	PCFA	.700	.076	.801	.061	.899	.045	1.000	.032
	PDPR	.700	.075	.802	.059	.900	.044	1.002	.031
	PDTR	.701	.075	.803	.060	.901	.044	1.003	.031
	PDCR	.699	.076	.801	.060	.899	.044	1.000	.031
	PSS	.588	.068	.673	.057	.755	.042	.841	.030
250	PCFA	.699	.043	.799	.038	.899	.028	1.001	.019
	PDPR	.699	.043	.799	.037	.900	.027	1.001	.019
	PDTR	.700	.043	.800	.037	.900	.027	1.002	.019
	PDCR	.699	.043	.799	.037	.899	.027	1.001	.019
	PSS	.588	.039	.672	.035	.757	.027	.842	.018
1,000	PCFA	.699	.023	.800	.018	.899	.014	1.000	.009
	PDPR	.699	.023	.800	.018	.899	.014	1.000	.009
	PDTR	.699	.023	.800	.018	.899	.014	1.000	.009
	PDCR	.699	.023	.800	.018	.899	.014	1.000	.009
	PSS	.589	.021	.674	.017	.756	.013	.842	.009
3	PCFA	.699	.043	.799	.038	.899	.028	1.001	.019
	PDPR	.699	.043	.799	.037	.900	.027	1.001	.019
	PDTR	.700	.043	.800	.037	.900	.027	1.002	.019
	PDCR	.699	.043	.799	.037	.899	.027	1.001	.019
	PSS	.588	.039	.672	.035	.757	.027	.842	.018
6	PCFA	.699	.039	.799	.030	.900	.019	1.000	.009
	PDPR	.699	.039	.799	.030	.900	.019	1.000	.009
	PDTR	.699	.039	.800	.030	.900	.019	1.000	.009
	PDCR	.699	.039	.799	.030	.900	.019	1.000	.009
	PSS	.639	.038	.731	.030	.823	.021	.914	.010
9	PCFA	.699	.038	.798	.028	.899	.016	1.000	.006
	PDPR	.699	.038	.798	.028	.899	.016	1.000	.006
	PDTR	.699	.038	.798	.028	.899	.016	1.001	.006
	PDCR	.699	.038	.798	.028	.898	.016	1.000	.006
	PSS	.657	.037	.751	.029	.845	.018	.941	.007
0.5, 0.4, 0.4	PCFA	.709	.156	.802	.161	.915	.157	1.012	.170
	PDPR	.727	.153	.819	.158	.934	.157	1.028	.168
	PDTR	.728	.153	.819	.158	.934	.157	1.029	.168
	PDCR	.711	.152	.802	.156	.917	.155	1.012	.169
	PSS	.288	.058	.325	.058	.369	.053	.407	.053
0.8, 0.7, 0.7	PCFA	.699	.051	.803	.042	.900	.035	1.001	.028
	PDPR	.701	.052	.805	.043	.903	.036	1.004	.029
	PDTR	.702	.052	.805	.043	.904	.036	1.004	.029
	PDCR	.699	.051	.802	.043	.900	.036	1.001	.028
	PSS	.544	.043	.624	.037	.699	.033	.778	.026

(continued)

Table 6. (continued)

Sample Size, Indicators, Loadings ^a	Technique	Factor Correlation							
		.7		.8		.9		1.0	
		M	SD	M	SD	M	SD	M	SD
0.8, 0.8, 0.8	ρ_{CFA}	.699	.043	.799	.038	.899	.028	1.001	.019
	ρ_{DPR}	.699	.043	.799	.037	.900	.027	1.001	.019
	ρ_{DTR}	.700	.043	.800	.037	.900	.027	1.002	.019
	ρ_{DCR}	.699	.043	.799	.037	.899	.027	1.001	.019
	ρ_{SS}	.588	.039	.672	.035	.757	.027	.842	.018
0.9, 0.6, 0.3	ρ_{CFA}	.697	.062	.800	.052	.901	.047	1.000	.036
	ρ_{DPR}	.767	.083	.879	.073	.989	.071	1.101	.065
	ρ_{DTR}	.768	.082	.879	.073	.990	.071	1.102	.066
	ρ_{DCR}	.698	.072	.799	.062	.901	.058	1.003	.049
	ρ_{SS}	.453	.049	.519	.044	.586	.040	.651	.034
0.9, 0.8, 0.8	ρ_{CFA}	.700	.043	.797	.032	.902	.023	1.000	.014
	ρ_{DPR}	.701	.043	.798	.033	.904	.025	1.002	.015
	ρ_{DTR}	.701	.043	.798	.033	.904	.024	1.002	.015
	ρ_{DCR}	.700	.043	.796	.033	.902	.024	1.000	.014
	ρ_{SS}	.610	.040	.695	.032	.787	.024	.872	.015

Note: M = mean; SD = standard deviation; Techniques: CFA = confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability; SS = summed scale without disattenuation.

a. When the sample size was varied, the number of indicators was 3, and all loadings were .8. When the number of indicators was varied, the sample size was 250, and all loadings were .8. When the loadings were varied, the sample size was 250, and the number of indicators was 3.

When the loadings varied, ρ_{DTR} and ρ_{DPR} became positively biased. The reason for this result is that these estimation techniques assume tau-equivalence or equal reliability of the indicators; when this assumption does not hold, the techniques have been shown to produce negatively biased reliability estimates, thus leading to overcorrection for attenuation and producing positive bias. The same results are mirrored in the second set of rows in Table 7; both CI_{DPR} and CI_{DTR} produced positively biased CIs with poor coverage and balance.

In summary, Table 6 and Table 7 show that the best-performing correlation estimate was ρ_{CFA} , followed by ρ_{DCR} and that the corresponding confidence intervals CI_{CFA} and CI_{DCR} outperform the others. When the condition of being tau-equivalent was violated (e.g., the loadings were .9, .6, and .3), ρ_{CFA} and ρ_{DCR} produced estimates that were more accurate than those produced by other methods, but ρ_{CFA} was slightly more precise, having smaller standard deviations. Similarly, CI_{CFA} and CI_{DCR} were largely unaffected and retained their performance from the tau-equivalent condition.

Inference Against a Cutoff

Our main results concern inference against a cutoff and are relevant when a researcher wants to make a yes/no decision about discriminant validity. We start by assessing the performance of the techniques that can be thought of as tests of a perfect correlation or as rules of thumb. Because the differences between ρ_{DPR} (i.e., HTMT) and ρ_{DTR} were negligible, only the former is reported. Similarly, CI_{DTR} is omitted due to nearly identical performance with CI_{DPR} .

Table 7. Coverage and Balance of 95% Confidence Intervals by Loadings and Sample Size.

Loadings	Sample Size	Technique	Factor Correlation							
			.7		.8		.9		1.0	
			Coverage	Balance	Coverage	Balance	Coverage	Balance	Coverage	Balance
0.8, 0.8, 0.8	50	ρ_{CFA}	.905	-.057	.899	-.085	.912	-.068	.955	-.023
		ρ_{DPR}	.921	-.037	.922	-.048	.908	-.056	.927	-.045
		ρ_{DTR}	.922	-.040	.915	-.059	.908	-.068	.909	-.067
		ρ_{DCR}	.927	-.027	.919	-.049	.916	-.046	.933	-.033
	100	ρ_{CFA}	.915	-.051	.925	-.055	.936	-.046	.955	-.025
		ρ_{DPR}	.917	-.021	.938	-.022	.944	-.030	.940	-.030
		ρ_{DTR}	.922	-.030	.932	-.028	.940	-.042	.938	-.038
		ρ_{DCR}	.925	-.019	.932	-.016	.943	-.027	.941	-.021
	250	ρ_{CFA}	.966	-.016	.938	-.022	.947	-.025	.949	-.015
		ρ_{DPR}	.967	-.011	.947	-.011	.949	-.015	.946	-.024
		ρ_{DTR}	.968	-.008	.949	-.013	.944	-.022	.942	-.026
		ρ_{DCR}	.967	-.005	.948	-.008	.948	-.012	.945	-.017
	1,000	ρ_{CFA}	.949	-.001	.955	-.017	.954	-.008	.950	-.002
		ρ_{DPR}	.938	.004	.956	-.012	.947	-.007	.947	-.007
		ρ_{DTR}	.947	.003	.956	-.010	.952	-.004	.953	-.009
		ρ_{DCR}	.946	.006	.956	-.008	.951	-.003	.951	-.005
0.9, 0.6, 0.3	50	ρ_{CFA}	.911	-.034	.919	-.030	.923	-.036	.954	.015
		ρ_{DPR}	.851	-.143	.859	-.141	.822	-.176	.806	-.194
		ρ_{DTR}	.852	-.142	.848	-.150	.810	-.188	.803	-.197
		ρ_{DCR}	.964	-.017	.957	-.023	.942	-.035	.952	-.023
	100	ρ_{CFA}	.942	-.016	.939	-.005	.949	-.019	.944	.002
		ρ_{DPR}	.859	-.135	.843	-.157	.810	-.190	.756	-.244
		ρ_{DTR}	.866	-.128	.842	-.158	.803	-.195	.747	-.253
		ρ_{DCR}	.956	-.010	.943	-.007	.945	-.017	.940	-.020
	250	ρ_{CFA}	.940	-.006	.943	-.017	.937	-.031	.964	-.002
		ρ_{DPR}	.832	-.164	.786	-.214	.694	-.306	.581	-.417
		ρ_{DTR}	.826	-.170	.781	-.219	.682	-.318	.568	-.430
		ρ_{DCR}	.943	-.007	.949	-.005	.937	-.023	.942	-.012
	1,000	ρ_{CFA}	.958	.006	.947	-.023	.950	-.008	.947	-.027
		ρ_{DPR}	.621	-.379	.453	-.545	.279	-.721	.117	-.883
		ρ_{DTR}	.613	-.387	.446	-.552	.264	-.736	.108	-.892
		ρ_{DCR}	.955	-.005	.947	-.003	.951	-.017	.945	-.017

Note: Balance is the difference between frequencies for the population value above and below the confidence interval. The number of indicators was always 3. Techniques: CFA= confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability.

Table 8 clearly shows that some of the techniques have either unacceptably low power or a false positive rate that is too high to be considered useful. Techniques that directly compare the point estimates of correlations with a cutoff (i.e., $\rho_{CFA}(1)$, $\rho_{DPR}(1)$, and $\rho_{CR}(1)$) have very high false negative rates because an unbiased and normal correlation estimate can be expected to be below the population value (here, 1) exactly half the time. In contrast, the AVE/SV technique that uses factor correlation following Fornell and Larcker's (1981a) original proposal has a very high false positive rate. As expected based on our analysis, the misused version using scale score correlation, AVE/SV_{SS} has a smaller false positive rate because the attenuation bias in the scale score correlation worked to offset the high false positive rate of the AVE comparison. Clearly, none of these techniques can be recommended.

Table 8. Detection Rates of the Discriminant Validity Problem as a Perfect Correlation by Technique.

Sample Size	Technique	Loadings and Factor Correlation								
		0.8, 0.8, 0.8				0.9, 0.6, 0.3				
		.7*	.8*	.9*	1.0 [†]	.7*	.8*	.9*	1.0 [†]	
50	$\rho_{CFA}(1)$.000	.001	.016	.513	.007	.016	.063	.503	
	$\rho_{DPR}(1)$.000	.001	.016	.530	.051	.110	.258	.770	
	$\rho_{DCR}(1)$.000	.000	.016	.511	.014	.031	.107	.529	
	$CI_{CFA}(1)$.013	.075	.293	.992	.108	.194	.388	.984	
	$CI_{DPR}(1)$.016	.072	.273	.986	.410	.616	.898	1.000	
	$CI_{DCR}(1)$.014	.066	.256	.983	.244	.397	.697	.988	
	AVE/SV _{CFA}	.161	.582	.968	1.000	.733	.941	.992	1.000	
	AVE/SV _{SS}	.014	.141	.692	.975	.178	.457	.729	.871	
	CFI(1)	.008	.037	.160	.935	.085	.140	.249	.916	
	$\chi^2(1)$.006	.043	.199	.967	.110	.185	.332	.968	
	$\chi^2(\text{merge})$.009	.056	.243	.967	.130	.210	.358	.968	
	100	$\rho_{CFA}(1)$.000	.000	.002	.520	.000	.001	.028	.510
		$\rho_{DPR}(1)$.000	.000	.002	.528	.016	.059	.204	.827
$\rho_{DCR}(1)$.000	.000	.002	.517	.001	.006	.051	.525	
$CI_{CFA}(1)$.000	.004	.106	.990	.022	.088	.241	.985	
$CI_{DPR}(1)$.000	.003	.105	.984	.241	.386	.773	.999	
$CI_{DCR}(1)$.000	.003	.093	.980	.076	.182	.483	.984	
AVE/SV _{CFA}		.063	.603	.993	1.000	.822	.980	.999	1.000	
AVE/SV _{SS}		.000	.071	.690	.990	.097	.438	.728	.878	
CFI(1)		.000	.006	.096	.973	.028	.083	.211	.955	
$\chi^2(1)$.000	.002	.081	.973	.024	.086	.234	.972	
$\chi^2(\text{merge})$.000	.005	.105	.970	.038	.114	.263	.977	
250		$\rho_{CFA}(1)$.000	.000	.000	.519	.000	.000	.006	.505
		$\rho_{DPR}(1)$.000	.000	.000	.527	.001	.016	.148	.920
	$\rho_{DCR}(1)$.000	.000	.000	.515	.000	.001	.017	.509	
	$CI_{CFA}(1)$.000	.000	.005	.983	.000	.005	.124	.984	
	$CI_{DPR}(1)$.000	.000	.005	.984	.074	.225	.516	.999	
	$CI_{DCR}(1)$.000	.000	.005	.982	.002	.030	.204	.980	
	AVE/SV _{CFA}	.009	.584	1.000	1.000	.911	.999	1.000	1.000	
	AVE/SV _{SS}	.000	.010	.671	.999	.028	.426	.697	.886	
	CFI(1)	.000	.000	.019	.994	.001	.020	.133	.987	
	$\chi^2(1)$.000	.000	.004	.974	.000	.006	.123	.976	
	$\chi^2(\text{merge})$.000	.000	.008	.972	.000	.014	.152	.976	
	1,000	$\rho_{CFA}(1)$.000	.000	.000	.491	.000	.000	.000	.510
		$\rho_{DPR}(1)$.000	.000	.000	.495	.000	.001	.106	.997
$\rho_{DCR}(1)$.000	.000	.000	.488	.000	.000	.000	.518	
$CI_{CFA}(1)$.000	.000	.000	.983	.000	.000	.001	.985	
$CI_{DPR}(1)$.000	.000	.000	.981	.000	.028	.320	1.000	
$CI_{DCR}(1)$.000	.000	.000	.980	.000	.000	.015	.979	
AVE/SV _{CFA}		.000	.583	1.000	1.000	.988	1.000	1.000	1.000	
AVE/SV _{SS}		.000	.000	.666	1.000	.001	.396	.667	.900	
CFI(1)		.000	.000	.000	1.000	.000	.000	.017	1.000	
$\chi^2(1)$.000	.000	.000	.979	.000	.000	.001	.981	
$\chi^2(\text{merge})$.000	.000	.000	.978	.000	.000	.003	.979	

Note: Averages over all indicator numbers. Techniques: ρ = comparing the correlations against a cutoff; CI = comparing if a cutoff is included in the 95% confidence interval; AVE/SV = comparing the AVE statistics against the squared correlation; CFA = confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability; CFI = nested model comparison using the CFI rule; χ^2 = nested model test; (1) = two factors are constrained to be perfectly correlated; (merge) = two factors are merged into one.

*False positive rate. [†]true positive rate = (1 - false negative rate).

The various model comparisons and CIs performed better. Among the three methods of model comparison (CFI(1), $\chi^2(1)$, and $\chi^2(\text{merge})$), $\chi^2(1)$ was generally the best in terms of both the false positive rate and false negative rate. While the difference was small, it is surprising that $\chi^2(1)$ was strictly superior to $\chi^2(\text{merge})$, having both more power and a smaller false positive rate. Our explanation for this finding is that although $\chi^2(\text{merge})$ imposes more constraints on the model, these constraints work differently when the factors are perfectly correlated and when they are not. When the factors are perfectly correlated, imposing more constraints means that the model can be declared to misfit in more ways, thus leading to lower power. In contrast, when the correlation between the factors is less than 1, the additional constraints are somewhat redundant because constraining the focal correlation to 1 will also bias all other correlations involving the focal variables. Thus, the amount of misfit produced by the first constraint is greater than the other constraints that $\chi^2(\text{merge})$ contributes. This phenomenon leads to a higher false positive rate because while the additional constraints contribute degrees of freedom, they contribute less misfit.

Another interesting finding is that although CFI(1) was proposed as an alternative to $\chi^2(1)$ based on the assumption that it had a smaller false positive rate, this assumption does not appear to be true: the false positive rates of these techniques were comparable, and in larger samples (250, 1,000), the false positive rate of CFI(1) even exceeded that of $\chi^2(1)$. Generalizing this finding to larger models requires caution because the CFI comparison depends on the fit of the null model, which depends on model size. Nevertheless, it is clear that the CFI comparison does not *generally* have a smaller false positive rate than $\chi^2(1)$.

The performance of the CIs ($CI_{CFA}(1)$, $CI_{DPR}(1)$, and $CI_{DCR}(1)$) was nearly identical in the tau-equivalent condition (i.e., all loadings at .8), but in the congeneric condition (i.e., the loadings at .3, .6, and .9), $CI_{DPR}(1)$ had an excessive false positive rate due to the positive bias explained earlier. $CI_{DCR}(1)$ had a larger false positive rate than $CI_{CFA}(1)$, particularly in small samples, possibly due to violating the large sample assumption of bootstrapping. In summary, Table 8 supports the use of $CI_{CFA}(1)$ and $\chi^2(1)$. The former had slightly more power but a larger false positive rate than the latter. The performance of these two techniques converged in large samples.

Table 9 considers cutoffs other than 1, using values of .85, .90, and .95 that are sometimes recommended in the literature, showing results that are consistent with those of the previous tables. The techniques that compared the CI against a cutoff (i.e., $CI_{CFA}(\text{cut})$, $CI_{DPR}(\text{cut})$ and $CI_{DCR}(1)$) had more power than those that compared an estimate against a cutoff (i.e., $\rho_{CFA}(\text{cut})$, $\rho_{DPR}(\text{cut})$, and $\rho_{DCR}(\text{cut})$), especially in small samples, and between the three, $CI_{CFA}(\text{cut})$ showed a lower false positive rate while having power similar to that of $CI_{DPR}(\text{cut})$ and to a smaller extent $CI_{DCR}(\text{cut})$. Among the techniques that compared model fit (i.e., CFI(cut) and $\chi^2(\text{cut})$), CFI(cut) had slightly more power but also a higher false positive rate in small samples than $\chi^2(\text{cut})$. In larger samples, the power of the two techniques was similar, but $\chi^2(\text{cut})$ generally had the lowest false positive rate. This finding and the sensitivity of the CFI tests to model size, explained earlier, make $\chi^2(\text{cut})$ the preferred alternative of the two. In summary, $CI_{CFA}(\text{cut})$ and $\chi^2(\text{cut})$ are generally the best techniques. They have different strengths: $CI_{CFA}(\text{cut})$ has slightly more power, but $\chi^2(\text{cut})$ enjoys a considerably lower false positive rate.

Effects of Model Misspecification

We now turn to the cross-loading conditions to assess the robustness of the techniques when the assumption of no cross-loadings is violated. The results shown in Table 10 show that all estimates become biased toward 1. ρ_{DCR} was slightly most robust to these misspecifications, but the differences between the techniques were not large. Table 11 presented the detection rates of different techniques using alternative cutoffs and over the cross-loading conditions and showed similar results.

Table 9. Detection Rates by Technique Using Alternative Cutoffs.

Sample Size	Technique	Cutoff and Factor Correlation											
		.85				.9				.95			
		.7*	.8*	.9 [†]	1.0 [†]	.7*	.8*	.9 [†]	1.0 [†]	.7*	.8*	.9*	1.0 [†]
50	$\rho_{CFA}(\text{cut})$.085	.289	.784	.950	.049	.126	.517	.925	.033	.064	.193	.866
	$\rho_{DPR}(\text{cut})$.121	.348	.818	.968	.075	.182	.576	.946	.051	.101	.270	.892
	$\rho_{DCR}(\text{cut})$.095	.301	.785	.958	.056	.140	.524	.930	.036	.073	.212	.857
	$CI_{CFA}(\text{cut})$.694	.945	.992	.996	.413	.788	.982	.995	.216	.439	.895	.992
	$CI_{DPR}(\text{cut})$.695	.940	.999	1.000	.477	.777	.990	1.000	.331	.522	.884	.999
	$CI_{DCR}(\text{cut})$.669	.929	.999	1.000	.441	.746	.988	1.000	.305	.476	.858	.999
	CFI(cut)	.416	.781	.971	.992	.227	.509	.915	.987	.146	.259	.649	.977
	$\chi^2(\text{cut})$.541	.884	.992	.998	.301	.629	.967	.997	.192	.333	.760	.995
100	$\rho_{CFA}(\text{cut})$.041	.194	.850	.973	.023	.068	.510	.953	.015	.034	.126	.907
	$\rho_{DPR}(\text{cut})$.066	.264	.878	.980	.037	.110	.582	.963	.022	.061	.206	.925
	$\rho_{DCR}(\text{cut})$.044	.211	.846	.975	.022	.072	.518	.952	.014	.036	.141	.897
	$CI_{CFA}(\text{cut})$.396	.891	.996	.999	.189	.525	.987	.998	.115	.230	.734	.996
	$CI_{DPR}(\text{cut})$.455	.888	.999	1.000	.279	.577	.989	1.000	.198	.337	.760	1.000
	$CI_{DCR}(\text{cut})$.410	.869	.999	1.000	.236	.528	.986	1.000	.166	.280	.721	.999
	CFI(cut)	.317	.793	.987	.998	.156	.417	.952	.996	.096	.188	.617	.990
	$\chi^2(\text{cut})$.314	.818	.997	1.000	.169	.419	.974	.999	.114	.209	.614	.997
250	$\rho_{CFA}(\text{cut})$.016	.102	.927	.989	.008	.029	.513	.977	.004	.014	.071	.945
	$\rho_{DPR}(\text{cut})$.027	.169	.942	.991	.013	.060	.589	.980	.006	.025	.143	.956
	$\rho_{DCR}(\text{cut})$.015	.110	.919	.989	.008	.030	.508	.977	.004	.014	.077	.939
	$CI_{CFA}(\text{cut})$.145	.704	.998	1.000	.079	.233	.985	1.000	.049	.114	.426	.998
	$CI_{DPR}(\text{cut})$.224	.743	.999	1.000	.139	.336	.987	1.000	.095	.192	.540	.999
	$CI_{DCR}(\text{cut})$.168	.698	.999	1.000	.097	.262	.981	1.000	.064	.137	.468	.999
	CFI(cut)	.208	.819	.996	.999	.093	.321	.981	.998	.063	.117	.571	.996
	$\chi^2(\text{cut})$.134	.635	.998	1.000	.078	.210	.975	1.000	.053	.111	.381	.999
1,000	$\rho_{CFA}(\text{cut})$.002	.030	.976	.999	.000	.006	.505	.995	.000	.002	.025	.980
	$\rho_{DPR}(\text{cut})$.003	.095	.980	.999	.000	.025	.622	.997	.000	.003	.093	.983
	$\rho_{DCR}(\text{cut})$.002	.033	.975	.999	.000	.006	.507	.995	.000	.002	.026	.979
	$CI_{CFA}(\text{cut})$.032	.253	1.000	1.000	.016	.071	.980	1.000	.006	.033	.174	1.000
	$CI_{DPR}(\text{cut})$.069	.378	1.000	1.000	.028	.146	.984	1.000	.011	.074	.309	1.000
	$CI_{DCR}(\text{cut})$.036	.279	1.000	1.000	.020	.081	.979	1.000	.008	.038	.198	1.000
	CFI(cut)	.063	.846	.999	1.000	.027	.132	.995	1.000	.013	.060	.400	1.000
	$\chi^2(\text{cut})$.033	.238	1.000	1.000	.018	.071	.975	1.000	.007	.035	.170	1.000

Note: Averages over all indicator numbers and loading. Techniques: ρ = comparing correlation against a cutoff; CI = comparing if a cutoff is included in the 95% confidence interval, CFA= confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability; CFI = nested model comparison using the CFI rule; χ^2 = nested model test.
 *False positive rate. [†]true positive rate = (1 – false negative rate).

All techniques were again affected, and both the power and false positive rates increased across the board when the correlation between the factors was less than one. As before, the effect was stronger for smaller population correlations.

In the cross-loading conditions, we also estimated a correctly specified CFA model in which the cross-loadings were estimated. Table 10 shows that the mean estimate was largely unaffected, but the variance of the estimates (not reported in the table) increased because of the increased model

Table 10. Mean Correlation Estimate Under Model Misspecification.

Cross-Loading Estimation	Technique	Population Cross-Loadings and Factor Correlation											
		No Cross-Loadings				1 Cross-Loading				2 Cross-Loadings			
		.7	.8	.9	1.0	.7	.8	.9	1.0	.7	.8	.9	1.0
Assumed zero	ρ_{CFA}	.699	.800	.901	1.001	.819	.878	.938	1.002	.885	.921	.957	1.000
	ρ_{DPR}	.707	.808	.911	1.012	.808	.876	.943	1.011	.879	.922	.962	1.008
	ρ_{DTR}	.708	.809	.911	1.013	.808	.876	.943	1.012	.879	.922	.963	1.008
	ρ_{DCR}	.700	.800	.901	1.002	.801	.868	.935	1.002	.871	.914	.955	1.000
	ρ_{SS}	.560	.640	.720	.801	.633	.692	.749	.801	.700	.740	.777	.805
Estimated	ρ_{CFA}					.700	.801	.903	1.009	.694	.790	.894	1.021

Note: Averages the correlations over the conditions shown in Table 6. Techniques: CFA= confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability; SS = summed scale without disattenuation.

complexity. This effect is seen in Table 11, where the pattern of results for CFA models was largely similar between the cross-loading conditions, but the presence of cross-loadings increased the false positive rate. This result is easiest to understand in the context of $CI_{CFA}(cut)$; when estimates became less precise, this also widened the confidence intervals and, consequently, increased the frequency of results where the cutoff fell within the interval.

The cross-loading results underline the importance of observing the assumptions of the techniques and that not doing so may lead to incorrect inference. However, two conclusions that are new to discriminant validity literature can be drawn: First, the lack of cross-loadings in the population (i.e., factorial validity) is not a strict prerequisite for discriminant validity assessment as long as the cross-loadings are modeled appropriately. Second, while the lack of factorial validity can lead scale-item pairs to have complete lack of discriminant validity (see Equation 2), this does not always invalidate scale-level discriminant validity (see Equation 3) as long as this is properly modeled.

Discussion

The original meaning of the term “discriminant validity” was tied to MTMM matrices, but the term has since evolved to mean a lack of a perfect or excessively high correlation between two measures after considering measurement error. We provided a comprehensive review of the various discriminant validity techniques and presented a simulation study assessing their effectiveness. There are several issues that warrant discussion.

Contradictions With Prior Studies

Our simulation results clearly contradict two important conclusions drawn in the recent discriminant validity literature, and these contradictions warrant explanations. First, Henseler et al. (2015) and Voorhees et al. (2016) strongly recommend ρ_{DPR} (HTMT) for discriminant validity assessment. We prove that the HTMT index is simply a scale score correlation disattenuated with parallel reliability (i.e., the standardized alpha) and thus should not be expected to outperform modern CFA techniques, which our simulation demonstrates. The different conclusions are due to the limitations of these prior studies. While Henseler et al. (2015) explain that ρ_{DPR} is a factor correlation estimate, they do not compare it against other factor correlation estimation techniques. Thus, their results do not indicate the superiority of ρ_{DPR} but simply indicate that AVE/SV, which was their main comparison, performs very poorly. The follow-up study by Voorhees et al. (2016) considered a broader set of techniques,

Table 11. Detection Rates by Technique Using Alternative Cutoffs Under Model Misspecification.

Cross-loadings and Whether Estimated		Cutoff and Factor Correlation											
		.85				.9				.95			
		.7*	.8*	.9 [†]	1.0 [†]	.7*	.8*	.9 [†]	1.0 [†]	.7*	.8*	.9*	1.0 [†]
No cross-loadings Assumed zero	$\rho_{CFA}(\text{cut})$.034	.152	.886	.980	.018	.055	.511	.965	.011	.026	.101	.927
	$\rho_{DPR}(\text{cut})$.052	.216	.906	.986	.029	.091	.592	.973	.018	.045	.175	.941
	$\rho_{DCR}(\text{cut})$.037	.161	.883	.982	.019	.060	.514	.966	.012	.029	.111	.920
	$CI_{CFA}(\text{cut})$.313	.697	.997	.999	.169	.400	.984	.998	.091	.199	.555	.997
	$CI_{DPR}(\text{cut})$.356	.735	.999	1.000	.225	.455	.988	1.000	.152	.276	.620	1.000
	$CI_{DCR}(\text{cut})$.315	.691	.999	1.000	.192	.400	.983	1.000	.129	.227	.558	.999
	CFI(cut)	.248	.810	.989	.998	.121	.341	.961	.996	.075	.151	.557	.991
	$\chi^2(\text{cut})$.251	.641	.997	.999	.137	.328	.973	.999	.086	.167	.478	.998
1 cross-loading Assumed zero	$\rho_{CFA}(\text{cut})$.445	.837	.958	.980	.168	.442	.895	.965	.038	.098	.450	.925
	$\rho_{DPR}(\text{cut})$.253	.757	.962	.986	.109	.297	.883	.974	.051	.115	.368	.938
	$\rho_{DCR}(\text{cut})$.194	.739	.954	.982	.067	.229	.863	.965	.029	.067	.290	.919
	$CI_{CFA}(\text{cut})$.900	.996	.999	.999	.621	.917	.997	.999	.311	.541	.940	.997
	$CI_{DPR}(\text{cut})$.732	.996	1.000	1.000	.469	.787	.999	1.000	.288	.462	.867	.999
	$CI_{DCR}(\text{cut})$.727	.996	1.000	1.000	.425	.781	.999	1.000	.237	.408	.859	.999
	CFI(cut)	.912	.985	.995	.998	.596	.930	.989	.996	.214	.482	.946	.991
	$\chi^2(\text{cut})$.880	.995	.999	1.000	.563	.895	.997	.999	.254	.469	.918	.998
1 cross-loading Estimated	$\rho_{CFA}(\text{cut})$.058	.191	.848	.962	.037	.088	.518	.938	.025	.051	.160	.883
	$CI_{CFA}(\text{cut})$.385	.750	.994	.998	.245	.488	.978	.996	.155	.287	.641	.994
	CFI(cut)	.367	.871	.986	.996	.216	.498	.964	.993	.147	.267	.745	.987
	$\chi^2(\text{cut})$.342	.711	.996	.999	.221	.433	.972	.998	.157	.264	.591	.995
2 cross-loadings Assumed zero	$\rho_{CFA}(\text{cut})$.855	.943	.970	.977	.411	.819	.936	.961	.067	.203	.740	.918
	$\rho_{DPR}(\text{cut})$.779	.940	.975	.983	.307	.761	.941	.969	.110	.230	.709	.932
	$\rho_{DCR}(\text{cut})$.760	.930	.970	.979	.235	.733	.929	.961	.057	.143	.663	.914
	$CI_{CFA}(\text{cut})$.996	.998	.999	.999	.938	.996	.998	.998	.492	.773	.992	.996
	$CI_{DPR}(\text{cut})$.997	1.000	1.000	1.000	.803	.996	.999	1.000	.459	.693	.995	.999
	$CI_{DCR}(\text{cut})$.996	1.000	1.000	1.000	.793	.995	1.000	1.000	.397	.660	.993	.999
	CFI(cut)	.986	.994	.996	.997	.941	.984	.992	.995	.434	.858	.976	.990
	$\chi^2(\text{cut})$.996	.998	.999	.999	.916	.994	.998	.998	.415	.720	.989	.996
2 cross-loadings Estimated	$\rho_{CFA}(\text{cut})$.105	.257	.791	.922	.071	.150	.530	.895	.045	.091	.239	.837
	$CI_{CFA}(\text{cut})$.510	.814	.987	.993	.384	.614	.973	.990	.291	.444	.752	.987
	CFI(cut)	.565	.921	.983	.992	.408	.704	.968	.988	.303	.466	.885	.980
	$\chi^2(\text{cut})$.482	.790	.992	.997	.361	.573	.971	.995	.276	.408	.709	.990

Note: Averages over all conditions. Techniques: ρ = comparing the correlations against a cutoff; CI = comparing if a cutoff is included in the 95% confidence interval; CFA= confirmatory factor analysis; Dxx = disattenuated correction using PR = parallel reliability, TR = tau-equivalent reliability, and CR = congeneric reliability; CFI = nested model comparison using the CFI rule; χ^2 = nested model test.

*False positive rate. [†]true positive rate = (1 – false negative rate).

including $CI_{CFA}(1)$ and $\chi^2(1)$. The problem in their study was that the different techniques were applied using different cutoffs: ρ_{DPR} was used with cutoffs of .80, .85, and .90, whereas the other techniques always used the cutoff of 1 and were thus predestined to fail in a study where a correlation of .90 was used as a discriminant validity problem condition. A more complete study would have used the same cutoffs (.80, .85, and .90) that were applied to ρ_{DPR} with χ^2 and CI_{CFA} as well, as we did.

Second, our results also challenge J. A. Shaffer et al.'s (2016) recommendation that discriminant validity should be tested by a CFI comparison between two nested models (CFI(1)). The reason for this contradiction is that their recommendation is based on the assumption that if the CFI difference

works well for measurement invariance assessment, it should also work well when discriminant validity assessment. The assumption appears to be invalid as it ignores an important difference between these uses: Whereas the degrees of freedom of an invariance test scale roughly linearly with the number of indicators, the degrees of freedom in CFI(1) are always one.

Magnitude of the Discriminant Validity Correlations

In the discriminant validity literature, high correlations between scales or scale items are considered problematic. However, the literature generally has not addressed what is high enough beyond giving rule of thumb cutoffs (e.g., .85). Our definition of discriminant validity suggests that the magnitude of the estimated correlation depends on the correlation between the constructs, the measurement process, and the particular sample, each of which has different implications on what level should be considered high. To warn against mechanical use, we present a scenario where high correlation does not invalidate measurement and a scenario where low correlation between measures does not mean that they measure distinct constructs.

A large correlation does not always mean a discriminant validity problem if one is expected based on theory or prior empirical observations. For example, the correlation between biological sex and gender identity can exceed .99 in the population.¹⁷ However, both variables are clearly distinct: sex is a biological property with clear observable markers, whereas gender identity is a psychological construct. These two variables also have different causes and consequences (American Psychological Association, 2015), so studies that attempt to measure both can lead to useful policy implications. In cases such as this where the constructs are well defined, large correlations should be tolerated when expected based on theory and prior empirical results. Of course, large samples and precise measurement would be required to ensure that the constructs can be distinguished empirically (i.e., are empirically distinct).

A small or moderate correlation (after correcting for measurement error) does not always mean that two measures measure concepts that are distinct. For example, consider two thermometers that measure the same temperature, yet one is limited to measuring only temperatures above freezing, whereas the other can measure only temperatures below freezing. While both measure the same quantity, they are correlated only by approximately .45 because the temperature would always be out of the range of one of the thermometers that would consequently display zero centigrade.¹⁸ In the social sciences, a well-known example is the measurement of happiness and sadness, two constructs that can be thought of as opposite poles of mood (D. P. Green et al., 1993; Tay & Jebb, 2018). Consequently, any evaluation of the discriminant validity of scales measuring two related constructs must precede the theoretical consideration of the existence of a common continuum. If this is the case, the typical discriminant validity assessment techniques that are the focus of our article are not directly applicable, but other techniques are needed (Tay & Jebb, 2018).

As the two examples show, a moderately small correlation between measures does not always imply that two constructs are distinct, and a high correlation does not imply that they are not. Like any validity assessment, discriminant validity assessment requires consideration of context, possibly relevant theory, and empirical results and cannot be reduced to a simple statistical test and a cutoff no matter how sophisticated. These considerations highlight the usefulness of the continuous interpretation of discriminant validity evidence.

On Choosing a Technique and a Cutoff

While a general set of statistics and cutoffs that is applicable to all research scenarios cannot exist, we believe that establishing some standards is useful. Based on our study, $CI_{CFA}(\text{cut})$ and $\chi^2(\text{cut})$ appear to be the leading techniques, but recommending one over another solely on a statistical basis

is difficult due to the similar performance of the techniques. However, we recommend $CI_{CFA}(cut)$ for practical reasons. First, $CI_{CFA}(cut)$ is less likely to be misused than $\chi^2(cut)$. While we found some evidence of misapplication of $\chi^2(cut)$ due to incorrect factor scaling, we did not see any evidence of the same when factor correlations were evaluated; these can be obtained postestimation simply by requesting standardized estimates from the software. Second, $CI_{CFA}(cut)$ makes it easier to transition from testing of discriminant validity to its evaluation because the focal statistic is a correlation, which organizational researchers routinely interpret in other contexts. In sum, $CI_{CFA}(cut)$ is simpler to implement, easier to understand, and less likely to be misapplied. $\chi^2(cut)$ is slightly more accurate, but considering that even the simpler $\chi^2(1)$ is often misapplied, we do not think that the potential precision gained by using $\chi^2(cut)$ is worth the cost of risking misapplication.

Equally important to choosing a technique is the choice of a cutoff if the technique requires one. Of the recent simulation studies, Henseler et al. (2015) suggested cutoffs of .85 and .9 based on prior literature (e.g., Kline, 2011). Voorhees et al. (2016) considered the false positive and false negative rates of the techniques used in their study and concluded that the cutoff of .85 had the best balance of high power and an acceptable Type I error rate. However, as explained in Online Supplement 5, such conclusions are to a large part simply artifacts of the simulation design. In sum, it seems that deriving an ideal cutoff through simulation results is meaningless and must be established by consensus among the field.

Empirical studies seem to agree that correlations greater than .9 indicate a problem and that correlations less than .8 indicate the lack of a problem. For example, Le et al. (2010) diagnosed a discriminant validity problem between job satisfaction and organizational commitment based on a correlation of .91, and Mathieu and Farr (1991) declared no problem of discriminant validity between the same variables on the basis of a correlation of .78. However, mixed judgments were made about the correlation values between .8 and .9. Lucas et al. (1996) acknowledged discriminant validity between self-esteem and optimism based on ρ_{DTR} of .83 ($\rho_{SS} = .72$), but Créde et al. (2017) criticized the conceptual redundancy between grit and conscientiousness based on a disattenuated correlation of .84 ($\rho_{SS} = .66$). Thus, if there is a threshold, it is likely to fall between .8 and .9.

Sources of Error Other Than the Random Error and Item-Specific Factor Error

Our focus on the common scenario of single-method and one-time measurements limits the researcher to techniques that operationalize measurement error as item uniqueness either directly (e.g., CFA) or indirectly (e.g., ρ_{DTR}). However, several articles point out that this may be a simplistic view of measurement error that only considers random and item-specific factor errors, ignoring time-specific transient errors (Le et al., 2009; Woehr et al., 2012). Because transient error can correlate between items and scales, they can either inflate or attenuate correlation estimates calculated using single-administration reliability estimates. While there is little that can be done about this issue if one-time measures are used, researchers should be aware of this limitation. Of course, if multiple measurement occasions are possible, the CFA-based techniques can also be used to model these other sources of error (Le et al., 2009; Woehr et al., 2012), and these more complex models can then be applied with the guideline that we present next.

A Guideline for Assessing Discriminant Validity

From a Cutoff to a Classification System

While many articles about discriminant validity consider it as a matter of degree (e.g., “the extent to . . .”) instead of a yes/no issue (e.g., “whether . . .”), most guidelines on evaluation techniques,

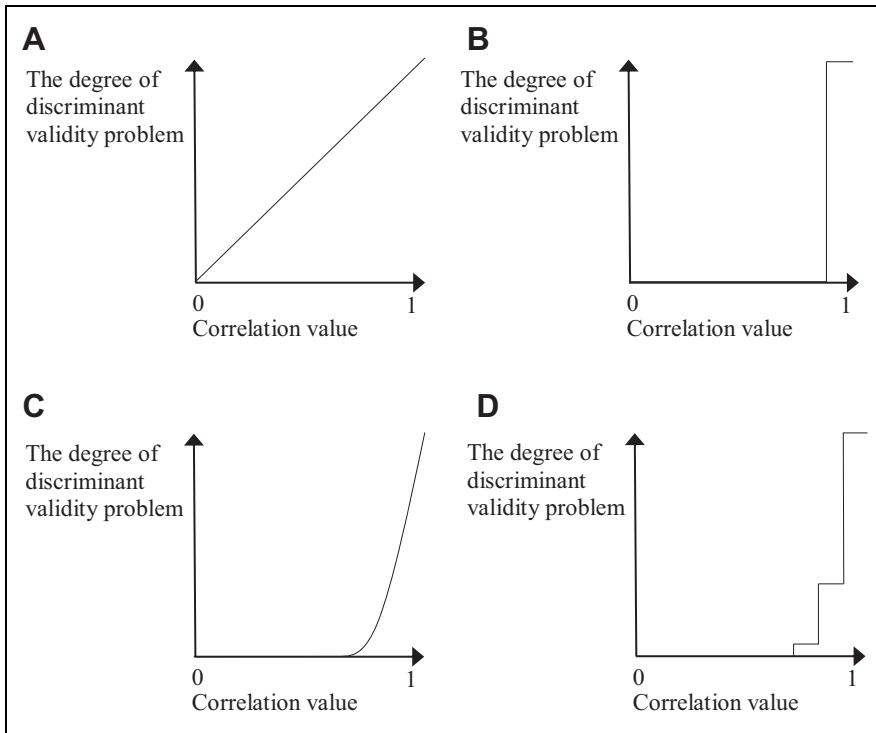


Figure 6. Relationship between correlation values and the problem of discriminant validity. (A) Linear model (implied by existing definitions), (B) Dichotomous model (existing techniques), (C) Threshold model (implied by the definition of this study), (D) Step model (proposed evaluation technique).

Table 12. Proposed Classification and Cutoffs.

Classification	CI_{CFA} (sys)	χ^2 (sys)
Severe problem	$1 \leq UL$	$\chi_1^2 - \chi_{org}^2 < 3.84$
Moderate problem	$.9 \leq UL < 1$	Not “Marginal problem” AND $\chi_1^2 - \chi_{org}^2 > 3.84$
Marginal problem	$.8 \leq UL < .9$	Not “No problem” AND $\chi_{.9}^2 - \chi_{org}^2 > 3.84$
No problem	$UL < .8$	$\rho_{CFA} < .8$ AND $\chi_{.8}^2 - \chi_{org}^2 > 3.84$

Note: ρ_{CFA} is the correlation obtained using CFA, UL is the 95% upper limit of ρ_{CFA} when $\rho_{CFA} > 0$, and the absolute value of the 95% lower limit of ρ_{CFA} when $\rho_{CFA} < 0$, χ_{org}^2 is the chi-square value of the original model, and χ_c^2 is the chi-square value of the comparison model where the focal correlation is fixed to c when $\rho_{CFA} > 0$ and $-c$ when $\rho_{CFA} < 0$.

including Campbell and Fiske’s (1959) original proposal, focus on making a dichotomous judgment as to whether a study has a discriminant validity problem (B in Figure 6). This inconsistency might be an outcome of researchers favoring cutoffs for their simplicity, or it may reflect the fact that after calculating a discriminant validity statistic, researchers must decide whether further analysis and interpretation is required. This practice also fits the dictionary meaning of *evaluation*, which is not simply calculating and reporting a number but rather dividing the object into several qualitative categories based on the number, thus requiring the use of cutoffs, although these do not need to be the same in every study.

To move the field toward discriminant validity evaluation, we propose a system consisting of several cutoffs instead of a single one. Discriminant validity is a continuous function of correlation

values (C in Figure 6), but because of practical needs, correlations are classified into discrete categories indicating different degrees of a problem (D in Figure 6). Similar classifications are used in other fields to characterize essentially continuous phenomena: Consider a doctor's diagnosis of hypertension. In the past, everyone was divided into two categories of normal and patient, but now hypertension is classified into several levels. That is, patients with hypertension are further subdivided into three stages according to their blood pressure level, and each level is associated with different treatments.

Table 12 shows the classification system we propose. We emphasize that these are guideline that can be adjusted case-by-case if warranted by theoretical understanding of the two constructs and measures, not strict rules that should always be followed. Based on our review, correlations below .8 were seldom considered problematic, and this is thus used as the cutoff for the first class, "No problem," which strictly speaking is not a proof of no problem, just no evidence of a problem. When correlations fall into this class, researchers can simply declare that they did not find any evidence of a discriminant validity problem. The next three steps are referred to as Marginal, Moderate, and Severe problems, respectively. The Severe problem is the most straightforward: two items or scales cannot be distinguished empirically, and researchers should rethink their concept definitions, measurement, or both. In empirical applications, the correlation level of .9 was nearly universally interpreted as a problem, and we therefore use this level as a cutoff between the Marginal and Moderate cases. In both cases, the high correlation should be acknowledged, and its possible cause should be discussed. In the Marginal case, the interpretation of the scales as representations of distinct constructs is probably safe. In the Moderate case, additional evidence from prior studies using the same constructs and/or measures should be checked before interpretation of the results to ensure that the high correlation is not a systematic problem with the constructs or scales.

How to Implement the Proposed Techniques

The proposed classification system should be applied with $CI_{CFA}(\text{cut})$ and $\chi^2(\text{cut})$, and we propose that these workflows be referred to as $CI_{CFA}(\text{sys})$ and $\chi^2(\text{sys})$, respectively. Both workflows start by estimating a CFA model that includes all scales that are evaluated for discriminant validity. Instead of using the default scale setting option to fix the first factor loadings to 1, scale the latent variables by fixing their variances to 1 (A in Figure 2); this should be explicitly reported in the article. The covariances between factors obtained in the latter way equal the correlations; alternatively, when using $CI_{CFA}(\text{sys})$, the standardized factor solution can be inspected. Next, inspect the upper limits (lower limits for negative correlations) of the 95% CIs of the estimated factor correlations and compare their values against the cutoffs in Table 12.¹⁹

Implementing $\chi^2(\text{sys})$ requires testing every correlation against the lower limit of each class in the classification system. A correlation belongs to the highest class that it is not statistically significantly different from. If the model tests cannot be automated, we suggest the following alternative workflow. First, exclude all correlation pairs whose upper limit of the CI is less than .80. Second, for the remaining correlations, determine the initial class for each by comparing correlation estimates against the cutoffs in Table 12. For example, a correlation of .87 would be classified as Marginal. Third, use $\chi^2(\text{cut})$ to compare the estimated model against a model where the correlation is constrained to the high cutoff, .9 in the example, using a nested model χ^2 test. If significantly different, the correlation is classified into the current section. If the correlation is not significantly different, repeat the model comparison by selecting the high cutoff for the next higher section (in this case 1).

Online Supplement 4 provides a tutorial on how to implement the techniques described in this article using AMOS, LISREL, Mplus, R, and Stata. Because implementing a sequence of

comparisons is cumbersome and prone to mistakes, we have contributed a function that automates the $\chi^2(\text{cut})$ tests to the semTools R package (Jorgensen et al., 2020). For users of other software, we developed MQAssessor,²⁰ a Python-based open-source application.

What to Do When Discriminant Validity Fails?

If problematically high correlations are observed, their sources must be identified. We propose a three-step process: First, suspect conceptual redundancy. We suggest starting by following the guidelines by J. A. Shaffer et al. (2016) and Podsakoff et al. (2016) for assessing the conceptual distinctiveness of the constructs (see also M. S. Krause, 2012). If two constructs are found to overlap conceptually, researchers should seriously consider dropping one of the constructs to avoid the confusion caused by using two different labels for the same concept or phenomenon (J. A. Shaffer et al., 2016).

Second, scrutinize the measurement model. An unexpectedly high correlation estimate can indicate a failure of model assumptions, as demonstrated by our results of misspecified models. Check the χ^2 test for an exact fit of the CFA model. If this test fails, diagnose the model with residuals and/or modification indices to understand the source of misspecification (Kline, 2011, chap. 8). If the model is modified based on these considerations, the wording of the items that led to these decisions should be explicitly reported, and how the item wordings justify the modifications should be explained to reduce the risk of data mining.

Third, collect different data. If conceptual overlap and measurement model issues have been ruled out, the discriminant validity problem can be reduced to a multicollinearity problem. For example, if one wants to study the effects of hair color and gender on intelligence but samples only blonde men and dark-haired women, hair color and gender are not empirically distinguishable, although they are both conceptually distinct and virtually uncorrelated in the broader population. This can occur either because of a systematic error in the sampling design or due to chance in small samples. If a systematic error can be ruled out, the most effective remedy is to collect more data. Alternatively, the data can be used as such, in which case large standard errors will indicate that little can be said about the relative effects of the two variables, or the two variables can be combined as an index (Wooldridge, 2013, pp. 94–98). If a researcher chooses to interpret results, he or she should clearly explain why the large correlation between the latent variables (e.g., $>.9$) is not a problem in the particular study.

Reporting

We also provide a few guidelines for improved reporting. First, researchers should clearly indicate what they are assessing when assessing discriminant validity by stating, for example, that “We addressed discriminant validity (whether two scales are empirically distinct).” Second, the correlation tables, which are ubiquitous in organizational research, are in most cases calculated with scale scores or other observed variables. However, most studies use only the lower triangle of the table, leaving the other half empty (*AMJ* 93.6%, *JAP* 83.1%). This practice is a waste of scarce resources, and we suggest that this space should be used for the latent correlation estimates, which serve as continuous discriminant validity evidence. Third, if nested model comparisons (e.g., $\chi^2(1)$) are used, researchers should explicitly report that the model was rescaled from the default option by stating, for example, “We used the χ^2 nested model comparison for assessing discriminant validity by comparing our CFA model against models that were more constrained, where all factor loadings were freely estimated, the factor variances were constrained to 1 and each factor correlation was constrained to 1 one at a time.” These reporting practices should considerably reduce the ambiguity in the literature and prevent the common misapplication of the $\chi^2(1)$ test.

Concluding Remarks

There is no shortage of various statistical techniques for evaluating discriminant validity. Such an abundance of techniques is positive if techniques have different advantages, and they are purposefully selected based on their fit with the research scenario. The current state of the discriminant validity literature and research practice suggests that this is not the case. Instead, it appears that many of the techniques have been introduced without sufficient testing and, consequently, are applied haphazardly. Because direct criticism of existing techniques is often avoided, there appears to be a tendency in which new techniques continue to be added without clarifying the problems of previously used techniques. This *technique proliferation* causes confusion and misuse. This study draws an unambiguous conclusion about which method is best for assessing discriminant validity and which methods are inappropriate. We hope that the article will help readers discriminate valid techniques from those that are not.

Appendix: Proofs

Proof That the HTMT Index Is Algebraically Equivalent to Disattenuated Correlation Using Parallel Reliability (i.e., the Standardized Alpha)

The HTMT index for scales X and Y was originally defined as follows (Henseler et al., 2015):

$$\text{HTMT} = \frac{1}{K_X K_Y} \sum_{g=1}^{K_X} \sum_{h=1}^{K_Y} r_{Xg, Yh} \div \left(\frac{2}{K_X(K_X - 1)} \sum_{g=1}^{K_X-1} \sum_{h=g+1}^{K_X} r_{Xg, Xh} \frac{2}{K_Y(K_Y - 1)} \sum_{g=1}^{K_Y-1} \sum_{h=g+1}^{K_Y} r_{Yg, Yh} \right)^{\frac{1}{2}} \quad (\text{A1})$$

The equation can be simplified considerably by expressing it as a function of three algebraic means (i.e., the sum divided by the count):

$$\text{HTMT} = \frac{\bar{r}_{XY}}{\sqrt{\bar{r}_{XX} \bar{r}_{YY}}}, \quad (\text{A2})$$

where \bar{r} is the mean of *nonredundant* correlations.

We will now prove that the HTMT index is equivalent to the scale score correlation disattenuated with the parallel reliability coefficient. Parallel reliability (i.e., the standardized alpha) is given as follows:

$$\rho_P = \frac{k^2 \bar{\sigma}}{k^2 \bar{\sigma} + k(1 - \bar{\sigma})}, \quad (\text{A3})$$

where K is the number of scale items (Cho, 2016). The disattenuated correlation between two unit-weighted composites X and Y of p and q items using parallel reliability as reliability estimates is given as follows:

$$\rho_{DPR} = \frac{\frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}}{\sqrt{\alpha_X \alpha_Y}}, \quad (\text{A4})$$

where

$$\text{cov}(X, Y) = \sum_{i=1}^p \sum_{j=p+1}^{p+q} r_{ij} = pq \bar{r}_{XY} \quad (\text{A5})$$

$$\text{var}(X) = \sum_{i=1}^p \sum_{j=i}^p r_{ij} = \overline{r_{XX}}p(p-1) + p \quad (\text{A6})$$

$$\text{var}(Y) = \sum_{i=1}^q \sum_{j=i}^q r_{ij} = \overline{r_{YY}}q(q-1) + q \quad (\text{A7})$$

Substituting Equations A5, A6, and A7 into Equation A4, the equation is as follows:

$$\rho_{DPR} = \frac{\frac{pq\overline{r_{XY}}}{\sqrt{\left(\overline{r_{XX}}p(p-1)+p\right)\left(\overline{r_{YY}}q(q-1)+q\right)}}}{\sqrt{\left(\frac{p\overline{r_{XX}}}{1+(p-1)\overline{r_{XX}}}\right)\left(\frac{q\overline{r_{YY}}}{1+(q-1)\overline{r_{YY}}}\right)}} \quad (\text{A8})$$

Rearranging:

$$\frac{\frac{pq\overline{r_{XY}}}{\sqrt{p\left(1+(p-1)\overline{r_{XX}}\right)q\left(1+(q-1)\overline{r_{YY}}\right)}}}{\sqrt{\left(\frac{p\overline{r_{XX}}}{1+(p-1)\overline{r_{XX}}}\right)\left(\frac{q\overline{r_{YY}}}{1+(q-1)\overline{r_{YY}}}\right)}} \quad (\text{A9})$$

Simplifying:

$$\frac{\frac{pq\overline{r_{XY}}}{\sqrt{pq}}}{\sqrt{p\overline{r_{XX}}q\overline{r_{YY}}}} = \frac{\sqrt{pq}\overline{r_{XY}}}{\sqrt{p\overline{r_{XX}}q\overline{r_{YY}}}} = \frac{\overline{r_{XY}}}{\sqrt{\overline{r_{XX}}\overline{r_{YY}}}}, \quad (\text{A10})$$

which equals the HTMT index shown in Equation A2.

Proof That CFI(1) Is Equivalent to Alternative Critical Values for $\chi^2(1)$

Based on a study of measurement invariance assessment by Meade et al. (2008), J. A. Shaffer et al. (2016) suggest that comparing the differences in the CFIs between the two models instead of χ^2 can produce a test whose result is less dependent on sample size than the $\chi^2(1)$ test. Their recommended cutoff for the difference was .002. We will now prove that the CFI comparison is equivalent to a χ^2 test that uses a critical value based on the null model instead of the χ^2 distribution.

CFI is defined as follows:

$$\text{CFI} = 1 - \frac{\chi_M^2 - df_M}{\chi_B^2 - df_B}, \quad (\text{A11})$$

where M is the model of interest and B is the baseline or null model. In the CFI(1) test, both the constrained and unconstrained models are evaluated against the same baseline. Thus, the CFI difference can be written as follows:

$$\begin{aligned} \Delta\text{CFI} &= \left(1 - \frac{\chi_M^2 - df_M}{\chi_B^2 - df_B}\right) - \left(1 - \frac{\chi_C^2 - df_C}{\chi_B^2 - df_B}\right) \\ &= \frac{(\chi_C^2 - df_C) - (\chi_M^2 - df_M)}{\chi_B^2 - df_B} \end{aligned} \quad (\text{A12})$$

$$\begin{aligned}
 &= \frac{\chi_C^2 - \chi_M^2 - 1}{\chi_B^2 - df_B} \\
 &= \frac{\Delta\chi^2 - 1}{\chi_B^2 - df_B},
 \end{aligned}$$

where C is the constrained model in which a correlation value is fixed to 1 in the model of interest (i.e., M).

Therefore, the comparison against the .002 cutoff can be written as a $\Delta\chi^2$ test where the reference point is not a critical value of the χ^2 distribution, but discriminant validity holds conditional on the following:

$$\frac{\Delta\chi^2 - 1}{\chi_B^2 - df_B} > .002 \quad (\text{A13})$$

$$\Delta\chi^2 - 1 > .002(\chi_B^2 - df_B)$$

$$\Delta\chi^2 > 1 + .002(\chi_B^2 - df_B)$$

Acknowledgments

We are deeply grateful to Louis Tay, associate editor, and two anonymous reviewers for their invaluable guidance and constructive comments. We acknowledge the computational resources provided by the Aalto Science-IT project.


Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by a grant from the Academy of Finland (Grant 311309) and the Research Grant of Kwangwoon University in 2019.

ORCID iD

Mikko Rönkkö  <https://orcid.org/0000-0001-7988-7609>

Eunseong Cho  <https://orcid.org/0000-0003-1818-0532>

Notes

1. The existence of constructs independently of measures (realism), although often implicit, is commonly assumed in the discriminant validity literature. This assumption was also present in the original article by Campbell and Fiske (1959) that assumed a construct to be a source of variation in the items thus closely corresponding to the definition of validity by Borsboom et al. (2004).
2. We are grateful for the comments by the anonymous reviewer who helped us come up with this definition.
3. The desirable pattern of correlations in a factorial validity assessment is similar to the pattern in discriminant validity assessment in an MTMM study (Spector, 2013), so in practice the difference between discriminant validity and factorial validity is not as clear-cut. Nevertheless, there is a clear conceptual difference between the two. In fact, the concept of factorial validity predates discriminant validity (Guilford, 1946), and neither Campbell and Fiske (1959) nor any of the reviewed articles presented discriminant

validity and factorial validity as the same concept. Importantly, factorial validity is an attribute of “a test” (Guilford, 1946), whereas only pairs of measures can exhibit discriminant validity.

4. Of the *AMJ* and *JAP* articles reviewed, most reported a correlation table (*AMJ* 96.9%, *JAP* 89.3%), but most did not specify whether the reported correlations were scale score correlations or factor correlations (*AMJ* 100%, *JAP* 98.5%). A plausible expectation is that studies that do not use SEMs report scale score correlations and that in studies that use SEMs, the presented correlations are factor correlations. To verify this assumption, we took a random sample of 49 studies out of the 199 studies that applied SEMs and emailed the authors to ask for the type of correlation used in the study. The responses of the 21 available replies were all scale score correlations. That is, a correlation that is not specified as a factor correlation can almost always be regarded as a scale score correlation.
5. The disattenuation equation shows that the scale score correlation is constrained to be no greater than at the geometric mean of the two reliabilities. For example, if a researcher uses the commonly used cutoff of .9 to make a yes/no decision about discriminant validity, a no decision can never be reached unless both scales are very reliable (i.e., the square root of the product of the reliabilities exceeds .9).
6. Different variations of disattenuated correlations can be calculated by varying how the scale score correlation is calculated, how reliabilities are estimated, or even the disattenuation equation itself. The reliability indices that we discuss assume that the scale score is calculated using equal weights. Another common approach is to apply unit weights. In this approach, the observed variables are first standardized before taking a sum or a mean; alternatively, a weighted sum or mean with $1/\sigma_{x_i}$ is taken as the weights (i.e., $X = \sum_i X_i/\sigma_{x_i}$) (Bobko et al., 2007). Using factor scores in this context is not a good idea because the reliability will be positively biased (Aguirre-Urreta et al., 2019), and, consequently, the correlation will be undercorrected.
7. We use the term “single-admission reliability” (Cho, 2016; Zijlmans et al., 2018) instead of the more commonly used “internal consistency reliability” because the former is more descriptive and less likely to be misunderstood than the latter (Cho & Kim, 2015).
8. Strictly speaking, tau-equivalence implies that item means are equal and the qualifier essentially relaxes this constraint. The constraint itself does not affect the value of reliability coefficients. We focus on essentially tau-equivalent, essentially parallel, and essentially congeneric conditions, but we omit the term essentially for convenience.
9. We follow the terminology from Cho (2016) because the conventional names provide (a) inaccurate information about the original author of each coefficient and (b) confusing information about the nature of each coefficient. Following Cho’s (2016) suggestion and including the assumption of each reliability coefficient in the name will hopefully also reduce the chronic misuse of these reliability coefficients.
10. Notably, Bagozzi (1981) wrote a critical commentary, to which Fornell and Larcker (1981b) published a rejoinder, but neither of these articles addressed the issues that we raise in this article.
11. A full discriminant validity analysis requires the pairwise comparisons of all possible factor pairs. The number of required model comparisons is the number of unique correlations between the variables, given by $k(k - 1)/2$, where k is the number of factors. In large models, manually specifying all these models and calculating model comparisons is tedious and possibly error prone.
12. If factor variances are estimated a correlation constraint can be implemented with a nonlinear constraint ($\rho_{12} = \frac{\phi_{12}}{\sqrt{\phi_{11}\phi_{22}}} = 1$). Because this is complicated, $\chi^2(1)$ has been exclusively applied by constraining the factor covariance to be 1.
13. While it is nearly impossible to identify scaling errors without access to the actual analysis and CFA result files or the item level covariance matrix, which would allow different specifications to be tested by replication, there exists indirect evidence of this problem. By analyzing the relationships between the reported item scales (e.g., 5 vs. 7 point) or variances, correlation estimates, and $\Delta\chi^2$ values, we could find instances where a model with estimated correlation that was not close to 1 (e.g., .8) did not fit statistically significantly better than a model for which the covariance was constrained to be 1; however, a model with a

correlation close to 1.0 (e.g., .95) was significantly different from the constrained model, while at the same time the $\Delta\chi^2$ clearly depended on the item scales. Based on this indirect evidence, we conclude that erroneous specification of the constraint is quite common in both methodological guidelines and empirical applications. Of course, this problem is not unique to the χ^2 test but applies to all nested model comparisons regardless of which statistic is used to compare the models.

14. We thank Terrence Jorgensen for pointing this out.
15. In empirical applications, the term “loading” typically refers to pattern coefficients, a convention that we follow. In a CFA, the model parameters are pattern coefficients, and these are also more commonly reported in EFA applications (Henson & Roberts, 2006).
16. For example, Henseler et al. (2015) defined a cross-loading when the loading (i.e., structure coefficient) between an item and its unintended factor is greater than the loading between the item and its intended factor. If the pattern coefficients have no cross-loadings, this condition is equivalent to say that the factor correlation is greater than 1 (see Table 5). This mathematical fact is why the cross-loading technique produced strange results in their simulation, which was not explained in the original paper.
17. Consider two binary variables “to which gender do you identify” and “what is your biological sex.” If 0.5% of the population are transgender or gender nonconforming (American Psychological Association, 2015) and half of these people indicate identification to a gender opposite to their biological sex, the correlation between the two variables would be .995.
18. The two hypothetical measures have a floor and ceiling effect, which leads to nonrandom measurement errors and a violation of the assumption underlying the disattenuation. This example demonstrates that researchers who use systematically biased measures cannot accurately assess discriminant validity. Thanks to the reviewer for pointing this out.
19. These CIs are reported as part of the default output of most modern SEM software, but if they are not available, they can be calculated using the estimates and standard errors as follows:

$$UL = \rho_{CFA} + 1.96 \times SE(\rho_{CFA}) \text{ and } LL = \rho_{CFA} - 1.96 \times SE(\rho_{CFA}).$$
20. The software is available at <https://github.com/eunscho/MQAssessor>.

References

- Aguirre-Urreta, M. I., Rönkkö, M., & McIntosh, C. N. (2019). A cautionary note on the finite sample behavior of maximal reliability. *Psychological Methods, 24*(2), 236-252. <https://doi.org/10.1037/met0000176>
- American Psychological Association. (2015). Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist, 70*(9), 832-864. <https://doi.org/10.1037/a0039906>
- Anderson, J., & Gerbing, D. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423. <https://doi.org/10.1037/0033-2909.103.3.411>
- Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeier et al. *Journal of Management, 41*(6), 1561-1577. <https://doi.org/10.1177/0149206315591075>
- Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: A comment. *Journal of Marketing Research, 18*(3), 375-381. <https://doi.org/10.2307/3150979>
- Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly, 27*(3), 459-489. <https://doi.org/10.2307/2392322>
- Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421-458. <https://doi.org/10.2307/2393203>
- Bartholomew, D. J. (2007). Three faces of factor analysis. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 9-21). Lawrence Erlbaum.
- Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689-709. <https://doi.org/10.1177/1094428106294734>

- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *1904-1920*, *3*(3), 296-322. <https://doi.org/10.1111/j.2044-8295.1910.tb00207.x>
- Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, *15*(8), 546-553. <https://doi.org/10.1037/h0048255>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. <https://doi.org/10.1037/h0046016>
- Castañeda, M. B., Levin, J. R., & Dunham, R. B. (1993). Using planned comparisons in management research: A case for the Bonferroni procedure. *Journal of Management*, *19*(3), 707-724. [https://doi.org/10.1016/0149-2063\(93\)90012-C](https://doi.org/10.1016/0149-2063(93)90012-C)
- Chang, H., & Cartwright, N. (2008). Measurement. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (pp. 367-375). Routledge.
- Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*(2), 206-226. <https://doi.org/10.1037/1082-989X.10.2.206>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651-682. <https://doi.org/10.1177/1094428116656239>
- Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207-230. <https://doi.org/10.1177/1094428114555994>
- Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511. <https://doi.org/10.1037/pspp0000102>
- De Vries, T. A., Walter, F., Van Der Vegt, G. S., & Essens, P. J. M. D. (2014). Antecedents of individuals' interteam coordination: Broad functional experiences as a mixed blessing. *Academy of Management Journal*, *57*(5), 1334-1359. <https://doi.org/10.5465/amj.2012.0360>
- Edwards, J. R. (2003). Construct validation in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: The state of the science* (p. 327-371). Lawrence Erlbaum.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research*, *63*(3), 324-327. <https://doi.org/10.1016/j.jbusres.2009.05.003>
- Fornell, C., & Larcker, D. F. (1981a). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50. <https://doi.org/10.2307/3151312>
- Fornell, C., & Larcker, D. F. (1981b). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, *18*(3), 382-388. <https://doi.org/10.2307/3150980>
- Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*(3), 465-486. <https://doi.org/10.1177/1094428116689708>
- Gefen, D., & Straub, D. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems*, *16*(5), 91-109.
- Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, *6*(3), 258-269. <https://doi.org/doi:10.1037/1082-989X.6.3.258>

- 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. <https://doi.org/10.1037/0022-3514.64.6.1029>
- Green, J. P., Tonidandel, S., & Cortina, J. M. (2016). Getting through the gate: Statistical and methodological issues raised in the reviewing process. *Organizational Research Methods, 19*(3), 402-432. <https://doi.org/10.1177/1094428116631417>
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427-438. <https://doi.org/10.1177/001316444600600401>
- Hamann, P. M., Schiemann, F., Bellora, L., & Guenther, T. W. (2013). Exploring the dimensions of organizational performance: A construct validity study. *Organizational Research Methods, 16*(1), 67-87. <https://doi.org/10.1177/1094428112470007>
- Hancock, G. R., & Klockars, A. J. (1996). The quest for α : Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*(3), 269-306. <https://doi.org/10.2307/1170524>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115-135. <https://doi.org/10.1007/s11747-014-0403-8>
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. <https://doi.org/10.1177/0013164405282485>
- Hu, J., & Liden, R. C. (2015). Making a difference in the teamwork: Linking team prosocial motivation to team processes and effectiveness. *Academy of Management Journal, 58*(4), 1102-1127. <https://doi.org/10.5465/amj.2012.1142>
- Jeon, M., & Rijmen, F. (2014). Recent developments in maximum likelihood estimation of MTMM models for categorical data. *Frontiers in Psychology, 5*. <https://doi.org/10.3389/fpsyg.2014.00269>
- John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339-369). Cambridge University Press.
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). *semTools: Useful tools for structural equation modeling*. <https://CRAN.R-project.org/package=semTools>
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12*(3), 247-252. [https://doi.org/10.1016/0022-1031\(76\)90055-X](https://doi.org/10.1016/0022-1031(76)90055-X)
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford.
- Krause, M. S. (2012). Measurement validity is fundamentally a matter of definition, not correlation. *Review of General Psychology, 16*(4), 391-400. <https://doi.org/10.1037/a0027701>
- Krause, R., Whitler, K. A., & Semadeni, M. (2014). Power to the principals! An experimental look at shareholder say-on-pay voting. *Academy of Management Journal, 57*(1), 94-115. <https://doi.org/10.5465/amj.2012.0035>
- Kuppelwieser, V. G., Putinas, A.-C., & Bastounis, M. (2019). Toward application and testing of measurement scales and an example. *Sociological Methods & Research, 48*(2), 326-349. <https://doi.org/10.1177/0049124117701486>
- Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*(2), 112-125. <https://doi.org/10.1016/j.obhdp.2010.02.003>
- Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12*(1), 165-200. <https://doi.org/10.1177/1094428107302900>

- Lucas, R. E., Diener, E., & Suh, E. (1996). Discriminant validity of well-being measures. *Journal of Personality and Social Psychology, 71*(3), 616-628. <https://doi.org/10.1037/0022-3514.71.3.616>
- Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology, 31*(1), 32-42. <https://doi.org/10.1016/j.newideapsych.2011.02.006>
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>
- Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology, 76*(1), 127-133. <https://doi.org/10.1037/0021-9010.76.1.127>
- McDonald, R. P. (1985). *Factor analysis and related methods*. Lawrence Erlbaum.
- McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods, 16*(1), 152-184. <https://doi.org/10.1177/1094428112459910>
- McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. <https://doi.org/10.1037/met0000144>
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592. <https://doi.org/10.1037/0021-9010.93.3.568>
- Morin, A. J. S., Boudrias, J.-S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2017). Complementary variable- and person-centered approaches to the dimensionality of psychometric constructs: Application to psychological wellbeing at work. *Journal of Business and Psychology, 32*(4), 395-419. <https://doi.org/10.1007/s10869-016-9448-7>
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63-75. <https://doi.org/10.1177/0013164496056001004>
- Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology, 3*(102). <https://doi.org/10.3389/fpsyg.2012.00102>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 409-428. <https://doi.org/10.1080/10705511.2013.797820>
- Padilla, M. A., & Veprinsky, A. (2012). Correlation attenuation due to measurement error: A new approach using the bootstrap procedure. *Educational and Psychological Measurement, 72*(5), 827-846. <https://doi.org/10.1177/0013164412443963>
- Padilla, M. A., & Veprinsky, A. (2014). Bootstrapped deattenuated correlation nonnormal distributions. *Educational and Psychological Measurement, 74*, 823-830. <https://doi.org/10.1177/0013164414531780>
- Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*(7139), 1236-1238. <https://doi.org/10.1136/bmj.316.7139.1236>
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods, 19*(2), 159-203. <https://doi.org/10.1177/1094428115624965>
- Reichardt, C. S., & Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivariate Behavioral Research, 30*(4), 513-538. https://doi.org/10.1207/s15327906mbr3004_3
- Reise, S. P., & Revicki, D. A. (2014). *Handbook of item response theory modeling*. Taylor & Francis.
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. <https://doi.org/10.1037/met0000045>

- Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods, 16*(3), 425-448. <https://doi.org/10.1177/1094428112474693>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 226-238). Taylor & Francis. <https://doi.org/10.4324/9780429273872-19>
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. <https://doi.org/10.1037/1082-989X.8.2.206>
- Schmitt, N. (1978). Path analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 2*, 157-173. <https://doi.org/10.1177/014662167800200201>
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. <https://doi.org/10.1037/1040-3590.8.4.350>
- Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*(1), 1-22. <https://doi.org/10.1177/014662168601000101>
- Schwab, D. P. (2013). *Research methods for organizational studies*. Psychology Press.
- Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*(1), 80-110. <https://doi.org/10.1177/1094428115598239>
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*(1), 561-584. <https://doi.org/10.1146/annurev.ps.46.020195.003021>
- Spector, P. E. (2013). Survey design and measure development. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (pp. 170-188). Oxford University Press.
- Straub, D., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13*(1). <https://doi.org/10.17705/1CAIS.01324>
- Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25. <https://doi.org/10.1146/annurev.clinpsy.032408.153639>. Construct
- Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science, 1*(3), 375-388. <https://doi.org/10.1177/2515245918775707>
- Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement, 57*(1), 5-19. <https://doi.org/10.1177/0013164497057001001>
- Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-208. <https://doi.org/10.1177/0013164496056002001>
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29. <https://doi.org/10.1037/h0071663>
- Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44*(1), 119-134. <https://doi.org/10.1007/s11747-015-0455-4>
- Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin, 74*(3), 193-212. <https://doi.org/10.1037/h0029778>
- Wetecher-Hendricks, D. (2006). Adjustments to the correction for attenuation. *Psychological Methods, 11*(2), 207-215. <http://dx.doi.org/10.1037/1082-989X.11.2.207>
- Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait-multimethod data clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods, 15*(1), 134-161. <https://doi.org/10.1177/1094428111408616>

- Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.
- Zijlmans, E. A. O., van der Ark, L. A., Tijmstra, J., & Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied Psychological Measurement, 42*(7), 553-570. <https://doi.org/10.1177/0146621618758290>
- Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement, 67*(6), 920-939. <https://doi.org/10.1177/0013164406299132>
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. <https://doi.org/10.1007/s11336-003-0974-7>

Author Biographies

Mikko Rönkkö is associate professor of entrepreneurship at Jyväskylä University School of Business and Economics (JSBE) and a docent at Aalto University School of Science. He is a department editor for *Journal of Operations Management* handling methodological articles and on the editorial boards of *Organizational Research Methods* and *Entrepreneurship Theory and Practice*. His methodological research addresses quantitative research methods broadly. He runs a research methods-focused YouTube channel at <https://www.youtube.com/mronkko>.

Eunseong Cho is a professor of marketing in the College of Business Administration, Kwangwoon University, Republic of Korea. He earned his PhD from the Korea Advanced Institute of Science and Technology in 2004. With methodological research focusing on reliability and validity, he is the awardee of the 2015 Organizational Research Methods Best Paper Award.