Improving the discriminant validation of multi-item scales

DOI: 10.1177/00222437251388994

JOURNAL of MARKETING RESEARCH

Improving the Discriminant Validation of Multi-Item Scales

Journal:	*Journal of Marketing Research*
Manuscript ID	JMR-24-0273.R2
Manuscript Type:	Revised Submission
Topics and Methods:	Research and analysis methods < Topics, Surveys < Methods

Improving the Discriminant Validation of Multi-Item Scales

Constant Pieters[a,1]

Hans Baumgartner[b]

Rik Pieters[c]

[a] Assistant Professor of Marketing, Department of Marketing, Copenhagen Business School, Solbjerg Plads 3, 2000 Frederiksberg, Denmark. Tel. +4538152127. E-mail: cpi.marktg@cbs.dk.

[b] Smeal Chair Professor of Marketing Emeritus, Smeal College of Business, The Pennsylvania State University, 482 Business Building, 16802 University Park, Pennsylvania, United States of America. Tel. 814-863-3559. E-mail: jxb14@psu.edu.

[c] Emeritus Professor of Marketing, Department of Marketing, Tilburg School of Economics and Management, Tilburg University, Warandelaan 2, PO Box 90153 5000 LE Tilburg, The Netherlands; and Adjunct Professor of Marketing, Department of Marketing and Consumer Behavior, Católica Lisbon, Palma de Cima, 1649-023 Lisbon, Portugal. Tel. +31134663022. E-mail: f.g.m.pieters@tilburguniversity.edu.

[1] Corresponding author.

Acknowledgements: The authors are grateful to the *JMR* review team for their thoughtful and constructive comments, which substantially improved the paper. The authors also thank the colleagues from the Department of Marketing at Copenhagen Business School, participants of the 2024 Association of Consumer Research annual conference, and participants of the 2025 European Marketing Academy annual conference for useful suggestions and feedback.

Data availability statement: The data that support the findings of this article are publicly available on *OSF* (https://osf.io/57srv/).

Improving the Discriminant Validation of Multi-Item Scales

Abstract

Discriminant validation examines to what extent constructs measured with multi-item scales, which are hypothesized to be conceptually distinct, are empirically distinct. A literature review of published scale development studies shows that a variety of criteria and approaches to assess discriminant validity are in use. However, the requirements for an appropriate criterion have not been spelled out, which has led to the use of problematic criteria. The present research introduces three requirements that an appropriate discriminant validation criterion should satisfy, concerning the correlation, comparison standard, and comparison method. It shows that the common Fornell and Larcker criterion is based on an inappropriate comparison standard and method and that alternative criteria have weaknesses as well. The authors therefore propose an improved comparison standard, congeneric reliability (CR), and develop a systematic discriminant validation procedure based on CR and an existing criterion (Phi), both of which satisfy the three requirements. The procedure provides continuous measures of support for discriminant validity and accounts for measurement and sampling error. A detailed case study and reanalyses of seven published scale development articles demonstrate the application and strengths of the procedure. Example code and an online application facilitate its implementation.

Keywords: scale development, discriminant validity, congeneric reliability, Fornell-Larcker criterion

Discriminant validation examines to what extent constructs measured with multi-item scales, which are hypothesized to be conceptually distinct, are empirically distinct. Discriminant validation is crucial for both the development of new scales and the deployment of existing ones (Churchill 1979; Gerbing and Anderson 1988; Haws et al. 2023; MacKenzie et al. 2011). Constructs should be empirically distinct from conceptually related constructs, and the same holds for the dimensions of multidimensional constructs. Lack of support for discriminant validity encourages the proliferation of semantically distinct but empirically indistinguishable constructs (Shaffer et al. 2016). Such conceptual fragmentation hampers theoretical progress, complicates syntheses of the literature, and may give the illusion of theoretical richness where conceptual redundancy prevails (Albert and Thomson 2023; Morrow 1983). It also reduces the effectiveness of managerial interventions that are based on the assumed distinctiveness of constructs and their measures. Establishing discriminant validity between similar but distinct constructs therefore enables developing and testing meaningful theories and practical interventions based on these theories.

Various criteria to assess discriminant validity have been proposed and are being used (Fornell and Larcker 1981; Franke and Sarstedt 2019; Henseler et al. 2015; Rönkkö and Cho 2022; Voorhees et al. 2016). However, current practice faces two challenges. First, the requirements that an appropriate discriminant validation criterion should meet have not been spelled out explicitly. This has resulted in a plethora of criteria (Rönkkö and Cho 2022) and conflicting recommendations. For instance, Voorhees et al. (2016, p. 133) conclude that the widely used criterion proposed by Fornell and Larcker (1981) "should be the standard for discriminant validity testing." In stark contrast, Rönkkö and Cho (2022, p. 27) advocate against the use of this criterion because it has "a very high false positive rate" (i.e., it flags discriminant validity problems when none are present). Second, it is unclear how appropriate individual criteria can be combined to comprehensively assess the discriminant validity of

constructs. Recently, Rönkkö and Cho (2022) argued that discriminant validation procedures should refrain from drawing "all or nothing" conclusions and instead assess the degree of distinctiveness of constructs. A systematic discriminant validation procedure incorporating multiple criteria enables this but is currently unavailable.

The present research addresses these challenges as follows. First, based on the measurement and scale development literatures (Fornell and Larcker 1981; Lord and Novick 1968; Rönkkö and Cho 2022), we identify three requirements that an appropriate discriminant validation criterion should satisfy. Specifically, a conceptually valid criterion should to use an appropriate correlation, comparison standard, and method of comparison.

Second, we document the use of discriminant validation criteria in a review of 79 recent scale development articles published between 2000 and 2024, which updates earlier reviews in marketing (Voorhees et al. 2016) and management (Shaffer et al. 2016). It shows that commonly used and newly proposed discriminant validation criteria fail one or more of the three requirements and therefore cannot be recommended for discriminant validation. This is particularly the case for the Fornell and Larcker (1981) criterion. Although this criterion is routinely recommended (Albert and Thomson 2023; Farrell 2010; Pieters 2017; Voorhees et al. 2016) and the article in which it was proposed is the most highly cited article in the *Journal of Marketing Research* (having garnered over 70,000 cumulative citations according to the *Web of Science* as of this writing), we will demonstrate that the criterion is inappropriate for discriminant validation of multi-items scales and should no longer be used.

Third, we propose a new two-step procedure for discriminant validation that integrates two criteria, the existing Phi criterion and a new congeneric reliability (CR) criterion, both of which meet all three requirements. The proposed procedure offers a graded approach to discriminant validation and moves beyond binary "yes" or "no" conclusions. Annotated *R*-code and an online *Shiny* application (app) implement the proposed procedure.

Fourth, we demonstrate empirically that using the CR criterion instead of the Fornell and Larcker (1981) criterion matters in practice. We do so with a detailed case analysis and by reanalyzing seven articles that report discriminant validity issues. The results demonstrate that the proposed procedure can lead to conclusions about discriminant validity that differ from and are more informative than those reached by currently used approaches.

The following section presents the three requirements for an appropriate discriminant validation criterion. We then assess the extent to which current criteria satisfy the three requirements. Next, we describe the new discriminant validation criterion and a procedure that incorporates the new criterion and an existing one. Next, we report the empirical applications. The final section offers recommendations. Supplemental material, including data and *R*-code, is available on *OSF* (https://osf.io/57srv/). The *Shiny* app can be found at https://constantpieters.shinyapps.io/discriminantvalidation/.

Three Requirements for Discriminant Validation Based on the Confirmatory Factor Model

Discriminant validation assesses to what extent constructs are empirically distinct to a sufficient degree (Campbell and Fiske 1959; Fornell and Larcker 1981; Jöreskog 1971; Lord 1957; Rönkkö and Cho 2022). Both direct and indirect approaches have been proposed for this purpose. Direct approaches examine whether the correlation between two constructs is small enough relative to some comparison standard (Campbell and Fiske 1959). Various indirect approaches complement the direct approaches; they examine, for example, whether a construct C relates differently to constructs A and B, or whether C moderates the association between A and B. Such findings support the distinctiveness of A and B (Franke et al. 2021; Tesser and Krauss 1976). Because the direct approach is more common, and the indirect approach does not provide direct evidence about the discriminant validity between A and B

and requires information about C, our research focuses on direct discriminant validation.

Below, we describe three requirements that a discriminant validation criterion under the direct approach should satisfy. Because direct discriminant validation entails comparing a correlation with a comparison standard, the three requirements (which are jointly sufficient) focus on the correlation, the comparison standard, and the comparison method.

To guide the analysis, Figure 1 presents a confirmatory factor model with hypothetical population-level data as an illustration. It contains two correlated constructs, each represented by a factor and measured with multi-item scales. Both factors are measured with three items: F_1 with $x_1 - x_3$ and F_2 with $x_4 - x_6$. The model readily extends to situations with additional factors and more or varying numbers of items for each factor. For all items x_i, the loadings λ_{i1} or λ_{i2} quantify how well an item measures F_1 or F_2, respectively. In the example, the loadings of $x_1 - x_3$ on F_1 are .55, .75, and .95, respectively, and the loadings of $x_4 - x_6$ on F_2 are all .70. For ease of exposition, the observed items are assumed to be standardized so that random measurement error ε_i has variance $\theta_{ii} = 1 - \lambda_{ij}^2$. The factor variances, VAR_{F_1} and VAR_{F_2}, are fixed to 1 for model identification. This means

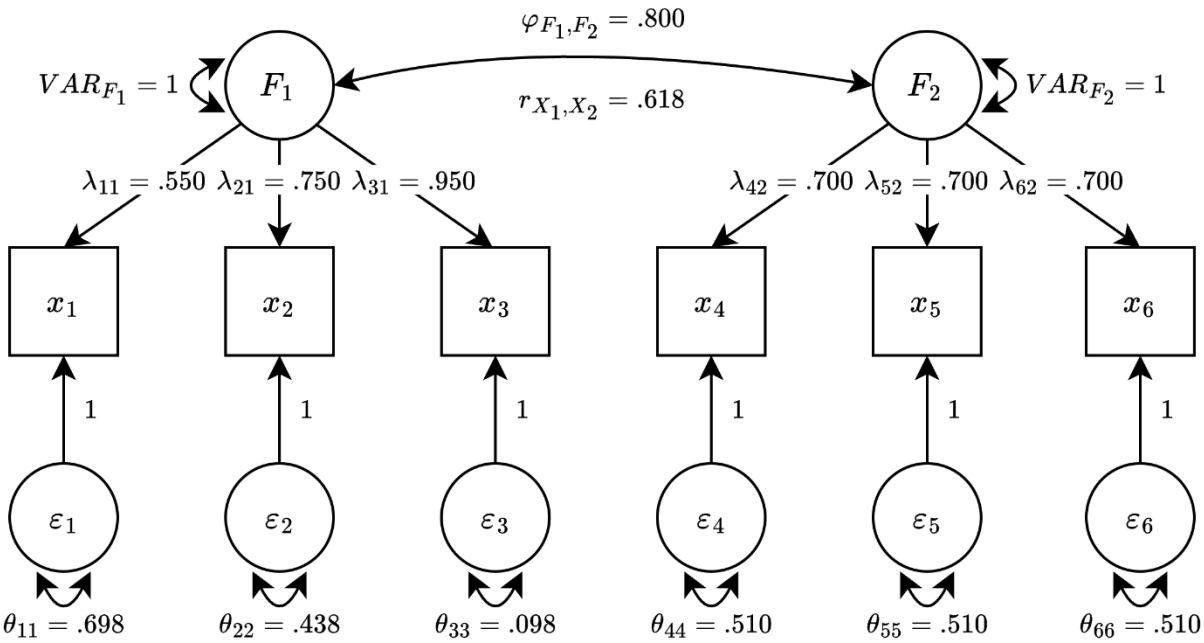


Figure 1. Illustrative Discriminant Validation Model.
Note: Factor variances are fixed to 1 for model identification.

that the factor covariance, φ_{F_1,F_2}, is a factor correlation, which is φ_{F_1,F_2} = .80 in Figure 1. Without loss of generality, the mean structure is ignored, and positive loadings and correlations, as well as a correctly specified and well-fitting model, are assumed.

Three classic measurement models have been distinguished in the psychometric literature (Lord and Novick 1968). The *parallel* measurement model, which is most restrictive, assumes that all items of a given factor have the same loadings and error variances. The (essentially) *tau-equivalent* measurement model, which is less restrictive, allows the error variances of the items to vary, but it assumes that the loadings of all items measuring a given factor are the same. The *congeneric* measurement model, which is the most general and flexible model, allows both the loadings and the error variances of the items to vary. Applying these concepts to the example in Figure 1, the three items measuring factor F_2 comprise a parallel measurement model: all three loadings are .70 and all three error variances are .51. The three items measuring F_1 comprise a congeneric measurement model because both the loadings and error variances vary. This example also shows that with standardized items, the parallel and tau-equivalent models are the same because identical loadings across items imply identical error variances.

Requirement 1: Use a Factor Correlation Derived from a Well-Fitting Confirmatory Factor Model

An appropriate correlation estimates the true correlation between constructs by accounting for measurement error. A correctly specified and well-fitting confirmatory factor analysis (CFA) model explicitly accounts for measurement error, making the resulting factor correlation (φ, Phi) an appropriate correlation measure. In contrast, correlations between observed scale scores are biased because they do not sufficiently account for measurement error and should therefore be avoided in discriminant validation.

There are two sources of measurement error. First, random measurement error

reduces the reliability of multi-item measures and thereby attenuates the correlation between scale scores. This exaggerates discriminant validity because the bivariate correlation between observed scale scores is almost always smaller than the true correlation and never larger. For two constructs (subscripted 1 and 2), the relationship between the true correlation ρ_{T_1,T_2} and the observed correlation r_{X_1,X_2} is (Spearman 1904, p. 90):

$$\rho_{T_1,T_2} = \frac{r_{X_1,X_2}}{\sqrt{\rho^2_{X_1,T_1} \times \rho^2_{X_2,T_2}}}, \quad (1)$$

where X_1 and X_2 are the observed multi-item scale scores of the two constructs, T_1 and T_2 are the true scores of the two constructs, and $\rho^2_{X_1,T_1}$ and $\rho^2_{X_2,T_2}$ are the multi-item scale reliabilities of X_1 and X_2, respectively. Unless the scale reliabilities are 1, in which case the scale scores are identical to true scores, the observed correlation underestimates the true correlation. In the context of the example in Figure 1, a CFA correctly estimates the true correlation of .80 by accounting for random measurement error in the items, whereas the observed correlation of .618 is a downward-biased (attenuated) estimate of the factor correlation. Therefore, the correlation between observed scale scores overestimates discriminant validity.

Second, systematic measurement error can either attenuate or inflate the correlation between the observed scores of multi-items measures. For example, individual differences in stylistic responding (e.g., acquiescence) or properties of items such as their keying direction can bias the correlation upward or downward (Baumgartner and Weijters 2022). Also, items that primarily capture one construct may cross-load on another construct, which can inflate the factor correlation when the cross-loadings are restricted to 0 (Asparouhov and Muthén 2009). A correctly specified CFA and well-fitting factor model provide evidence that the estimated factor correlation is unconfounded by systematic measurement error. Ideally, the chi-square value that assesses overall model (mis)fit will be sufficiently low, but when the sample size is relatively large, even relatively minor misspecifications, such as small non-

target loadings or small error covariances, can produce large chi-square values. Alternative fit indices such as RMSEA, CFI, or TLI may then be used to assess whether the fit of the model is good enough from a practical perspective (see Baumgartner and Weijters 2022, for details, including robust test statistics and flexible cutoff values for alternative fit indices). In the example of Figure 1 systematic error is absent and the fit of the model is perfect.

There are several alternative approaches besides CFA to account for measurement error, which are not recommended. First, the correlation between observed scale scores can be corrected for measurement unreliability post hoc via a model-free aproach using Equation 1. For the example of Figure 1, using the congeneric reliabilities of .804 for F_1 and .742 for F_2 to correct the observed correlation of .618 recovers the true correlation of .80 perfectly because the model fit is perfect and appropriate reliability estimates are used (see below). However, such post hoc corrections of the observed correlation often rely on simplifying assumptions and therefore may not correctly estimate the true correlation. For instance, using Cronbach's coefficient alpha assumes that the items are at least tau-equivalent and underestimates reliability if this assumption is violated. An underestimated reliability then overcorrects the observed correlation. Cronbach's coefficient alpha is .785 for F_1 and .742 for F_2 in the example. Coefficient alpha and congeneric reliability are identical for F_2 because tau-equivalence holds in the present case. However, coefficient alpha underestimates the true (congeneric) reliability of F_1 (.785 vs. .804) and therefore (slightly) overestimates the true correlation (.81 vs. .80). In practical applications, the bias could be larger, but the presence and magnitude of the bias is unknown without an explicit measurement model as in CFA. Moreover, model-free approaches to estimating the factor correlation, such as post hoc corrections, cannot assess the presence and effect of systematic measurement error (Rönkkö and Cho 2022).

Second, although exploratory factor analysis (EFA) can be used to estimate the factor

correlation or ascertain whether the hypothesized number of factors can be recovered, it is inappropriate for discriminant validation. EFA does not fully account for measurement error and thus produces biased correlation estimates (Bollen 1989). Moreover, EFA with orthogonal factors, such as when using the common varimax rotation, is uninformative for discriminant validation because it imposes uncorrelated factors a priori, even though the factor correlation is the metric of interest. Recovering the hypothesized number of factors in an EFA or not-finding substantial cross-loadings of items on non-target factors is also inappropriate because it does not assess discriminant validity but instead explores the dimensionality of constructs measured with multiple items (Gerbing and Anderson 1988).

Requirement 2: Compare the Factor Correlation to a Principled Maximum Correlation

The comparison standard is pivotal in discriminant validation and widely debated (Franke and Sarstedt 2019; Henseler et al. 2015; Rönkkö and Cho 2022; Voorhees et al. 2016). An overly lenient standard risks false claims of discriminant validity, whereas an overly strict standard risks false rejections of discriminant validity.

We distinguish three categories of comparison standards: (1) a perfect correlation of 1, (2) correlations smaller than 1, and (3) scale reliability. First, using a perfect correlation of 1 as a comparison standard ensures that conceptually distinct constructs are not perfectly correlated (Lord 1957). This is an appropriate standard because two perfectly correlated constructs (after correction for attenuation) cannot be distinct. However, since it is relatively easy to establish that two factors are less than perfectly correlated (as explained in greater detail below), more stringent standards have been proposed.

A second category of comparison standards uses cutoffs that are smaller than a perfect correlation of 1. These comparison standards are arbitrary, based on subjective judgments rather than theoretical or logical arguments, which may encourage researchers to selectively choose standards based on the obtained outcome. Among others, the following comparison

standards have been proposed: .70 (Haws et al. 2023), .71 (MacKenzie et al. 2011), .85 (Henseler et al. 2015; Voorhees et al. 2016), and .95 (Bagozzi and Yi 1988). Based on a review of comparison standards used in practice, Rönkkö and Cho (2022) proposed the following classification: a correlation below .80 indicates no problem with discriminant validity; .80–.90 is a marginal issue; above .90 but below 1 is a moderate problem; and a correlation not statistically different from 1 constitutes a severe problem. Henseler et al. (2015, p. 129) concluded that the choice of a comparison standard depended on "how conservative the researcher is in his or her assessment of discriminant validity". Bagozzi and Yi (1988, p. 77) pointed out that "[w]hat constitutes 'too high' a correlation is a somewhat arbitrary issue". It is apparent that consensus on what constitutes a defensible comparison standard that is smaller than 1 remains elusive.

Some prior work has relied on simulations to recommend comparison standards smaller than a perfect correlation (Franke and Sarstedt 2019; Voorhees et al. 2016). However, since the true population correlation for discriminant validity is unknown (Rönkkö and Cho 2022, p. 34), such simulations use arbitrary standards as inputs to judge the outputs. To illustrate, Voorhees et al. (2016, p. 130, Table 4) conducted a simulation study assuming that a population correlation of .75 between factors was discriminant-valid whereas a correlation of .90 was not. Based on a comparison standard of .85 and a sample size of n = 500, all 200 replications in the simulation flagged discriminant validity issues at the larger correlation of .90 and only 1 replication flagged issues at the smaller correlation of .75. This led to the recommendation that .85 should be used as a comparison standard. However, the assumptions about problematic and unproblematic population correlations predetermined the conclusions derived from the simulation. For example, if a correlation of .70 had been deemed problematic (following Haws et al. 2023), a population correlation of .75 would have violated discriminant validity in many replications; and if .95 had been used as the cutoff (Bagozzi

and Yi 1988), a population correlation of .90 would have confirmed discriminant validity in at least some of the replications. In sum, fixed comparison standards less than 1 are problematic and simulation evidence based on judgmental input is futile for choosing a comparison standard.

Third, comparison standards can be based on the reliability of multi-item measures. A reliability-based comparison standard is based on the notion that a construct should have more in common with its own multi-item measure than with another, conceptually related but supposedly distinct construct. Classical Test Theory specifies that the reliability of a multi-item measure is the shared variance between a construct and its measures (Lord and Novick 1968), and the shared variance between two constructs is equal to the squared factor correlation. Thus, using a reliability-based comparison standard requires that the squared factor correlation be smaller than the reliabilities of the multi-item measures of the two constructs. Equivalently, the factor correlation should be smaller than the square roots of the reliabilities of the two multi-item measures. Besides being intuitive, an advantage of a reliability-based comparison standard is that it is estimated from the data and not determined subjectively. A reliability-based comparison standard is also smaller than 1, except for boundary cases of perfect scale reliability, and thus more stringent than a comparison standard of perfect correlation. Fornell and Larcker (1981) can be credited with first proposing a reliability-based comparison standard, but unfortunately their measure of reliability is inappropriate for discriminant validation, as shown below.

Requirement 3: Provide a Continuous Measure of Support for Discriminant Validity and Account for Sampling Error

We propose the discriminant validity index (DVI) as a simple and continuous measure of the extent of support for discriminant validity. It is defined as the comparison standard (CS) minus the (assumed to be positive) factor correlation φ:

$$DVI = CS - \varphi. \quad (2)$$

DVI is theoretically bounded between −1 and +1. More positive values provide greater evidence in support of discriminant validity. Negative values point to a violation of discriminant validity because the factor correlation exceeds the standard. Numerical or statistical comparisons can be used to assess whether the DVI is sufficiently positive.

The DVI in Equation 2 holds at the population level. At the sample level, the estimated support for discriminant validity may differ from the population value due to sampling error e: $DVI_{sample} = DVI_{population} + e$. Even when the DVI is 0 in the population, sample values provide directional support for discriminant validity in about 50% of cases (assuming a symmetric distribution of the errors). Researchers commonly use the algebraic difference between CS and φ to evaluate discriminant validity (Campbell and Fiske 1959; Fornell and Larcker 1981), but this ignores sampling error. When sampling error is ignored, a .84 factor correlation would be judged to be smaller than a comparison standard of .85, even though there is little support for discriminant validity in this case.

Statistical comparisons account for sampling error and are therefore recommended (Franke and Sarstedt 2019; Pieters 2017; Rönkkö and Cho 2022). They provide a *P*-value of the extent of incompatibility of the data with the model as well as a confidence interval (CI) to indicate the uncertainty associated with the point estimate of DVI. Statistical tests can be conducted using a Wald test (i.e., an estimate of DVI divided by an estimate of its standard error based on the multivariate delta method) or a likelihood ratio (chi-square difference) test (Bollen 1989; Cheung and Pesigan 2023; Jöreskog 1971), and corresponding confidence intervals are available for both tests. If the sample size is small and/or the normality assumption is questionable, resampling approaches such as the bootstrap yield more accurate confidence intervals (Efron and Tibshirani 1994).

Because discriminant validity is a matter of degree, researchers can quantify the

evidence supporting the discriminant validity of constructs by reporting DVIs, confidence intervals, and precise *P*-values, and interpreting these statistics continuously. To enable continuous and robust inference, we report DVIs with 95% confidence intervals based on percentile bootstrapping with 10,000 resamples. The *R* code on *OSF* allows setting alternative confidence levels. We also report exact (up to .001) one-sided *P*-values of statistical tests of DVI = 0 based on the proportion of bootstrap resamples with a negative DVI (or the proportion of positive DVIs if the estimate is negative). Results based on Wald or likelihood ratio tests are an option in the *OSF* code.

In sum, from the discussion of the three requirements it follows that an appropriate discriminant validation criterion should use a factor correlation derived from a well-fitting CFA model; compare the factor correlation to a principled maximum (a perfect correlation or an appropriate measure of scale reliability); and provide a continuous measure of support for discriminant validity that accounts for sampling error.

Evaluating Current Discriminant Validation Criteria

This section examines the extent to which current discriminant validation criteria satisfy or violate the three requirements. Panel A in Table 1 summarizes the findings. The example in Figure 1 is used to demonstrate each criterion. We also report the results of a literature review of 79 scale development articles published in the 2000-2024 volumes of the *Journal of Consumer Psychology (JCP)*, *Journal of Consumer Research (JCR), Journal of Marketing (JM)*, and *Journal of Marketing Research (JMR)* to illustrate current discriminant validation and reporting practices (Web Appendix A has details).

Phi Criterion

The Phi criterion is dominant in marketing and management (Franke and Sarstedt 2019; Voorhees et al. 2016). It was used by 39% of the studies in an earlier review of 621 survey-

Table 1. Analysis of Discriminant Validation Criteria.

Criterion	# (%) of articles	Summary of criterion: Support for discriminant validity or discriminant validity index (DVI)	Summary of criterion: Comparison of correlation with comparison standard	Requirements satisfied by criterion: Requirement 1: Uses a correlation derived from a well-fitting CFA that accounts for measurement error	Requirements satisfied by criterion: Requirement 2: Compares the correlation to a principled maximum	Requirements satisfied by criterion: Requirement 3: Provides a continuous measure of support for discriminant validity and accounts for sampling error
Panel A: Discriminant validation criteria used in the literature						
Phi criterion	47 (59%)	$\varphi_{F_1,F_2} < 1$	Conduct a likelihood ratio test of H_0:$\varphi_{F_1,F_2} = 1$ vs. H_1:$\varphi_{F_1,F_2} < 1$ to compare a model with freely estimated factor correlation with a restricted model that fixes the factor correlation to 1 or, equivalently, a single-factor model that combines F_1 and F_2 (Jöreskog 1971). Alternatively, estimate a Wald confidence interval around φ (Anderson and Gerbing 1988).	Yes	Yes	Yes
Heterotrait-monotrait (HTMT) criterion	4 (5%)	$HTMT < .85$	Estimate a Wald (Franke and Sarstedt 2019) or bootstrapped (Henseler et al. 2015) confidence interval around $HTMT$.	No, HTMT is a potentially biased (model-free) estimate of the factor correlation	No, uses an arbitrary value as a comparison standard	No, the 4 reviewed articles reporting the criterion ignore sampling error
Fornell-Larcker (FL) criterion	43 (54%)	$AVE > \varphi^2_{F_1,F_2}$	Usually, a numerical comparison is used, although a Wald confidence interval around $AVE - \varphi^2_{F_1,F_2}$ could be estimated (Franke and Sarstedt 2019; Pieters 2017).	Yes	No, uses an inappropriate reliability measure as a comparison standard	No, the 43 reviewed articles reporting the criterion ignore sampling error
Panel B: Discriminant validation criterion proposed in this paper						
Congeneric reliability (CR) criterion	0 (0%)	$DVI_{CR_{F_1}} > 0$ and $DVI_{CR_{F_2}} > 0$, where $DVI_{CR_{F_1}} = \sqrt{CR_{F_1}} - \varphi_{F_1,F_2}$ and $DVI_{CR_{F_2}} = \sqrt{CR_{F_2}} - \varphi_{F_1,F_2}$	Assess both H_0:$\varphi_{F_1,F_2} = \sqrt{CR_{F_1}}$ vs. H_1:$\varphi_{F_1,F_2} < \sqrt{CR_{F_1}}$ and H_0:$\varphi_{F_1,F_2} = \sqrt{CR_{F_2}}$ vs. H_1: $\varphi_{F_1,F_2} < \sqrt{CR_{F_2}}$ by using continuous P-values and/or constructing (bootstrapped) confidence intervals around $DVI_{CR_{F_1}}$ and $DVI_{CR_{F_2}}$.	Yes	Yes	Yes

Notes: Panel A focuses on discriminant validation in the existing literature and Panel B on the proposal of this paper. Data in the second column are based on a review of 79 scale development articles that reported a discriminant validation, detailed in Web Appendix A. Percentages (within parentheses) do not sum to 100% due to rounding and articles reporting multiple discriminant validation criteria (e.g., 26 of the 79 articles [33%] reported both the Phi and FL criteria). Notation is on the population level and replaced by sample estimates in practice, and φ_{F_1,F_2} refers to the factor correlation.

based articles in seven major marketing journals between 1996 and 2012 (Voorhees et al. 2016) and by 47 (59%) of the 79 articles in our review. The correctly implemented Phi criterion meets all three requirements (Table 1): (1) the factor correlation based on a well-fitting CFA model accounts for measurement error; (2) the factor correlation is compared with an appropriate (perfect correlation) standard ($DVI_1 = 1 - \varphi$, with the subscript 1 referring to a perfect correlation of 1); and (3) the comparison accounts for sampling error because a Wald confidence interval is constructed around DVI_1 (Anderson and Gerbing 1988) or a likelihood ratio test is used to compare the fit of a model with a freely estimated factor correlation to the fit of a restricted model that combines the two factors into a single factor or, equivalently, fixes the factor correlation to 1 (Jöreskog 1971). The more the upper limit of the confidence interval deviates from 1 or the greater the misfit of the restricted model relative to the unrestricted model, the greater the support for discriminant validity. The name "Phi criterion" is ingrained in the discipline (Voorhees et al. 2016) and we retain it for legacy reasons even though most criteria compare the factor correlation to some standard.

To illustrate, consider the example in Figure 1 at a large sample size of n = 1,000, about double the mean sample size in the literature review (M = 504, Median = 201, SD = 1,297, range = 34–12,263). The 95% CI around DVI_1 = 1 − .80 = .20 is [.160, .240] (p < .001). The lower limit of the interval is substantially larger than zero, and the p-value is very small, so there is convincing support for the Phi criterion. Setting the sample size to a moderate n = 500 yields a similar result: 95% CI [.143, .260] and p < .001. Even at a modest sample of n = 100, the lower limit of the 95% CI of DVI [.059, .355] exceeds zero (p = .002).

A drawback of the Phi criterion is its leniency because it can be met even when the correlation is very large. Based on follow-up simulations, the factor correlation would have to be larger than about .92 (DVI_1 = .08 with 95% CI [-.007, .189]), .96 (DVI_1 = .04 with 95% CI [-.004, .086]), or .98 (DVI_1 = .02 with 95% CI [-.006, .046]) for the 95% CI of DVI to

overlap zero at sample sizes of 100, 500, and 1,000, respectively. Web Appendix B demonstrates the leniency of the Phi criterion more systematically.

Heterotrait-Monotrait (HTMT) Criterion

The heterotrait-monotrait ratio (HTMT) was used by 4 studies (5%) in the literature review. The HTMT estimates the factor correlation from the ratio of the arithmetic mean inter-item correlation *between* pairs of indicators of two multi-item scales in the numerator and the geometric mean of the arithmetic mean inter-item correlation *within* each of the scales in the denominator (Henseler et al. 2015). The original HTMT assumes parallel items, which overestimates the factor correlation and thus underestimates the support for discriminant validity when this assumption is violated (Rönkkö and Cho 2022). An updated HTMT relaxes the parallel item assumption (Ringle et al. 2023):

$$HTMT = \frac{\overline{R}_{F_1,F_2}}{\sqrt{\overline{R}_{F_1} \times \overline{R}_{F_2}}}, \quad (3)$$

where $\overline{R}_{F_1,F_2}$ is the geometric mean correlation between pairs of F_1 and F_2 scale items, and $\overline{R}_{F_1}$ and $\overline{R}_{F_2}$ are the geometric mean item correlations within the F_1 and F_2 scales (where absolute values of the correlations should be used when there are correlations with different signs). The four articles in the literature review using the HTMT reported the original HTMT because they were published before the updated version was proposed.

The updated HTMT criterion still has four problems (Table 1). First, the model-free approach is unnecessary because it estimates the factor correlation indirectly, whereas a CFA automatically provides an estimated factor correlation. Second, the HTMT is a potentially biased estimate of the true correlation because it makes assumptions in line with the congeneric factor model and the model-free approach does not allow researchers to verify these assumptions or make adjustments, if necessary. This violates the first requirement. As a case in point, the updated HTMT is .80 for the example in Figure 1, identical to the factor

correlation estimated with CFA. The two correlations are identical because the assumptions required for using the HTMT approach happen to be satisfied. However, there is no way to assess whether systematic measurement error is present or whether remedial action to improve the fit of the model is necessary. Web Appendix C demonstrates that factor correlations estimated via the HTMT approach can be substantially biased in the presence of systematic error, whereas a CFA model would point to the need for respecifying the factor model to obtain trustworthy correlation estimates. Third, the HTMT criterion usually violates the second requirement that a discriminant validation criterion should have a principled maximum to which the factor correlation is compared. The literature recommends a perfect correlation of 1 or arbitrary standards of .85 or .90 (Franke and Sarstedt 2019; Henseler et al. 2015; Voorhees et al. 2016). All four articles in the review that reported the HTMT criterion used a .85 comparison standard. Fourth, it is recommended that a Wald test or bootstrapping be used for statistical inference when implementing the HTMT criterion (Franke and Sarstedt 2019; Henseler et al. 2015). However, all four articles in the review that reported the HTMT criterion used a numerical comparison without a statistical comparison, violating the third requirement.

The HTMT criterion was initially proposed for variance-based structural equation modeling (SEM), such as partial least squares or PLS (Henseler et al. 2015). Henseler et al. (2015) found based on simulations that the Fornell and Larcker (1981) criterion, discussed subsequently, performed very poorly in that context and proposed the HTMT criterion as an alternative. However, some researchers have misapplied the HTMT criterion to covariance-based SEM under the mistaken belief that it is a new and improved approach for assessing the discriminant validity of constructs in general. It should be clear that, based on the foregoing discussion, there is little reason to recommend the HTMT criterion in the context of the more common covariance-based SEM.

Fornell and Larcker (FL) Criterion

The Fornell and Larcker (FL) criterion (Fornell and Larcker 1981) is generally viewed as an essential part of an assessment of discriminant validity (Albert and Thomson 2023; Farrell 2010; Pieters 2017; Voorhees et al. 2016), and it is commonly reported in the literature. Forty-nine percent of 621 survey-based articles in seven major marketing journals between 1996 and 2012 used the FL criterion (Voorhees et al. 2016) and 43 of the 79 articles (54%) in the current review did. However, the FL criterion is inappropriate for discriminant validation and should no longer be used for this purpose, as explained next and summarized in Table 1.

The FL criterion specifies that the average variance extracted (AVE) by a construct from the items in a scale should be larger than the variance that the construct shares with other constructs. It relies on CFA to estimate the squared factor correlation as an estimate of the shared variance (SV) between factors, thus meeting the first requirement. However, it numerically compares AVE with the squared factor correlation (AVE > SV), thus failing the third requirement. Although it is straightforward to implement a statistical test to account for sampling error (Franke and Sarstedt 2019; Pieters 2017), none of the articles in the literature review reported a statistical test. The FL criterion also fails the second requirement, which is more problematic. Because the FL criterion occupies a central position in discriminant validation, we elaborate on this deficiency of the FL criterion in more detail.

First, AVE does not capture scale reliability but the reliability of a hypothetical "average" item, which is uninformative for discriminant validation. Specifically, the AVE for factor F_1 with J items, each with loading λ_{j1} and error-variance θ_{jj}, is specified as:

$$AVE_{F_1} = \frac{\sum_{j=1}^{J} \lambda_{j1}^2}{\sum_{j=1}^{J} \lambda_{j1}^2 + \sum_{j=1}^{J} \theta_{jj}}. \quad (4)$$

This expression can be rewritten as follows:

$$AVE_{F_1} = \sum_{j=1}^{J} IIR_j \times \frac{VAR[x_j]}{\sum_{j=1}^{J} VAR[x_j]}, \quad (5)$$

where $IIR_j = \lambda_{j1}^2/(\lambda_{j1}^2+\theta_{jj})$ is the individual-item reliability of the j^{th} item. IIR is the proportion of the total variance of an item in the denominator accounted for by the factor in the numerator. Equation 5 shows that AVE is a weighted average of the IIRs of the items measuring a construct, where the weights are the item variances ($VAR[x_j]$) as a proportion of the sum of all item variances. For standardized items (for which all variances are 1), AVE simplifies to the unweighted mean IIR. For parallel items, AVE is equal to the constant IIR. Based on this analysis, the FL criterion asks the following question: Does a factor account for more of the variance in a single, hypothetical "average" item than it shares with another factor that is presumed to measure a conceptually distinct construct? This is an irrelevant question for discriminant validation, which focuses on the multi-item scale, not on a single item.

Second, to demonstrate that AVE is not a measure of scale reliability and an inappropriate comparison standard for discriminant validation, consider again the example in Figure 1. The AVEs of F_1 and F_2 are .589 and .490, respectively, and the observed correlation between the multi-item scale scores is .618. Substituting these values into Equation 1 yields a corrected correlation of 1.15, which is much larger than the true factor correlation of .80; in fact, it is an improper correlation estimate because it is larger than 1.

Third, the AVE of a multi-item measure is always lower than a measure of scale reliability, except for boundary cases where all loadings are 1 or 0. Web Appendix D provides additional detail. Therefore, the FL criterion systematically underestimates the extent of discriminant validity. As a consequence, discriminant validity may appear to be violated when this is not the case. To illustrate, the FL criterion in Figure 1 is violated even though the factor correlation of .80 does not seem overly large: the squared factor correlation of .640 numerically exceeds both the AVE of F_1, which is .589, and the AVE of F_2, which is .490.

In sum, the FL criterion is an inappropriate discriminant validation criterion because it violates requirements 2 and 3. The FL criterion uses the average item reliability instead of the total scale reliability and therefore underestimates the support for discriminant validity. Using a statistical test would further exacerbate the issue because it makes an already overly strict criterion even more difficult to meet. The next section proposes to replace FL's comparison standard by a more appropriate metric based on the reliability of the entire multi-item scale.

The CR Criterion: An Improved Reliability-Based Discriminant Validation Criterion

Congeneric reliability (CR), also known as composite reliability, is a common estimate of scale reliability in the context of CFA (Bollen 1989; Cho 2021; Fornell and Larcker 1981; Jöreskog 1971; Werts et al. 1974). Cho (2021) compared 18 scale reliability coefficients and concluded that congeneric reliability mapped best onto the Classical Test Theory definition (Lord and Novick 1968), had the least restrictive assumptions compared to alternatives such as Cronbach's coefficient alpha, and performed well across datasets. To our knowledge, it has not been used as a comparison standard in discriminant validation.

Equation 6 specifies congeneric reliability for a construct F_1 measured with J items (Fornell and Larcker 1981; Jöreskog 1971; Werts et al. 1974):

$$CR_{F_1} = \frac{\left(\sum_{j=1}^{J} \lambda_{j1}\right)^2}{\left(\sum_{j=1}^{J} \lambda_{j1}\right)^2 + \sum_{i=1}^{J} \theta_{jj}}. \quad (6)$$

This expression looks superficially similar to AVE in Equation 4, but it differs because the numerator and denominator in Equation 6 both contain the square of the summed loadings rather than the sum of the squared loadings as in Equation 4. This makes a crucial difference. CR is the proportion of the total true score variance captured by the multi-item scale as a whole, rather than the average true variance captured by the individual items. Thus, CR correctly reconstructs the factor correlation when an observed correlation is disattenuated. In contrast to AVE, CR also increases with increasing scale length, as it should, because scales

with more items generally are more reliable. Web Appendix D provides additional details.

CR assumes a congeneric measurement model and therefore makes less restrictive assumptions than Cronbach's coefficient alpha, which assumes at least tau-equivalent items. In the example of Figure 1, the loadings of the items measuring F_1 differ, resulting in a CR of .804, which is larger than the coefficient alpha of .785 (which is a lower bound on reliability in the present case). The item loadings for F_2 (as well as the error variances, since the items are standardized) are the same and therefore the coefficient alpha of .742 is identical to CR.

The square root of CR is the correlation of the true score of the construct with the observed multi-item scale score. We propose to use it as comparison standard as follows:

$$DVI_{CR} = \sqrt{CR} - \varphi. \quad (7)$$

The subscript CR of the DVI indicates that the factor correlation is compared to the square root of CR. The CR criterion is stricter than the Phi criterion and more lenient than the FL criterion, except for corner cases with perfect reliability. The bootstrap can be used to construct a confidence interval and *P*-value around DVI. Web Appendix E shows that bootstrap intervals outperform Wald- and likelihood-based confidence intervals, particularly when the items are non-normally distributed.

Applying the CR criterion to the example in Figure 1, $DVI_{CR_{F_1}} = \sqrt{.804} - .80 = .097$ for F_1, and $DVI_{CR_{F_2}} = \sqrt{.742} - .80 = .062$ for F_2, both numerically larger than zero. At a sample size of n = 100, $DVI_{CR_{F_1}}$ has a 95% CI of [-.042, .247] (p = .092) and $DVI_{CR_{F_2}}$ has 95% CI [-.094, .221] (p = .225). At larger sample sizes of n = 500 ([.041, .154], p < .001 for F_1; [-.001, .126], p = .028 for F_2; note that the CI is two-sided and the p-value is one-sided) and n = 1,000 ([.057, .137], p < .001 for F_1; [.017, .104], p = .003 for F_2), the 95% CIs are narrower and the p-values are smaller.

In sum, the CR criterion meets all three requirements for an appropriate discriminant validation criterion (see Panel B of Table 1). It uses the factor correlation obtained from a

well-fitting CFA as the correlation metric to account for measurement error. It compares the factor correlation to the square root of CR, which captures the reliability of the multi-item scale and relies on less restrictive assumptions compared to alternatives such as Cronbach's coefficient alpha. The DVI based on CR quantifies the support for discriminant validity, with a precise *P*-value indicating the extent of incompatibility of the data with the model, and a confidence interval around DVI quantifying its estimated uncertainty. Next, we integrate the CR criterion into our proposed discriminant validation procedure.

A Proposed Discriminant Validation Procedure

Figure 2 summarizes our proposed discriminant validation procedure based on two appropriate criteria that are ordered from easier to more difficult to meet. The Phi criterion is a minimum criterion because it uses a perfect correlation as a comparison standard. When it is met, the more stringent CR criterion is tested next and can be met for none, one, or both focal constructs. Together, the two steps provide a four-tiered classification to facilitate decision-making and drawing conclusions. The two steps in the procedure and the four possible conclusions are described next. Annotated *R* code on *OSF* (https://osf.io/57srv/) implements the procedure when raw data are available. A *Shiny* app (https://constantpieters.shinyapps.io/discriminantvalidation/) implements the procedure when only summary statistics are available, which requires more restrictive assumptions (Web Appendix F has details).

Step 1 uses CFA to estimate the factor correlation and assess the extent to which it is smaller than 1. Only factor correlations from well-fitting models should be considered, as discussed earlier. A $(1-\alpha)$ bootstrapped confidence interval provides the estimated uncertainty associated with the point estimate of $DVI_1 = 1-\varphi$, and a test of the null hypothesis that $DVI_1 = 0$ vs. the alternative hypothesis $DVI_1 > 0$ yields a *P*-value. If the Phi

criterion is supported (i.e., DVI_1 is sufficiently larger than zero), move to Step 2, else end the procedure. In the latter case, there is insufficient support for discriminant validity of the two constructs because even the minimum Phi criterion is not met (Conclusion 1 in Figure 2).

Step 2 compares the factor correlation with the square roots of the CRs of the two constructs. Again, *P*-values are available and a $(1-\alpha)$ bootstrapped confidence interval around $DVI_{CR} = \sqrt{CR} - \varphi$ quantifies the uncertainty around the point estimate. The null hypotheses for the tests are that $DVI_{CR_{F_1}} = 0$ and $DVI_{CR_{F_2}} = 0$, and the respective alternative hypotheses are $DVI_{CR_1} > 0$ and $DVI_{CR_2} > 0$, respectively.

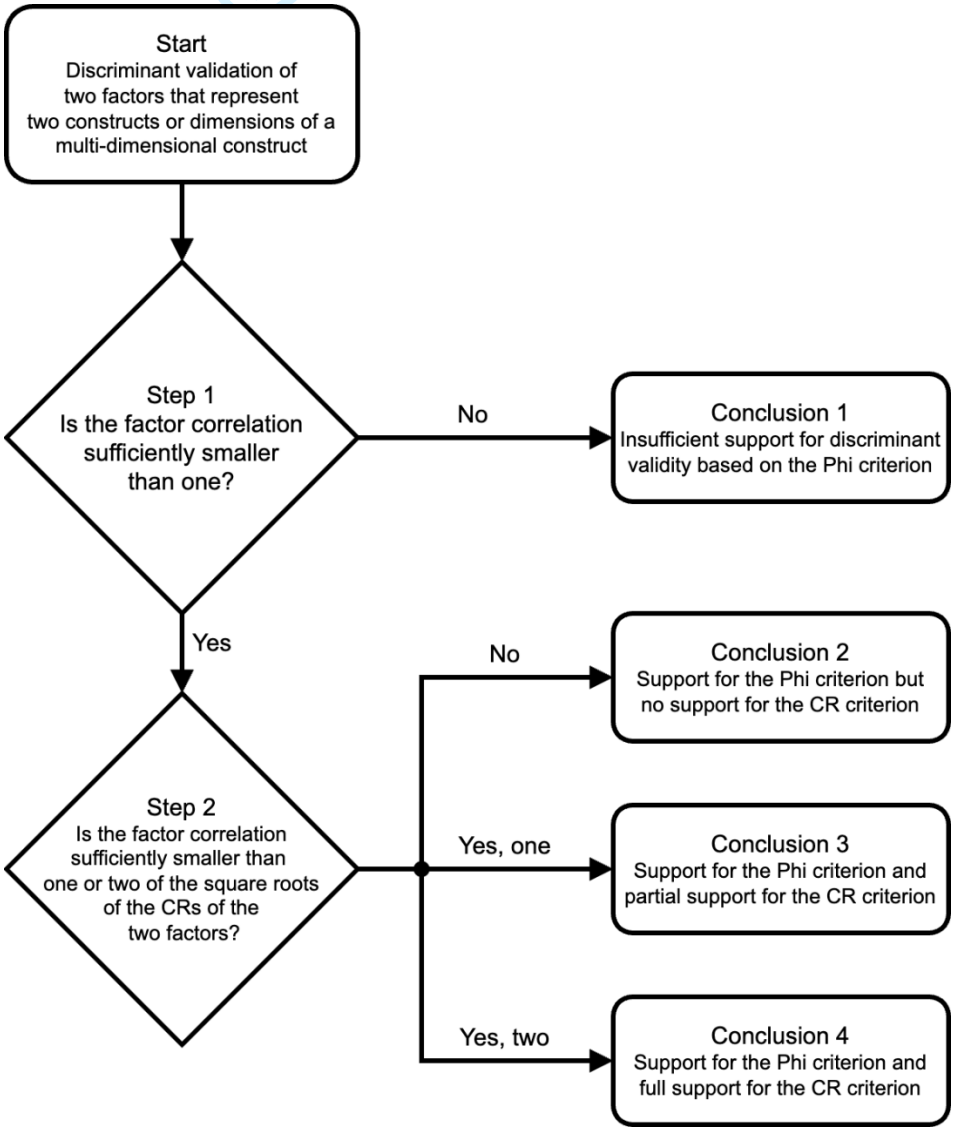


Figure 2. Proposed Discriminant Validation Procedure.
Note: CR is congeneric reliability.

A construct pair only meets the Phi criterion if neither of the two factors meet the CR criterion (Conclusion 2). A construct pair partially meets the CR criterion if there is sufficient support for only one factor (Conclusion 3), which can occur when the scale reliabilities differ. A construct pair fully meets the CR criterion if both DVIs are sufficiently positive while accounting for sampling error (Conclusion 4). These conclusions are drawn based on continuous statistical results (McShane et al. 2024). In both steps, those are the DVIs, their confidence intervals, and the *P*-values of tests of the DVIs against zero. Since discriminant validity is a precondition for meaningful theory testing (Pieters 2017), the four conclusions in the proposed procedure aid reporting and decision-making, a point to which the General Discussion returns.

When applying the proposed procedure to the example in Figure 1 at a modest sample size of $n = 100$, only the Phi criterion is met (Conclusion 2) based on $DVI_1 = .20$ with 95% CI [.059, .355] and $p = .002$, but the CR criterion is not met for either F_1 ($DVI_{CR_{F_1}} = .097$ with 95% CI [-.042, .247], $p = .092$) or F_2 ($DVI_{CR_{F_2}} = .062$ with 95% CI [-.094, .221], $p = .225$). At a larger sample size of $n = 500$, the CR criterion is partially met (Conclusion 3) because the DVI is sufficiently positive for F_1 (95% CI [.041, .154], $p < .001$) but not for F_2 (95% CI [-.001, .126], $p = .028$). At an even larger sample size of $n = 1{,}000$, the CR criterion is fully met (Conclusion 4) for the DVI of both F_1 (95% CI [.057, .137], $p < .001$) and F_2 (95% CI [.017, .104], $p = .003$).

The proposed procedure provides a systematic approach to assessing the extent to which two constructs are discriminant. It provides continuous measures of discriminant validity and a decision-theoretic procedure with four conclusions that go beyond drawing binary conclusions based on a single criterion. The next section presents empirical applications of the procedure. It shows that the procedure can result in conclusions that differ from the conclusions obtained via currently used criteria.

Empirical Applications

We first present a case study to illustrate the proposed procedure with actual rather than synthetic data and to empirically compare it with other criteria. Then, we reanalyze seven articles from the 79 in the literature review, each one of which reported at least one discriminant validity issue based on the Phi or FL criterion.

Case Study

Wilson and Bellezza (2022) conceptualize the construct of consumer minimalism with a Minimalist Consumer Scale that has three conceptually distinct facets, each measured with four items: Number of Possessions (NoP; e.g., "I restrict the number of things I own"), Sparse Aesthetic (SA; e.g., "I prefer simplicity in design"), and Mindfully Curated Consumption (MCC; e.g., "I am mindful of what I own"). In their Study 3, n = 394 valid *Prolific* panelists completed the Minimalist Consumer Scale, and CFAs supported its dimensionality and reliability. Our reanalysis based on raw data focuses on NoP and SA, which had the largest factor correlation. Wilson and Bellezza (2022) concluded discriminant validity of the Minimalist Consumer Scale, and our reanalysis strengthens this conclusion.

Figure 3 plots the estimated DVI of NoP with SA (based on the factor correlation) on the vertical axis against a continuous range of comparison standards from .65 to 1 on the horizontal axis, which furthermore denotes the estimated factor correlation of .780 (the HTMT is .789). The blue diamond also points out that factor correlation, at which DVI = 0. The DVI slopes upwards: the higher the comparison standard, the more the DVI exceeds 0, and the greater the support for discriminant validity. DVIs below 0 indicate violations of discriminant validity. Also visualized are the 90%, 95%, and 99% bootstrapped confidence intervals around DVI. The goal of Figure 3 is to compare several discriminant validation criteria, and we discuss them one by one. Web Appendix G tabulates the detailed results.

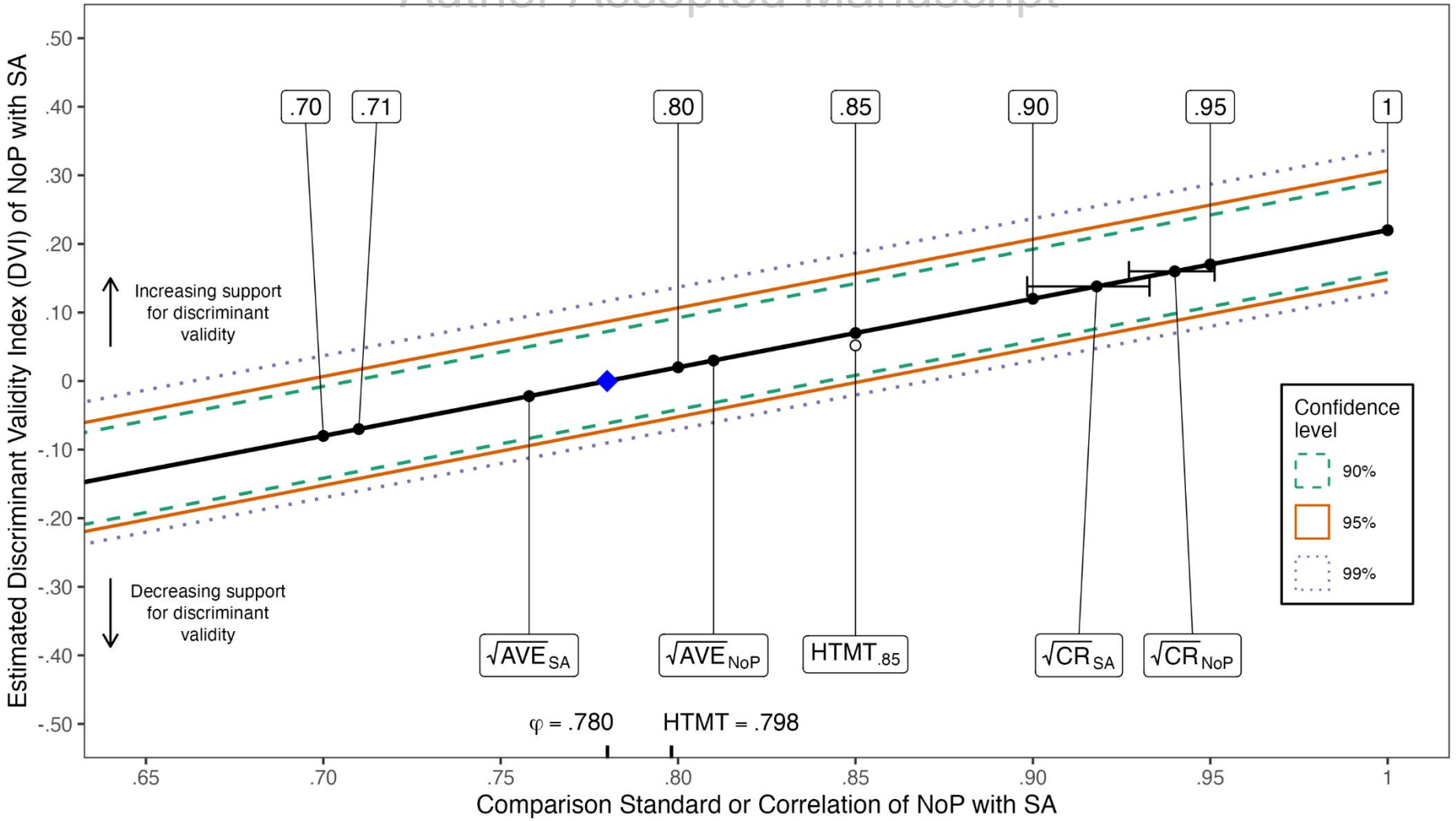


Figure 3. Case Study of Wilson and Bellezza (2022, Study 3): Support for Discriminant Validity Depends on the Comparison Standard.
Notes: NoP refers to Number of Possessions and SA to Sparse Aesthetics. The horizontal axis displays the estimated factor correlation (at which DVI = 0, denoted by the blue diamond) and the heterotrait-monotrait ratio (HTMT). Error bars denote 95% bootstrapped confidence intervals around $\sqrt{CR}$. The white point represents the DVI for the HTMT criterion with a .85 standard. AVE is average variance extracted and CR is congeneric reliability.

First, the Phi criterion uses a comparison standard of unity and has the maximum support for discriminant validity: DVI_1 = .220, with the lower limit of the confidence intervals much greater than zero (e.g., 95% CI [.148, .307], p < .001), visualized at the right side of Figure 3. This is intuitive because the estimated factor correlation is a moderate .780, and the Phi criterion only flags issues with very large correlations that are close to 1.

Second, the estimated congeneric reliabilities are .884 for NoP and .842 for SA, and their square roots are much larger than the factor correlation. For instance, the $DVI_{CR_{NoP}}$ = .160 with 95% CI [.088, .245] (p < .001), and $DVI_{CR_{SA}}$ = .138 with 95% CI [.064, .223] (p < .001). Thus, applying our proposed procedure, both the Phi and CR criteria are fully met such that there is convincing support for discriminant validity (Conclusion 4 in Figure 2).

Third, the result based on CR is in stark contrast to that for the FL criterion, which is met for NoP but not for SA: the estimated AVEs are .656 ($DVI_{AVE_{NoP}}$ = .030) and .574 ($DVI_{AVE_{SA}}$ = -.022). The 95% CI for the DVI of NoP is [-.044, .115] (p = .227) and for SA it is [-.101, .068] (p = .299). Thus, if Wilson and Bellezza had reported the FL criterion (which they did not), discriminant validity would have been rejected for SA.

Fourth, the estimate of the HTMT ratio is .798, very similar to the .780 factor correlation, as it should be. Using .85 as a comparison standard, $DVI_{HTMT_{.85}}$ = .052 (denoted by the white point in Figure 3). There is no convincing support for discriminant validity since the confidence intervals overlap zero (e.g., 95% CI [-.009, .126], p = .051). This conclusion is not very informative, however, because the .85 standard is arbitrary.

Fifth, Figure 3 includes fixed comparison standards less than 1 as suggested in the literature. Conclusions about discriminant validity depend much on the judgmental standard chosen. For example, discriminant validity is supported for a comparison standard of .90 ($DVI_{.90}$ = .120 with 95% CI [.048, .207], p < .001), but not for a comparison standard of .80 ($DVI_{.80}$ = .020 with 95% CI [-.052, .107], p = .316). Without a clear theoretical foundation,

drawing unambiguous conclusions based on judgmental standards is hard even if multiple standards are used. The factor correlation of .78 is substantially smaller than .90, but not much smaller than .80, which would yield a "marginal" discriminant validity problem according to Rönkkö and Cho (2022). In contrast, the CR criterion estimates the comparison standard from the data based on a clear theoretical foundation. The CR and judgmental criteria only coincide if the judgmental standard happens to be equal to the square root of CR.

Reanalysis of Seven Articles Reporting Discriminant Validity Violations

Eight articles in the literature review reported at least one discriminant validity violation. One article reported not meeting the Phi criterion (Bloch et al. 2003). Seven articles reported failures to meet the FL criterion (Drolet et al. 2021; Fischer et al. 2010; Garbinsky et al. 2020; Grohmann 2009; Homburg and Pflesser 2000; Morhart et al. 2015; Nenkov et al. 2008). None of the four articles reporting the HTMT criterion identified discriminant validity issues.

Whereas the case analysis relied on raw data, this reanalysis used sufficient summary data reported in the articles, assuming well-behaved data and adequately fitting measurement models. One article did not report sufficient data (Homburg and Pflesser 2000). Seven articles did; three reported factor correlations and standardized loadings (Bloch et al. 2003; Morhart et al. 2015; Nenkov et al. 2008), and four reported factor correlations and AVEs or coefficient alpha. This enabled us to reconstruct the item correlations assuming parallel items. Although this assumption is restrictive, it was necessary given the available data. Our reanalysis replicated all reported violations of discriminant validity, which is reassuring and suggests that the parallel item assumption is reasonable. In total, we reanalyzed thirteen construct pairs based on seven articles, using Monte Carlo resampling to estimate confidence intervals (Preacher and Selig 2012). We report the results from the proposed procedure and focus on the comparison with the FL criterion because of its prominence in scale

development research. We focus on the numerical comparison based on the FL criterion, which is how the criterion is used in practice. Web Appendix G contains statistical results for the FL criterion as well as the results for the HTMT criterion and fixed comparison standards.

Table 2 summarizes the results. We refer to specific construct pairs by their row number and name of the first author in Table 2. Several findings stand out.

First, at a sample size of n = 136, the construct pairs in rows 1 and 2 (Bloch) failing the Phi criterion had factor correlations of .930 (DVI_1 = .070 with 95% CI [-.022, .164], p = .065) and .880 (DVI_1 = .120 with 95% CI [-.013, .255], p = .038; note again that the p-value is one-sided), respectively. Although the Phi criterion identified a violation of discriminant validity in these two cases, the reanalysis also demonstrates the relative leniency of the Phi criterion: The second largest factor correlation of .917 at n = 367 in row 4 (Nenkov) still supports the Phi criterion: DVI_1 = .083 with 95% CI [.032, .134] and p = .001.

Second, the FL criterion is not only inappropriate but also overly strict. It flags discriminant validity issues even at moderate factor correlations. In row 13 (Drolet), for instance, a factor correlation of .470 at a sample size of n = 65 had AVEs of .350 and .090, respectively. Using a numerical comparison, the FL criterion was violated for the second construct with $DVI_{AVE_{F_2}}$ = -.170 (but not for the first construct with $DVI_{AVE_{F_1}}$ = .122).

Third, the FL and CR criteria are not only conceptually but also empirically different. That is, 16 out of the 20 cases in Panel B of Table 2 violated the FL criterion when using a numerical comparison (all except one comparison in rows 10-13). However, the CR criterion was statistically supported in 9 out of 16 cases based on confidence intervals not overlapping zero. For example, returning to row 13 (Drolet), the CR criterion was supported: $DVI_{CR_{F_1}}$ = .431 with 95% CI [.172, .683] and p = .001, and $DVI_{CR_{F_2}}$ = .395 with 95% CI [.134, .647] and p = .002. The average difference between the DVIs based on the CR and FL criteria was .215 (range of .092 to .565).

Table 2. Reanalysis of Seven Articles Reporting Discriminant Validity Violations.

Row	Article and study or sample	*n*	Measures (number of items)	CRs	AVEs	φ	FL criterion: DVI_{AVE}	Proposed discriminant validation procedure: Step 1: DVI_1 [95% CI] (*p*)	Proposed discriminant validation procedure: Step 2: DVI_{CR} [95% CI] (*p*)	Conclusion
Panel A: Study reporting at least one construct pair not supporting the Phi criterion										
1	Bloch et al. (2003; Study 3)	136	Value (4)	.794	.493	.930	-.228	.070 [-.022, .164] (.065)	-.039 [-.136, .060] (.218)	1
			Acumen (4)	.726	.400		-.297		-.078 [-.186, .027] (.071)	
2	Bloch et al. (2003; Study 3)	136	Acumen (4)	.726	.400	.880	-.247	.120 [-.013, .255] (.038)	-.028 [-.171, .114] (.346)	1
			Response (3)	.655	.393		-.253		-.071 [-.229, .085] (.182)	
3	Bloch et al. (2003; Study 3)	136	Value (4)	.794	.493	.870	-.168	.130 [.009, .248] (.017)	.021 [-.104, .143] (.376)	2
			Response (3)	.655	.393		-.243		-.061 [-.205, .080] (.199)	
Panel B: Studies reporting at least one construct pair not supporting the FL criterion										
4	Nenkov et al. (2008; Part 1: Sample 1)	367	Generation (3)	.795	.564	.917	-.166	.083 [.032, .134] (.001)	-.025 [-.083, .031] (.190)	2
			Evaluation (3)	.791	.559		-.169		-.027 [-.084, .028] (.173)	
5	Grohmann (2009; Study 4a)	2,406	Female brand personality (6)	.907	.620	.890	-.103	.110 [.100, .121] (< .001)	.063 [.052, .074] (< .001)	4
			Female human personality (20)	.950	.487		-.192		.085 [.074, .095] (< .001)	
6	Morhart et al. (2015; Study 4)	810	Credibility (3)	.819	.603	.883	-.106	.117 [.089, .144] (< .001)	.022 [-.009, .051] (.081)	3
			Trustworthiness (5)	.895	.630		-.089		.063 [.034, .090] (< .001)	
7	Fischer et al. (2010; 2008 wave)	4,200	Symbolic value (2)	.745	.594	.879	-.108	.121 [.099, .142] (< .001)	-.016 [-.040, .008] (.104)	2
			Group identity (2)	.733	.579		-.118		-.023 [-.047, .002] (.037)	
8	Fischer et al. (2010; 2006 wave)	11,539	Symbolic value (2)	.734	.580	.862	-.100	.138 [.124, .152] (< .001)	-.005 [-.021, .010] (.259)	2
			Group identity (2)	.722	.565		-.110		-.012 [-.028, .004] (.065)	
9	Grohmann (2009; Study 4a)	2,406	Male brand personality (6)	.918	.650	.820	-.014	.180 [.166, .194] (< .001)	.138 [.124, .152] (< .001)	4
			Male human personality (20)	.960	.545		-.081		.160 [.146, .174] (< .001)	
10	Morhart et al. (2015; Study 4)	810	Credibility (3)	.819	.603	.755	.022	.245 [.203, .287] (< .001)	.150 [.107, .193] (< .001)	4
			Sincerity (11)	.850	.340		-.172		.167 [.125, .209] (< .001)	
11	Garbinsky et al. (2020; Study 2: Sample A)	309	Financial infidelity (12)	.926	.510	.500	.214	.500 [.407, .593] (< .001)	.462 [.369, .554] (< .001)	4
			Machiavellianism (20)	.857	.230		-.020		.426 [.333, .518] (< .001)	
12	Garbinsky et al. (2020; Study 2: Sample C)	311	Financial infidelity (12)	.959	.660	.490	.322	.510 [.411, .608] (< .001)	.489 [.390, .587] (< .001)	4
			Financial literacy (13)	.765	.200		-.043		.384 [.284, .482] (< .001)	
13	Drolet et al. (2021; Sample 1)	65	Preference for moderation (8)	.812	.350	.470	.122	.530 [.272, .785] (< .001)	.431 [.172, .683] (.001)	4
			Domain-specific risk (30)	.748	.090		-.170		.395 [.134, .647] (.002)	

Notes: Rows report measure pairs from the review of 79 scales (Web Appendix A) for which sufficient data were available for reanalysis, sorted in descending order by the absolute value of the factor correlation (φ). Also, *n* is sample size, CR is congeneric reliability, AVE is average variance extracted. FL refers to Fornell and Larcker (1981), DVI is the discriminant validity index. CIs of the Phi and CR criteria are reported with 95% Monte Carlo confidence interval (CI) within square brackets and one-sided *p*-value within parentheses. The conclusion column refers to the substantive conclusions in Figure 2.

Fourth, the rightmost column of Table 2 summarizes, for each construct pair, the final conclusion from the discriminant validation procedure. Eleven of the 13 pairs meet the minimum Phi criterion. Four pairs meet the Phi criterion but not the stricter CR criterion. One pair partially meets the CR criterion, and six pairs fully meet the CR criterion. In sum, the proposed four-tiered procedure provides new and informative insights.

General Discussion

Conclusions

This research outlines three requirements that an appropriate discriminant validation criterion should satisfy. First, an appropriate criterion uses a factor correlation from a well-fitting CFA that accounts for measurement error. Second, the factor correlation is compared to a principled maximum such as a perfect correlation or the square root of scale reliability (i.e., congeneric reliability or CR). Third, the comparison method accounts for sampling error and provides a continuous measure of support for discriminant validity.

The analysis of current criteria showed that the Phi criterion, which specifies that the factor correlation should be less than perfect, meets all requirements but is lenient. Moreover, the Fornell and Larcker (FL) criterion fails the second and third requirements. We therefore recommend that FL not be used in future scale validation research and recommend to replace it with the CR criterion, which meets all three requirements.

We integrated the more lenient Phi criterion and the more stringent CR criterion into a procedure that yields a four-tiered classification of discriminant validity. The procedure provides continuous measures of the degree of support for discriminant validity, as well as statistical results. Empirical applications demonstrated how the proposed procedure systematically assesses discriminant validity and can lead to conclusions that differ from those based on the FL criterion.

Implications

First, the CR criterion is easier to meet than the FL criterion that it replaces because multi-item scale reliability is generally higher than average item reliability (on average .328 across the reanalyses in Table 2). The CR criterion can therefore support the discriminant validity of more closely related constructs that would not pass muster with the FL criterion, which our empirical applications demonstrate. We speculate that valid scales of important constructs may not have reached the submission or publication stage because they failed to meet the inappropriate FL criterion. At the same time, the CR criterion flags discriminant validity issues between constructs that are too closely related but still meet the more lenient Phi criterion. Two constructs meeting the CR criterion strengthens their construct validity because it implies that each construct correlates more strongly with its own multi-item measure than with another, presumably distinct construct. This ensures meaningful theory testing (Pieters 2017) and prevents the proliferation of constructs that may be semantically but not empirically distinct (Albert and Thomson 2023; Morrow 1983; Shaffer et al. 2016).

Second, our procedure meaningfully combines the Phi and CR criteria to provide graded discriminant validity conclusions. It is common practice to report multiple discriminant criteria in scale development research; 60 out of the 79 reviewed articles (76%) did so, although there was little consistency in how different criteria were combined (Web Appendix A has details). Prior literature appears to have treated the various criteria as equivalent without clearly differentiating their relative stringency. In contrast, the proposed procedure systematically combines two increasingly stringent criteria and acknowledges that discriminant validity is a matter of degree, rather than a single all-or-nothing decision (Rönkkö and Cho 2022). Quantifying the extent of support for discriminant validity in terms of the DVI, together with a confidence interval around DVI and a precise *P*-value based on a test of DVI = 0, further supports the continuous reporting and interpretation.

Third, our analysis focused on the discriminant validation of multi-item scales using CFA with a reflective measurement model; the Phi and CR criteria are sufficient for this purpose, without the need for additional criteria. This is particularly the case for the FL criterion, which the CR criterion replaces. Our analysis showed that AVE should not be used as a comparison standard for discriminant validity, although it might still be reported as a useful summary measure of individual-item reliability or convergent validity. We see little reason for authors to report the HTMT criterion for reflective measurement models because it potentially violates the first requirement (at best it is an indirect way to estimate the factor correlation), and it uses an arbitrary comparison standard (usually .85), thus violating the second requirement. The HTMT criterion was initially developed for variance-based SEM (such as PLS), and it might be informative in that context (Henseler et al. 2015). However, it is unnecessary for reflective measurement models. Arbitrary fixed comparison standards such as .85 should also be avoided in general, although the systematic use of multiple standards alleviates the problem of selecting an appropriate single fixed standard to some extent (Rönkkö and Cho 2022).

Implementing the Proposed Procedure

To facilitate the implementation of the proposed procedure using raw data, researchers can use the *R* code described in Web Appendix F and available on *OSF* (https://osf.io/57srv/). For purposes of illustration, the code implements the case study reported above and in Figure 3. An online *Shiny* app (https://constantpieters.shinyapps.io/discriminantvalidation/) enables discriminant validation with menus when only summary data are available, assuming standardized and parallel items. Using raw data for discriminant validation is preferable: it relaxes the assumption that all items are equally good measures of the underlying construct, which provides the most appropriate assessment of discriminant validity.

We recommend that researchers fully report and continuously interpret the following

results of discriminant validation: factor correlations, CR estimates, estimated DVIs for the Phi and CR criteria, the estimated uncertainty of each DVI captured by confidence intervals, and *P*-values for the tests of each DVI against zero. Using our reanalysis of Study 3 in Wilson and Bellezza (2022) as an example, the results could be summarized as follows:

> *We assessed discriminant validity with the Phi and CR criteria. Number of Possessions (CR = .884) and Sparse Aesthetic (CR = .842) have a factor correlation of* φ *= .780, which meets the Phi criterion (*DVI_1 *= .220 with 95% bootstrapped CI based on 10,000 resamples [.148, .307], one-sided bootstrapped* $p < .001$*). Moreover, the factor correlation is substantially smaller than the square root of the CR for Number of Possessions (*DVI_{CR} *= .160 with 95% CI [.088, .245],* $p < .001$*) and Sparse Aesthetic (*DVI_{CR} *= .138 with 95% CI [.064, .223],* $p < .001$*). Thus, the Phi and CR criteria both support the discriminant validity of Number of Possessions with Sparse Aesthetic in the present study.*

When constructing new scales or using scales in subsequent theory testing, constructs or subscales of constructs ideally meet both the Phi and CR criteria. Since the CR criterion is more stringent, it should be met for constructs to be clearly distinct. At the opposite end of the spectrum, if two constructs do not even meet the minimum Phi criterion, it is hard to argue for their distinctiveness. How to proceed in this case depends on the specific context. Two subdimensions of a construct that are not distinct can be combined, as was done in Bloch et al. (2003) (see rows 1-3 in Table 2). Alternatively, one dimension can be dropped (Henseler et al. 2015). If a new scale of a construct cannot be empirically distinguished from an existing scale of the same construct, there is little utility in proposing the new scale, all other things such as the number of scale items and the ease of responding being equal. If a new construct is too similar to a different existing construct, then there is little incremental value in the new construct and in order to avoid construct proliferation, the new construct

should likely be dropped.

It is more difficult to offer normative recommendations when the Phi criterion is met but the CR criterion is not, for either one construct or both constructs. In these cases, either the reliability of one or both constructs is too low or the factor correlation is too high. If the scale reliabilities are comparatively low, efforts might be directed at improving measurement reliability before proceeding to the next stage of scale development. However, researchers should avoid the trap of inflating scale reliability artificially by adding synonyms of existing items, which leads to unnecessarily long scales with a narrow scope (Voorhees et al. 2016, p. 132). The CR criterion can be met for only one factor if the reliabilities differ. This insight is particularly useful when two competing scales of a construct are validated because it suggests that the scale for which the CR criterion is met is to be preferred based on the discriminant validation results. If the factor correlation is large, subscales might be combined, which Nenkov et al. (2008) did based on a correlation of .917 (also see row 4 in Table 2). In mature research areas, large factor correlations may be tolerated if it can be demonstrated that the constructs can be treated as distinct. In particular, if indirect approaches to discriminant validation show that two highly correlated constructs are differentially related to a third construct, or that a third construct moderates the relation between two highly correlated constructs (Franke et al. 2021; Tesser and Krauss 1976), then there is remedial evidence of discriminant validity despite the large factor correlation relative to the scale reliabilities.

To conclude, discriminant validation is a crucial step in the development and deployment of multi-item scales. This research proposed a new CR criterion and systematic procedure that combines the CR criterion with the existing Phi criterion. We hope that the new criterion and procedure contribute to improved discriminant validation of multi-item scales and thereby to better theory and practice.

References

Albert, Noel and Matthew Thomson (2023), "Epistemological Jangle and Jingle Fallacies in the Consumer–Brand Relationship Subfield: A Call to Action," *Journal of Consumer Research*, 51 (2), 383-407.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (3), 411-23.

Asparouhov, Tihomir and Bengt Muthén (2009), "Exploratory Structural Equation Modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, 16 (3), 397-438.

Bagozzi, Richard P. and Youjae Yi (1988), "On the Evaluation of Structural Equation Models," *Journal of the Academy of Marketing Science*, 16 (1), 74-94.

Baumgartner, Hans and Bert Weijters (2022), "Structural Equation Modeling," in *Handbook of Market Research*, Christian Homburg and Martin Klarmann and Arnd Vomberg, eds. Cham: Springer.

Bloch, Peter H., Frédéric F. Brunel, and Todd J. Arnold (2003), "Individual Differences in the Centrality of Visual Product Aesthetics: Concept and Measurement," *Journal of Consumer Research*, 29 (4), 551-65.

Bollen, Kenneth A. (1989), Structural Equations with Latent Variables. New York: Wiley.

Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (2), 81-105.

Cheung, Shu Fai and Ivan Jacob Agaloos Pesigan (2023), "semlbci: An R Package for Forming Likelihood-Based Confidence Intervals for Parameter Estimates, Correlations, Indirect Effects, and Other Derived Parameters," *Structural Equation Modeling: A Multidisciplinary Journal*, 30 (6), 985-99.

Cho, Eunseong (2021), "Neither Cronbach's Alpha nor McDonald's Omega: A Commentary on Sijtsma and Pfadt," *Psychometrika*, 86 (4), 877-86.

Churchill, Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (1), 64-73.

Drolet, Aimee, Mary Frances Luce, Li Jiang, Benjamin C. Rossi, and Reid Hastie (2021), "The Preference for Moderation Scale," *Journal of Consumer Research*, 47 (6), 831-54.

Efron, Bradley and Robert J. Tibshirani (1994), An Introduction to the Bootstrap. NY: Chapman and Hall.

Farrell, Andrew M. (2010), "Insufficient Discriminant Validity: A Comment on Bove, Pervan, Beatty, and Shiu (2009)," *Journal of Business Research*, 63 (3), 324-27.

Fischer, Marc, Franziska Völckner, and Henrik Sattler (2010), "How Important Are Brands? A Cross-Category, Cross-Country Study," *Journal of Marketing Research*, 47 (5), 823-39.

Fornell, Claes and David F. Larcker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (1), 39-50.

Franke, George R., Marko Sarstedt, and Nicholas P. Danks (2021), "Assessing Measure Congruence in Nomological Networks," *Journal of Business Research*, 130, 318-34.

Franke, George and Marko Sarstedt (2019), "Heuristics Versus Statistics in Discriminant Validity Testing: A Comparison of Four Procedures," *Internet Research*, 29 (3), 430-47.

Garbinsky, Emily N., Joe J. Gladstone, Hristina Nikolova, and Jenny G. Olson (2020), "Love, Lies, and Money: Financial Infidelity in Romantic Relationships," *Journal of Consumer Research*, 47 (1), 1-24.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," *Journal of Marketing Research*, 25 (2), 186-92.

Grohmann, Bianca (2009), "Gender Dimensions of Brand Personality," *Journal of Marketing Research*, 46 (1), 105-19.

Haws, Kelly L., Kevin L. Sample, and John Hulland (2023), "Scale Use and Abuse: Towards Best Practices in the Deployment of Scales," *Journal of Consumer Psychology*, 33 (1), 226-43.

Henseler, Jörg, Christian M. Ringle, and Marko Sarstedt (2015), "A New Criterion for Assessing Discriminant Validity in Variance-Based Structural Equation Modeling," *Journal of the Academy of Marketing Science*, 43 (1), 115-35.

Homburg, Christian and Christian Pflesser (2000), "A Multiple-Layer Model of Market-Oriented Organizational Culture: Measurement Issues and Performance Outcomes," *Journal of Marketing Research*, 37 (4), 449-62.

Jöreskog, Karl G. (1971), "Statistical Analysis of Sets of Congeneric Tests," *Psychometrika*, 36 (2), 109-33.

Lord, Frederic M. (1957), "A Significance Test for the Hypothesis That Two Variables Measure the Same Trait except for Errors of Measurement," *Psychometrika*, 22 (3), 207-20.

Lord, Frederic M. and Melvin R. Novick (1968), Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

MacKenzie, Scott B., Philip M. Podsakoff, and Nathan P. Podsakoff (2011), "Construct Measurement and Validation Procedures in *MIS* and Behavioral Research: Integrating New and Existing Techniques," *MIS Quarterly*, 35 (2), 293-334.

McShane, Blakeley B., Eric T. Bradlow, John G. Lynch, and Robert J. Meyer (2024),

"“Statistical Significance” and Statistical Reporting: Moving Beyond Binary," *Journal of Marketing*, 88 (3), 1-19.

Morhart, Felicitas, Lucia Malär, Amélie Guèvremont, Florent Girardin, and Bianca Grohmann (2015), "Brand Authenticity: An Integrative Framework and Measurement Scale," *Journal of Consumer Psychology*, 25 (2), 200-18.

Morrow, Paula C. (1983), "Concept Redundancy in Organizational Research: The Case of Work Commitment," *The Academy of Management Review*, 8 (3), 486-500.

Nenkov, Gergana Y., J. Jeffrey Inman, and John Hulland (2008), "Considering the Future: The Conceptualization and Measurement of Elaboration on Potential Outcomes," *Journal of Consumer Research*, 35 (1), 126-41.

Pieters, Rik (2017), "Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication," *Journal of Consumer Research*, 44 (3), 692-716.

Preacher, Kristopher J. and James P. Selig (2012), "Advantages of Monte Carlo Confidence Intervals for Indirect Effects," *Communication Methods and Measures*, 6 (2), 77-98.

Ringle, Christian M., Marko Sarstedt, Noemi Sinkovics, and Rudolf R. Sinkovics (2023), "A Perspective on Using Partial Least Squares Structural Equation Modelling in Data Articles," *Data in Brief*, 48, 1-21.

Rönkkö, Mikko and Eunseong Cho (2022), "An Updated Guideline for Assessing Discriminant Validity," *Organizational Research Methods*, 25 (1), 6-47.

Shaffer, Jonathan A., David DeGeest, and Andrew Li (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.

Spearman, Charles (1904), "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, 15 (1), 72-101.

Tesser, Abraham and Herbert Krauss (1976), "On Validating a Relationship between Constructs," *Educational and Psychological Measurement*, 36 (1), 111-21.

Voorhees, Clay M., Michael K. Brady, Roger Calantone, and Edward Ramirez (2016), "Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies," *Journal of the Academy of Marketing Science*, 44 (1), 119-34.

Werts, C. E., R. L. Linn, and K. G. Jöreskog (1974), "Intraclass Reliability Estimates: Testing Structural Assumptions," *Educational and Psychological Measurement*, 34 (1), 25-33.

Wilson, Anne V. and Silvia Bellezza (2022), "Consumer Minimalism," *Journal of Consumer Research*, 48 (5), 796-816.

Web Appendix

Improving the Discriminant Validation of Multi-Item Scales

Constant Pieters (cpi.marktg@cbs.dk)

Hans Baumgartner (jxb14@psu.edu)

Rik Pieters (f.g.m.pieters@tilburguniversity.edu)

OSF has additional details, data, and analysis code: https://osf.io/57srv/.

Web Appendix A: Literature Review

This Web Appendix summarizes the literature review.

Method

The review focuses on discriminant validation in scale development research because it is generally understood that the subdimensions of a scale that is under development should be empirically distinct. Moreover, the focal construct should be empirically distinct from related but theoretically distinct constructs (Churchill 1979; Gerbing and Anderson 1988).

The review covers scales that were published in the 2000-2024 volumes of the *Journal of Consumer Psychology (JCP)*, *Journal of Consumer Research (JCR), Journal of Marketing (JM)*, and *Journal of Marketing Research (JMR)*. To identify scale development articles, we extracted meta-data of all published articles from the *Web of Science* with the following query: "(SO = (JOURNAL OF CONSUMER PSYCHOLOGY) OR SO = (JOURNAL OF MARKETING) OR SO = (JOURNAL OF MARKETING RESEARCH) OR SO = (JOURNAL OF CONSUMER RESEARCH)) AND PY = (2000-2024)." We then searched all article titles with case-insensitive regular expressions related to scales, measures, measurement scales, and measure/scale construction/development as follows: "scale", "measure", "construct", "concept", "individual-difference" (with and without a hyphen), "trait", and "dimension". We then conducted similar searches in the article abstracts and article keywords. We supplemented these search results with a citation analysis of Fornell and Larcker (1981) and a list of 25 scale development articles identified by Baumgartner and Weijters (2019). We followed up with additional full-text and *Google Scholar* searches. During the selection process, articles were manually inspected for eligibility.

Following Haws et al. (2023), we define validated scales as those that are developed and vetted in accordance with established guidelines (e.g., Churchill 1979; Gerbing and

Anderson 1988). An article was eligible for further analysis if it met the following requirements. First, a major objective of the article had to be the development or validation of a scale. Second, the scale had to consist of multiple items. Third, the construction of the scale had to follow the generally recommended multi-item scale development steps such as generating an initial item pool, pruning, and subsequent validation of the proposed scale in new samples. These criteria distinguish formal scale development and validation from scale deployment or the use of ad-hoc scales (Haws et al. 2023). This led to a set of 87 articles for further analysis, which are listed on *OSF*.

Results

Overall, out of the 87 articles, 79 (91%) analyzed the discriminant validity of the focal scale and/or its subscales. The high incidence of discriminant validation reporting emphasizes that discriminant validity is a key facet of construct validity and that discriminant validation is an important step in the scale development process. The 79 articles reported an average of 2.20 criteria (Median = 2, SD = .992, range = 1-5). Most articles reported two criteria (36 articles, 46% out of 79), followed by one criterion (19 or 24%), three criteria (15 or 19%), four criteria (7 or 9%), and up to five criteria (2 or 3%).

Table WA1 summarizes the criteria reported in the 79 articles. Panel A summarizes the reported criteria. The percentages do not sum to 100% due to rounding and articles reporting multiple criteria. Twenty-seven (34% out of 79) articles subjectively assessed the size of the estimated correlation, for example by noting that the estimated correlations were "moderate," but without a predetermined or explicit cutoff.

Forty-one articles (52%) reported the Phi criterion based on a nested model test, which compares a baseline model with a freely estimated factor correlation with a restricted model that fixes the factor correlation to one. An alternative but equivalent specification of the restricted model is to combine the measures of two constructs into a single factor.

Table WA1:
Discriminant Validation Criteria Reported in Marketing Scale Development Research 2000-2024

Panel A: Discriminant validation criteria reported

Criterion	# articles	% articles (out of 79)	# violations	% violations (out of # articles)
No explicit criterion or subjective judgment about the size of the factor correlation	27	34%	0	0%
Factor correlation should be smaller than one, based on a nested model test of a model with a freely estimated factor correlation against a model with the factor correlation fixed to one (Phi: Nested)	41	52%	1	2%
Factor correlation should be smaller than one, based on the confidence interval of the estimated factor correlation not overlapping one (Phi: CI)	16	20%	0	0%
Factor correlation should be smaller than one, based on the *P*-value of a test of the factor correlation against one being sufficiently small (Phi: *p*)	2	3%	0	0%
Estimated average variance extracted should exceed the squared factor correlation (FL)	43	54%	7	16%
Estimated average variance extracted should exceed .50	2	3%	0	0%
The heterotrait-monotrait ratio of correlations, a model-free estimate of the factor correlation, should be smaller than .85 (HTMT)	4	5%	0	0%
Exploratory factor analysis extracts the hypothesized number of factors and/or does not estimate substantial cross-loadings	15	19%	0	0%
Focal constructs have incremental or different estimated associations with measures or manipulations of a third construct, or their association is moderated	24	30%	0	0%
Based on a multitrait-multimethod analysis, correlations between two measures of a construct that use different methods should (1) exceed correlations between measures that do not have a construct and a method in common, and (2) exceed correlations between measures of different constructs, using the same methods; and (3) correlations between measures of different constructs should be similar in size across methods	2	3%	0	0%

Panel B: Co-occurrence of focal discriminant validation criteria

	Phi	FL	HTMT	Other
Phi	47	26	2	28
FL	26	43	3	19
HTMT	2	3	4	1
Other	28	19	1	51

Notes: Panel A documents the use of criteria to assess discriminant validity in the 79 scale development articles. Percentages do not sum to 100% due to multiple criteria being used in the articles and rounding. In Panel B, off-diagonal entries denote the number of articles (out of 79) that report both criteria in the row and column. The diagonal corresponds to the number of times that a criterion was reported. Off-diagonal table entries do not sum to the number in the diagonal due to articles reporting fewer or more than two criteria.

If the restricted model fits substantially worse, discriminant validity is supported. The Phi criterion based on a confidence interval of the estimated factor correlation not overlapping one is reported in 16 articles (20%). Two articles (3%) report a *P*-value to test the factor correlation against one. These tests of the Phi criterion are (asymptotically) equivalent (Buse 1982). Taken together, 47 articles (59%) reported one or more of these variations of the Phi criterion. Forty-three articles (54%) reported the FL criterion, which is met if the squared factor correlation is smaller than the average variance extracted. None of these articles statistically tested the FL criterion. Together, the Phi and FL criteria are dominant: 64 articles (81%) reported at least one of the two criteria.

Two articles concluded discriminant validity based on the average variance extracted exceeding .50; this heuristic assesses whether the average item-reliability is adequate (Fornell and Larcker 1981), but it is not a test of discriminant validity. Four articles (5%) reported the heterotrait-monotrait (HTMT) criterion; all of these used a .85 comparison standard and none reported a statistical test. Fifteen articles (19%) used exploratory factor analysis to report evidence for discriminant validity; discriminant validity was judged to be met if the hypothesized number of factors was recovered, or substantial cross-loadings were absent. A total of 24 (30%) articles used nomological validation to demonstrate discriminant validity by showing that (a) a construct had explanatory power over and above another, related construct (Sechrest 1963), (b) two constructs correlated differently with a third construct (Franke et al. 2021), or (c) the relationship between two constructs was moderated by another construct (Tesser and Krauss 1976). Finally, one article reported the multitrait-multimethod (MTMM) criterion (Campbell and Fiske 1959), which is an indirect discriminant validation criterion because it requires the measurement of multiple constructs with multiple methods, which is uncommon. MTMM uses three subcriteria. First, the validity diagonal correlations (i.e., the correlations between the same construct measured with two different methods) should be

larger than the correlations between different constructs measured with different methods. Second, the validity diagonal correlations should be larger than the correlations between different constructs measured with the same method. Third, the pattern of intertrait correlations should be the same regardless of which method is used to measure the traits. One article used a criterion in the context of formative measurement referring to Klein and Rai (2009). Conceptually, this idea is very similar to the first and second MTMM subcriteria because discriminant validity holds if the intraconstruct item correlations are larger than the interconstruct item correlations.

Across all investigated articles, reported violations of the criteria were quite rare (8 or 12% out of 79), although publication bias could not be ruled out. In the eight articles in which a violation of discriminant validity occurred, seven reported at least one violation of the FL criterion and one reported that the Phi criterion was not met. The main text details and reanalyzes seven out of those eight articles for which sufficient data were available.

Given our finding that most articles report multiple criteria, the question remains which criteria were used in tandem. Panel B of Table WA1 summarizes the focal criteria in a co-occurrence matrix. This matrix has the number of articles in which a criterion is used on the diagonal, while the off-diagonal cells contain the number of articles that report the two criteria in the row and column. The Phi criterion co-occurs most commonly with the FL criterion (26 articles) or other criteria (28 articles). This suggests that authors (and/or reviewers) are often not satisfied when only the Phi criterion is met. Even though the FL criterion is quite strict, other criteria are still reported in 19 of the 43 articles that report the FL criterion. In terms of the HTMT criterion, three out of four articles that report HTMT also report the FL criterion. In sum, it is common to report both the Phi and FL criteria (33%) or at least one of them (81%).

Web Appendix B: Phi Criterion

This Web Appendix presents the results of a simulation study to explore the conditions under which the Phi criterion is (not) met. We generated data from a two-factor model with single-item scales that have equal reliability for simplicity but without loss of generality. The design had 20 (Factor correlation: 0 to .95 in steps of .05) × 4 (Reliability of both scales: .70, .80, .90, or 1) × 2 (Sample size [*n*]: 100 or 500) = 160 cells. Of primary interest are the asymptotic 95% confidence intervals of the estimated factor correlations. We estimated the confidence intervals using *R-lavaan* based on standard maximum likelihood estimation and Wald tests (R Core Team 2025; Rosseel 2012).

Figure WA1 shows the results for n = 100. The plot shows that, intuitively, the confidence intervals get narrower as the factor correlation increases (for all levels of scale reliability) and that for each level of factor correlation, the confidence intervals get wider as the reliabilities decrease. The horizontal dashed line indicates the fixed comparison standard of one for the Phi criterion, and the Phi criterion is met if the confidence interval of the factor correlation does not contain one. Thus, a factor correlation of .85 or higher is not statistically different from 1 when the reliabilities are .70 and a factor correlation of .95 or higher is not statistically different from one when the reliabilities are .80 or lower (assuming a standard 95% confidence level). However, factor correlations of up to .95 are statistically different from one for reliabilities of .90 or higher.

Thus, even a factor correlation as high as .95 is significantly different from one at a relatively small sample size of n =100 and for reliabilities of .90 or higher. On the one hand, this result is intuitive because in the absence of uncertainty due to measurement error, correlations cannot exceed the bound of one. On the other hand, while a statistical test may indicate that the two factors are not perfectly correlated and thus distinct, one may question whether two scales that are correlated .95 really measure constructs that are substantively and

practically distinct.

Figure WA2 shows the results for n = 500. As one would expect, the confidence intervals are much narrower across all levels of reliability when the sample size is n = 500 rather than n = 100. Now, a rather high correlation of .95 differs from one for reliabilities of .80 or higher. Even a factor correlation as high as .90 is distinct from one when reliabilities are .70, which is relatively low.

Against this background, it is not surprising that researchers have proposed criteria that are stricter than the Phi criterion, because the Phi criterion is almost always met for conventional reliabilities and sample sizes. The review by Voorhees et al. (2016) of 621 survey-based articles published in seven major marketing journals between 1996 and 2012 found that the Phi criterion was met in 99.8 percent of the cases. Similarly, the Phi criterion was met in 46 of the 47 articles in our review in which the criterion was used, see Table WA1 in Web Appendix A.

Figure WA1:
Leniency of the Phi Criterion (n = 100)

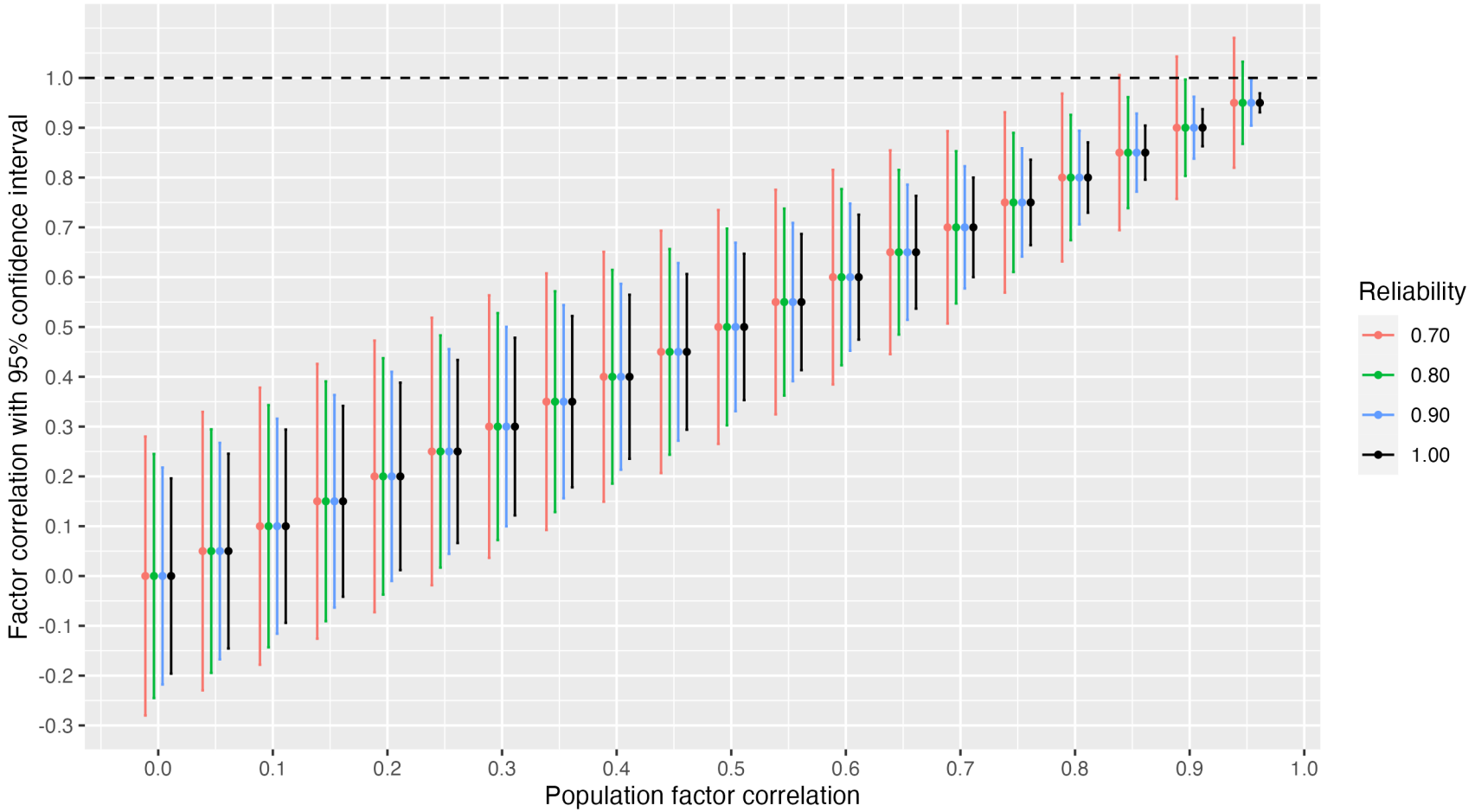
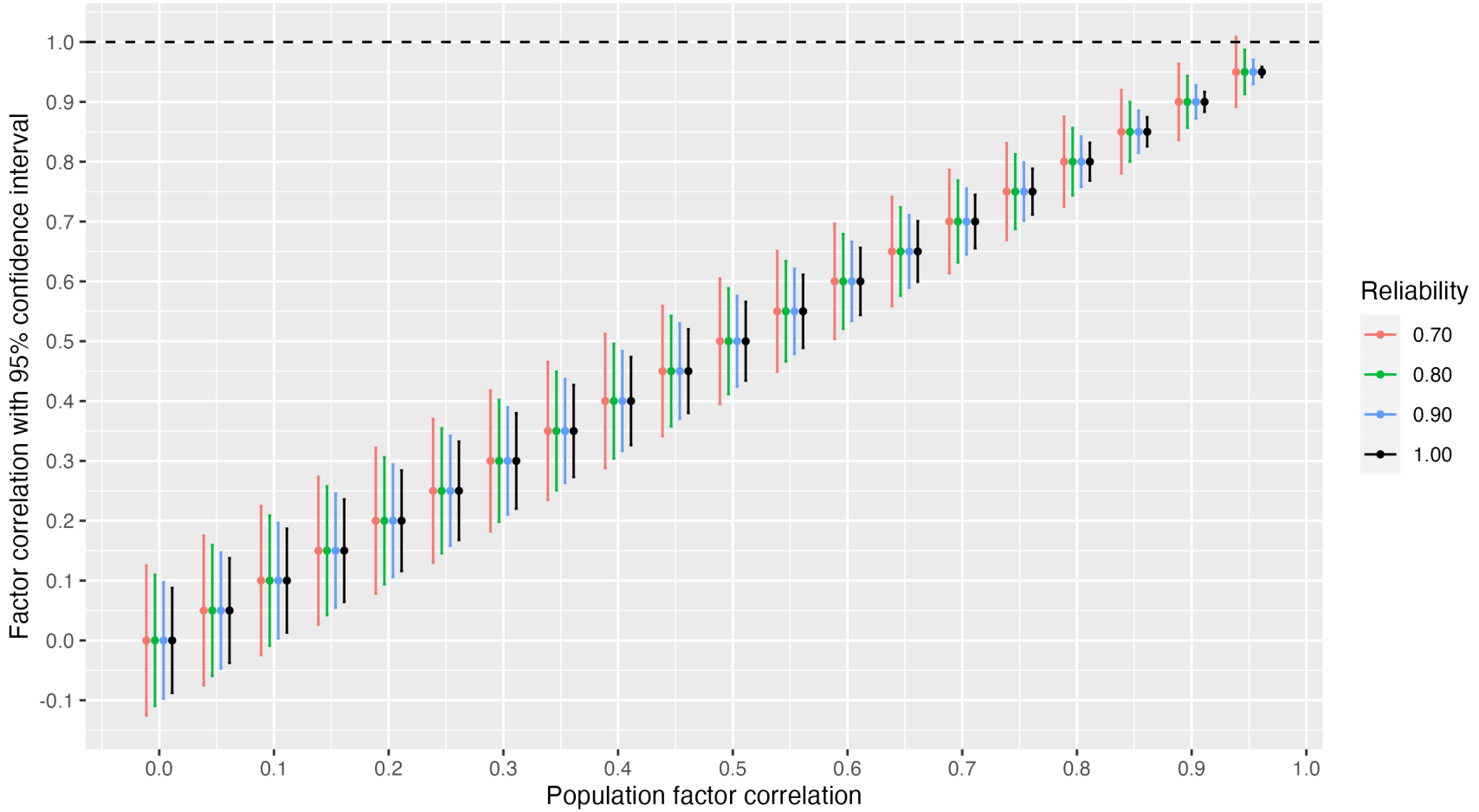


Figure WA2:
Leniency of the Phi Criterion (n = 500)



Web Appendix C: Comparison of Factor Correlation with HTMT

This Web Appendix shows that confirmatory factor analysis (CFA) is superior to the heterotrait-monotrait (HTMT) method for estimating the factor correlation. The HTMT ratio (HTMT for short) is a model-free method for estimating the factor correlation φ which cannot detect and does not account for systematic measurement error. In contrast, a correctly specified CFA does. We demonstrate this with simulated data generated with and without systematic measurement error. We subsequently demonstrate that the bias due to systematic measurement error is undetectable by HTMT, whereas it can be (a) diagnosed by inspecting the model fit of a CFA and (b) addressed with a correctly specified CFA.

We generated hypothetical data from two population models that differed in the absence vs. presence of systematic measurement error. The population models used a sample size of 200 and had two factors that correlated $\varphi = .80$. The two factors were both measured with four standardized and parallel items with loadings of .766. This generated factors with a scale reliability of .85, which is a good reliability typical in the marketing discipline (Peterson 1994; Pieters et al. 2022; Pieters 2017). The population model with systematic measurement error present had a third factor (e.g., a method factor) that loaded .50 on the first two items of both factors and -.50 on the remaining two items of both factors. This could reflect method variance due to, say, a mix of positively and negatively worded items, with 25% method variance in the items. This variance decomposition is based on a review of the empirical and methodological literature across domains (including marketing) in Lance et al. (2010), which estimated that method variance accounts for between 18% and 36% of item variance; 25% is in the center of this interval.

The simulated data were analyzed with three methods. The first method is a regular CFA without accounting for systematic measurement error. The second method uses the most recent version of the HTMT ratio (version 2) as recommended by Roemer et al. (2021) and

Ringle et al. (2023). It relaxes the parallel item assumption (i.e., that items are equally good measures of the underlying construct) of the initial version of the HTMT (Henseler et al. 2015; Rönkkö and Cho 2022). A third method uses a correctly specified bifactor CFA, which accounts for systematic measurement error by specifying a third factor and having all eight items freely load on this factor. We used *R* (R Core Team 2025) to generate the data and implemented all models in the *R-lavaan* library (Rosseel 2012) for a fair comparison.

We evaluated the results in terms of global model fit using the standard Chi-squared statistic (Bollen 1989), where a value of $\chi^2 = 0$ indicates a perfectly fitting model and larger values quantify the extent of model misfit. This information was unavailable for HTMT because it is a model-free method. We also assessed the ability of the methods to estimate the population factor correlation $\varphi = .80$ and quantified the percentage estimation bias as $\frac{\hat{\varphi}-\varphi}{\varphi} \times 100\%$, where $\hat{\varphi}$ is the estimated factor correlation and $\varphi = .80$ is the true correlation.

Panel A in Table WA2 has the results for the population model without systematic measurement error. The CFA fits perfectly ($\chi^2 = 0$) and estimates the true factor correlation of $\varphi = .80$ without bias. The HTMT also estimates the true factor correlation without bias, which is consistent with earlier findings that the HTMT is a consistent estimator of the factor

Table WA2:
HTMT Approximates the Factor Correlation but Cannot Detect Model Misspecification

	Model fit				Bias
Model	Number of parameters	Degrees of freedom	χ^2-statistic	$\hat{\varphi}$	$\hat{\varphi}$-bias (%)
Panel A: Absence of systematic measurement error					
1. CFA	17	19	0	.800	0
2. HTMT	-	-	-	.800	0
Panel B: Presence of systematic measurement error					
1. CFA (misspecified)	17	19	668.758	.850	6.250
2. HTMT (misspecified)	-	-	-	.871	8.875
3. CFA (correctly specified)	25	11	0	.800	0

Notes: Number of parameters and degrees of freedom sum to 36, the number of information elements in the population correlation matrix (8 variances and 28 covariances). Model fit information not available for the model-free HTMT, denoted by "-". Estimated factor correlation is $\hat{\varphi}$, with population $\varphi = .80$. The sample size was fixed to $n = 200$.

correlation when the underlying assumptions are satisfied (Henseler et al. 2015; Rönkkö and Cho 2022).

Panel B has the results for the model with systematic measurement error. The misspecified CFA without systematic measurement error accounted for fits poorly, with a Chi-squared statistic of 668.758. This suggests that a more complex measurement model is needed to account for the data. This poor fit also resulted in an overestimated factor correlation of .85, which is about 6% biased upward from the .80 population correlation. The HTMT was biased even more strongly (slightly less than 9%) based on an estimated factor correlation of .871. The difference with the CFA is that the possibility that the factor correlation is biased cannot be detected from model (mis)fit. In other words, the HTMT obfuscates that the underlying assumptions are not being met (i.e., the absence of a method factor or systematic measurement error). In contrast, the correctly specified CFA fit perfectly and the factor correlation was estimated without bias.

This example illustrates four key points. First, the HTMT cannot account for systematic measurement error, since it is a model-free estimate of the factor correlation. Second, the HTMT lacks an assessment of model fit and thus cannot assess to what extent the underlying assumptions are met. Third, in the presence of systematic measurement error that is unaccounted for (as in this example), both the regular CFA and HTMT are biased. However, the regular CFA can detect this misspecification with the Chi-squared statistic, a measure of model fit, which is unavailable for HTMT. Fourth, upon detecting misspecification of a regular CFA, the original model can be respecified and a correctly specified CFA accounts for systematic measurement error, which is also unavailable for the HTMT. In sum, CFA does, while HTMT does not, satisfy the first requirement that discriminant validation criteria should satisfy.

Web Appendix D: Comparison of CR with AVE

This Web Appendix builds on the analysis in the main text by further illustrating the differences between CR and AVE and the differences between the CR and FL criteria more generally. We want to emphasize that even though CR and AVE are strongly correlated, what matters is their numeric difference. CR as a measure of scale reliability is generally (quite a bit) larger than AVE as a measure of item reliability. Moreover, only CR (but not AVE) increases with the number of items in a scale.

The expressions for the average variance extracted (AVE) used by the FL criterion and the congeneric reliability (CR) used by the CR criterion based on unstandardized data are in the main text. For illustration purposes, the analysis here assumes two constructs measured with standardized parallel items (i.e., the items have equal loadings and unique factor variances or measurement errors). Under these assumptions, the expressions for CR and AVE are:

$$CR^* = \frac{(k\times\lambda^*)^2}{(k\times\lambda^*)^2+k\times\theta^*} = \frac{k\times(\lambda^*)^2}{1+(k-1)\times(\lambda^*)^2}, \quad \text{(WA1)}$$

$$AVE^* = IIR(x^*) = \frac{(\lambda^*)^2}{VAR(x^*)} = (\lambda^*)^2, \quad \text{(WA2)}$$

where for the k standardized parallel items (denoted by x^*), λ^* is the factor loading, θ^* is the unique factor variance, and IIR is the individual-item reliability. The asterisks indicate that the variables are standardized.

CR and AVE Are Correlated

Equation WA1 demonstrates that CR is a function of the (square of the sum of) factor loadings and the number of items. Equation WA2 demonstrates that when the items are standardized, AVE is equal to the (common) squared factor loading. Thus, the expressions for CR^* and AVE^* are based on the same factor loadings (and CR^* is also based on the number of items k), which suggests that their values should be highly correlated.

To examine this issue empirically, the correlation between CR^* and AVE^* across different values of λ^* ranging from $\sqrt{.50}$ = .71 to .99 in .01 increments is .995 (for k = 3 items), .993 (k = 4 items), .992 (k = 5 items), and .991 (k = 6 items). Thus, CR and AVE are indeed very highly correlated. However, what matters for discriminant validation is the difference between AVE and CR, which the following section explores.

CR is Larger than AVE

For standardized items, AVE^* as a function of CR^* for a scale of k items is:

$$AVE^* = CR^* + (k-1)\bar{r}_{ij} \times (CR^* - 1), \quad \text{(WA3)}$$

where $\bar{r}_{ij}$ is the average correlation of the standardized items. From Equation WA3 it becomes clear that CR and AVE are equal when CR equals zero or when it is one. This can only occur on the boundary, when all of the variance in the items is either due to unique factors (error) or the underlying substantive factors, respectively. In other situations, AVE is smaller than CR. If the items measure the same underlying factor, they should be positively correlated, i.e., $\bar{r}_{ij} > 0$. If this is the case, the second term on the right-hand side of Equation WA3 will be negative since $CR^* - 1$ is generally negative, which means that AVE is smaller than CR.

The Effect of Scale Length on CR and AVE

Equations WA1 and WA2 depict that CR depends on the number of items in a multi-item scale (denoted by k), while AVE does not. To illustrate, Figure WA3 visualizes CR (top plot) and AVE (middle plot) as a function of $(\lambda^*)^2$ ranging from 10 to .90 and for scales with k = 3 to 10 items. The figure also plots the difference between CR and AVE (bottom plot):

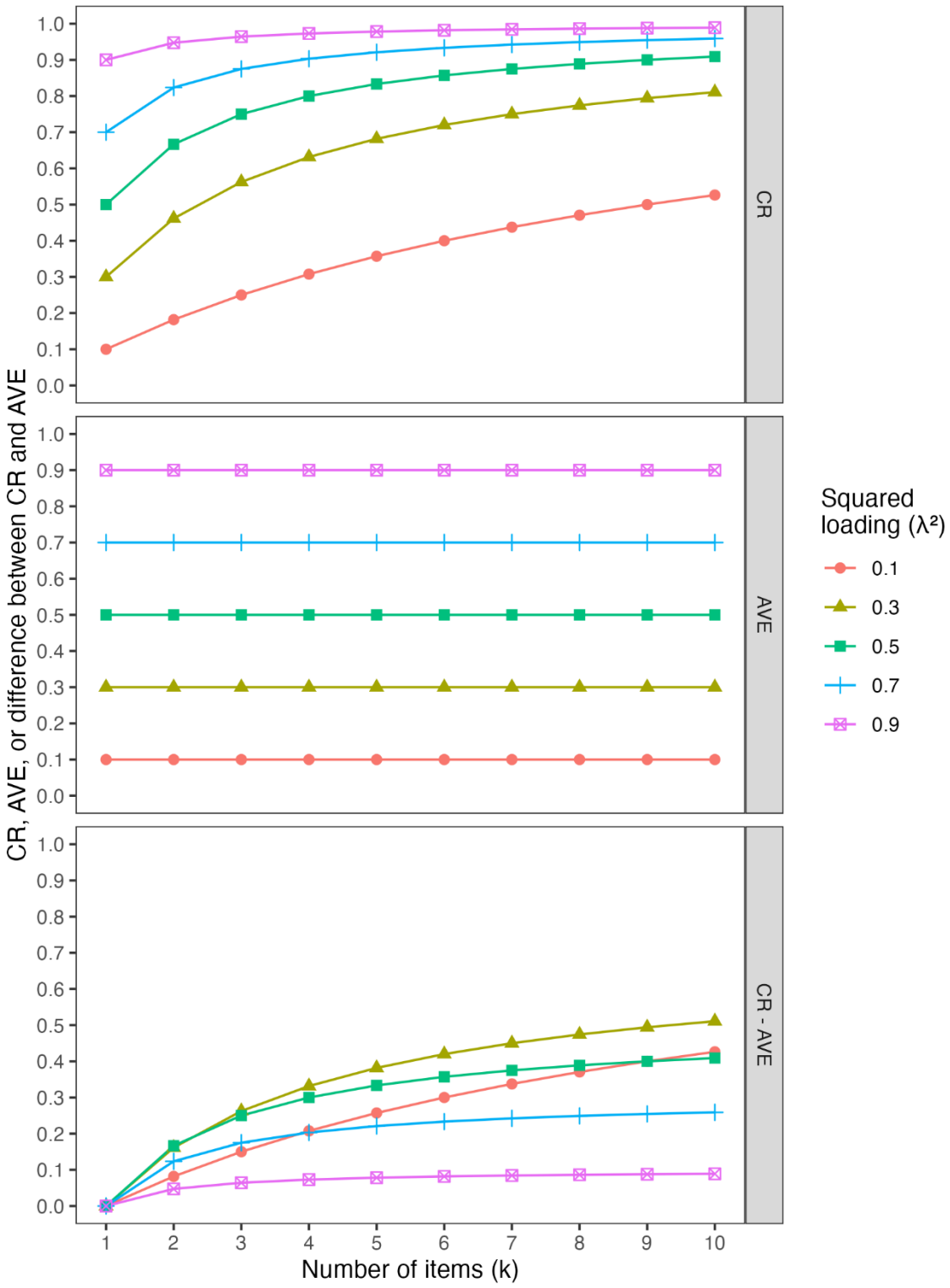
$$CR^* - AVE^* = \frac{k \times (\lambda^*)^2}{1 + (k-1) \times (\lambda^*)^2} - (\lambda^*)^2, \quad \text{(WA4)}$$

Of course, both average item (AVE) and scale reliability (CR) increase with the size of the item loadings. Moreover, the Figure illustrates that (under the focal conditions), holding

loadings constant, CR increases with the number of items, and that the relationship is stronger when items have smaller loadings (top plot). Longer scales generally have more systematic variance, which improves scale reliability. Moreover, AVE does not depend on the number of items (middle plot). Under the investigated standardized and parallel conditions, AVE is equal to the common squared loading. Once again, the difference between CR and AVE is always positive, and the difference is smallest when the loadings are very large and, across the investigated loadings, it increases when the number of items increases (bottom plot).

In sum, the derivations and simulations demonstrate that CR is almost always larger than AVE (unless the items are perfectly unreliable or perfectly reliable), which makes the FL criterion stricter than the CR criterion. When the factor correlation is larger, the likelihood that the CR criterion is met while the FL criterion is not met increases. When the factor correlation falls between the square roots of CR and AVE, the CR and FL criteria typically lead to different conclusions (as long as the sample size is large enough). The difference between CR and AVE (and hence between the corresponding CR and FL criteria) becomes larger when the number of items measuring a construct increases.

Figure WA3:
Difference between CR and AVE



Web Appendix E: Statistical Inference of the CR Criterion

This Web Appendix details statistical inference of the CR criterion. It discusses approaches to estimate a confidence interval of the discriminant validity index (DVI) as defined in the main text in an effort to meet the third requirement that discriminant validation criteria should satisfy (also discussed in the main text). The challenge is that the support for discriminant validity is not a model parameter but a non-linear function of the factor correlation and congeneric reliability, and congeneric reliability is also a non-linear function of the estimated loadings and standard errors. The remainder of this Web Appendix first summarizes three approaches to estimating a confidence interval for DVI. Simulations then compare the performance of the three approaches.

Wald Interval

The multivariate Delta method can be used to estimate a standard error of DVI, which can then be used to estimate a Wald confidence interval (e.g., Cheung 2009; Raykov 2002; Raykov and Marcoulides 2004). As described in the main text, suppose that $DVI_{CR_{F_1}} = \sqrt{CR}_{F_1} - \varphi_{F_1,F_2}$ is the support for discriminant validity of the correlation between F_1 and F_2 with respect to the CR of F_1 (the same logic applies to the CR of F_2). Based on the multivariate Delta method, the expression for the standard error ($SE[\cdot]$) is:

$$SE\left[DVI_{CR_{F_1}}\right] = \sqrt{D^2_{CR_{F_1}} \times SE\left[CR_{F_1}\right]^2 + D^2_{\varphi_{F_1,F_2}} \times SE\left[\varphi_{F_1,F_2}\right]^2 + 2 \times D_{CR_{F_1}} \times D_{\varphi_{F_1,F_2}} \times COV\left[CR_{F_1}, \varphi_{F_1,F_2}\right]}, \quad \text{(WA5)}$$

where $COV[\cdot]$ is a covariance and D refers to partial derivatives. The derivatives are:

$$D_{CR_{F_1}} = \frac{\partial DVI_{CR_{F_1}}}{\partial CR_{F_1}} = \frac{1}{2} \times CR_{F_1}^{-\frac{1}{2}}, \quad \text{(WA6)}$$

$$D_{\varphi_{F_1,F_2}} = \frac{\partial DVI_{CR_{F_1}}}{\partial \varphi_{F_1,F_2}} = -1. \quad \text{(WA7)}$$

The standard error of φ_{F_1,F_2} and the covariance of φ_{F_1,F_2} with CR_{F_1} are readily available from the parameter estimates based on maximum likelihood estimation (Bollen 1989; Yuan et al.

2010), and the Delta method also provides an estimate of $SE[CR_{F_1}]$ (see Cheung 2009, for details). Substitution leads to:

$$SE\left[DVI_{CR_{F_1}}\right] = \sqrt{\frac{SE[CR_{F_1}]^2}{4 \times CR_{F_1}} + SE\left[\varphi_{F_1,F_2}\right]^2 - \frac{COV_{CR_{F_1},\varphi_{F_1,F_2}}}{\sqrt{CR}_{F_1}}}. \quad \text{(WA8)}$$

Based on this standard error, and assuming asymptotic normality, a Wald test of DVI can be conducted (at the 5% or other predetermined α-level). The $(1-\alpha)$% Wald confidence interval is:

$$DVI_{CR_{F_1}} \pm Z_{1-\frac{\alpha}{2}} \times SE\left[DVI_{CR_{F_1}}\right], \quad \text{(WA9)}$$

where Z is the critical value of the standard normal distribution.

Although the Wald test and the corresponding confidence interval are based on asymptotic theory, are widely used, and are the default in statistical software packages (Cheung 2009; Raykov 2002; Raykov and Marcoulides 2004), there are alternative approaches to quantify the uncertainty of the DVI.

Profile Likelihood Interval

An alternative approach bases confidence intervals on the profile likelihood method (Cheung and Pesigan 2023). This method first specifies an unrestricted model that freely estimates the DVI as a function of parameter estimates that are obtained with maximum likelihood estimation. Let $\widehat{DVI}$ denote that maximum likelihood estimate. To estimate $\widehat{DVI}_L$, the lower bound of the confidence interval of $\widehat{DVI}$, the method specifies restricted models that systematically lower $\widehat{DVI}$ until the model fits substantially worse with a P-value that corresponds to the critical value at a $1-\alpha$ confidence level. To find this value of $\widehat{DVI}_L$, the method uses standard likelihood ratio tests with 1 degree of freedom (1 restriction): $2 \times (L_{Free} - L_{Restricted}) \sim \chi^2_{(1)}$, where Ls refer to the likelihoods of the free and restricted models with freely estimated $\widehat{DVI}$ and the fixed DVI respectively and $\chi^2_{(1)}$ is the critical value

of the Chi-squared distribution (Bollen 1989; Jöreskog 1971). The other parameters are estimated freely. The upper bound $\widehat{DVI}_U$ of the confidence interval of $\widehat{DVI}$ is found by doing the same thing, but now by specifying restricted models that systematically increase $\widehat{DVI}$. In other words, the profile likelihood confidence interval [$\widehat{DVI}_L$, $\widehat{DVI}_U$] is obtained so that the *P*-value of the likelihood ratio test is smaller than α for DVIs smaller than $\widehat{DVI}_L$ and larger than $\widehat{DVI}_U$. The *P*-values of this test are larger than α for DVIs within the interval [$\widehat{DVI}_L$, $\widehat{DVI}_U$]. The *P*-value is exactly α at the bounds of the interval. We refer to Cheung and Pesigan (2023) for implementation details.

The difference between the profile likelihood and Wald intervals is that the Wald interval computes a symmetric confidence interval by taking two points above and below $\widehat{DVI}$ based on the estimated standard error and the critical Z-value. The profile likelihood method searches for values of DVI above and below its estimate until the difference between likelihoods of the restricted model and the (maximum) likelihood of the unrestricted model is equal to the critical value of the Chi-squared distribution. Therefore, the confidence interval based on the profile likelihood can be asymmetric if the likelihood is asymmetric. This can occur, for example, in small samples and/or when the parameter of interest is skewed (Cheung and Pesigan 2023), which might be the case for DVI since it is a non-linear function of model parameters.

Bootstrap Interval

If there are doubts about the appropriateness of the normality assumption for the items, the distribution of the DVI could be estimated with resampling methods such as bootstrapping (Efron and Tibshirani 1994; Padilla and Veprinsky 2012). Bootstrapping uses empirical distributions of the DVI based on sampling from the original raw data with replacement, which is computationally more intensive than computing the Wald interval. A confidence interval at the desired α level is then based on the $\alpha/2$ and $1 - \alpha/2$ percentiles from this

empirical distribution. Bootstrapping requires raw data, although similar Monte Carlo intervals can be based on summary data (Preacher and Selig 2012).

Even though the Wald and profile likelihood methods are asymptotically equivalent (Buse 1982; Falk 2018), and the Bootstrapping approach might converge to the results from these methods in larger samples, we now compare the various approaches to estimate confidence intervals of the DVI under different conditions.

Empirical Comparison of Intervals

We conducted Monte Carlo simulations to empirically compare the three approaches to estimate a confidence interval around the support for discriminant validity because this comparison is difficult to do analytically.

The simulation design was 2 (Factor distribution: Normal [skewness = 0, kurtosis = 0] vs. non-normal [skewness = 3, kurtosis = 21]) × 7 (Factor correlation: .70 to 1 in steps of .05) × 7 (Congeneric reliability: .70 to 1 in steps of .05). This design varied the factor distribution because previous research found that factor non-normality was a key determinant of the accuracy of estimated standard errors and confidence intervals (Falk 2018; Nevitt and Hancock 2001; Pieters et al. 2022). The skewness and kurtosis values of the non-normal cells represent severe but realistic non-normality levels (Cain et al. 2017) and were based on previous simulations (Curran et al. 1996). The ranges of factor correlations and reliabilities reflect values that are common in empirical research and go all the way up to their limits. The sample size was fixed to 200, very close to the median sample size of 200.5 of the validation studies in the review of scale developments. For simplicity, all population models had two factors with three standardized and parallel items each.

We used *R* (R Core Team 2025) to generate 1,000 replications (datasets) per cell. Following Pieters et al. (2022), the Vale and Maurelli (1983) method implemented in the *semTools* library (Jorgensen et al. 2020) was used to generate non-normal factors. The *lavaan*

library (Rosseel 2012) enabled discriminant validation with Wald and percentage bootstrap intervals (we used 1,000 bootstrap resamples for each estimation, which trades off precision with computational burden), and the *semlbci* library estimated profile likelihood confidence intervals (Cheung and Pesigan 2023). We assessed the performance of the three approaches by calculating coverage of the 95% confidence interval of DVI (Muthén and Muthén 2002), or the proportion of datasets in which the estimated confidence intervals contained the true DVI.

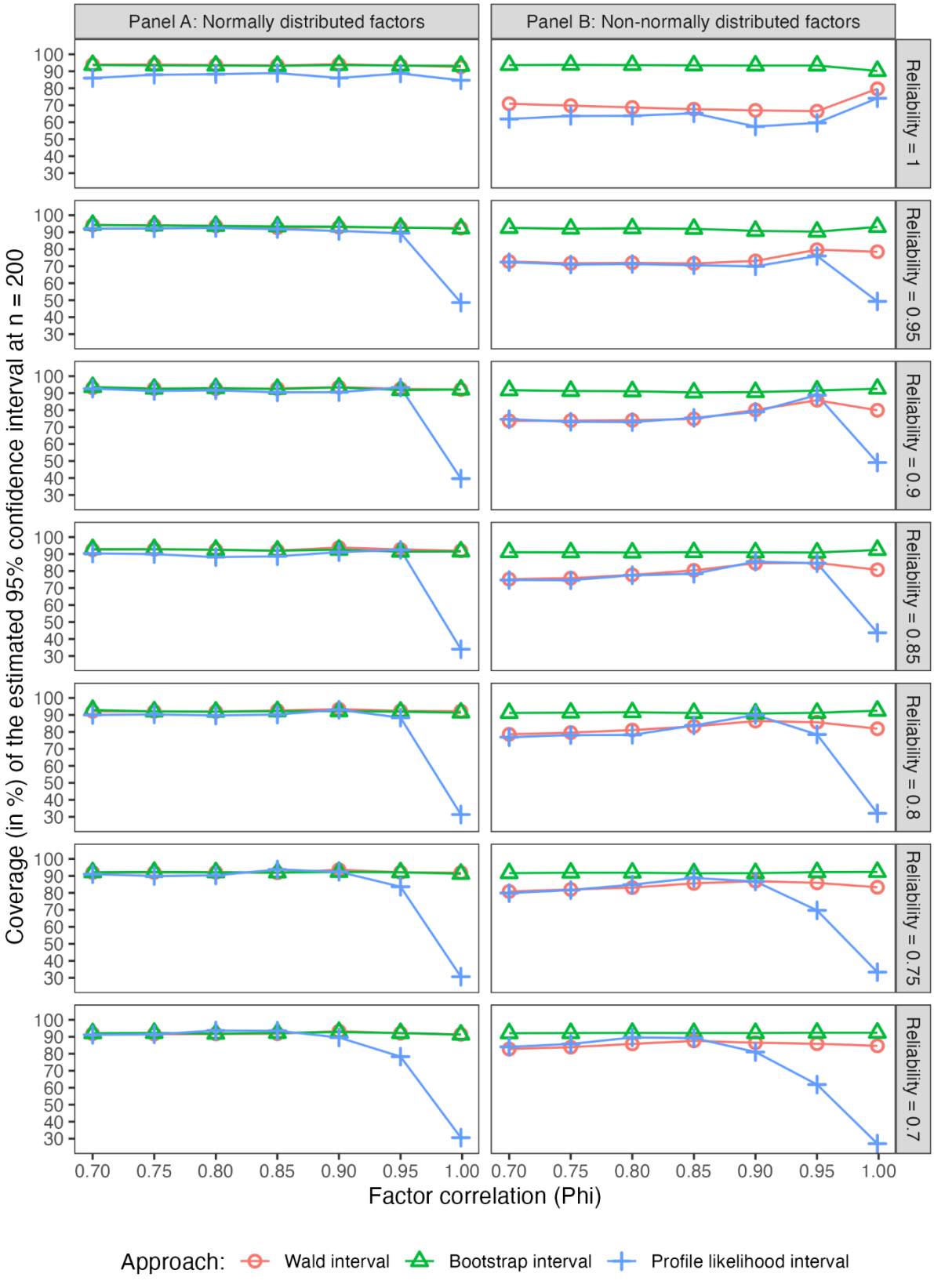
Figure WA4 visualizes the results. Panel A has the results for normally distributed factors and for reliabilities between 1 (top plot) and .70 (bottom plot). Across factor correlations and reliabilities, the Wald and bootstrap intervals have very similar and high coverage levels of about 95%. Thus, these approaches estimate an accurate confidence interval of DVI across the investigated conditions. The performance of the profile likelihood approach depends on the factor correlation and reliability. At smaller factor correlations of, say, .80, the profile likelihood approach has very similar coverage levels as the Wald and bootstrap approaches. When the factor correlation is very high, the coverage decreases. For example, the coverage rates were 78% and 31%, respectively, for factor correlations of .95 and 1 (for factor reliabilities of .70). These results are consistent with earlier findings that the test-statistic of the likelihood ratio test is not Chi-square distributed when parameters are at the boundary of the parameter space, which biases the likelihood-ratio test (Stoel et al. 2006). Under the focal conditions, this occurs for larger correlations because these intervals are more likely to approach the unity upper bound of the correlation scale, and for lower reliabilities because these yield wider confidence intervals that cover one. Thus, for normally distributed factors, the performance of Wald and Bootstrap intervals is best.

Panel B of Figure WA4 has the results for non-normally distributed factors. Across conditions, the bootstrap approach outperforms the Wald and profile likelihood intervals,

with high coverage rates close to 95%. These results reflect the fact that the multivariate normality assumptions of the asymptotic approaches are violated, while the empirical distribution used by the bootstrap is robust to non-normality. For example, coverage probabilities are 68% (Wald), 94% (bootstrap), and 65% (profile likelihood) at factor reliabilities of 1 (top plot) and a .85 factor correlation. Because the population model was specified to generate non-normally distributed factors and indicators based on these factors with normally distributed measurement errors, differences between methods are smaller at lower reliabilities because indicators are less severely non-normally distributed when more normally distributed measurement error is added. For instance, the coverage probabilities are 88% (Wald), 92% (bootstrap) and 89% (profile likelihood) at factor reliabilities of .70 (bottom plot) and at the same .85 factor correlation.

Overall, collapsing across conditions, the average coverage rates are 86% (Wald), 92% (bootstrap), and 83% (Profile likelihood). Thus, even though the differences between approaches are smaller for normally distributed factors, the bootstrap interval is the confidence interval with the best coverage across all the investigated conditions.

Figure WA4:
The Bootstrap Interval Has Equal Coverage as the Wald Interval Has for Normally Distributed Factors but Better Coverage for Non-Normally Distributed Factors



Web Appendix F: Implementation

This Web Appendix contains implementation details. We implement the procedure proposed in the main text by providing: (1) code for the two steps in the proposed procedure, (2) a function that automates these steps, and (3) an online *Shiny* app that uses menus and summary data rather than code (under somewhat restrictive assumptions as described in more detail in the main text and in the app).

All implementations use *R* (R Core Team 2025) and the *lavaan* library (Rosseel 2012). We use the Wilson and Bellezza (2022) case study presented in the main text as an illustrative example. As in the main text, we focus on the validation of Number of Possessions (NoP) with Sparse Aesthetics (SA) for brevity, but all code files validate the Mindfully Curated Consumption (MCC) subscale as well. All results support and strengthen the discriminant validity conclusion of the original article. Data and annotated code files are available on *OSF* (https://osf.io/57srv/), and the *Shiny* app is also hosted at https://constantpieters.shinyapps.io/discriminantvalidation. The remainder of this section shows selected output.

R *code*

Panel A of Figure WA5 has relevant output from the *R* code to implement Step 1 of the proposed procedure. It first shows the estimated factor correlations. It then reports estimates of DVI_1, the discriminant validity index of the Phi criterion. The output contains 95% percentile bootstrap confidence intervals bounded by `ci.lower` and `ci.upper` and bootstrapped *P*-values, as described in the main text. The estimated factor correlation between NoP and SA is .780, and the DVI_1 is a positive .220 with 95% bootstrap confidence interval [.148, .307], whose lower bound is quite different from zero, and a small p-value < .001, which implies that the Phi criterion is met.

Panel B implements Step 2. The results show that the CR of NoP is .884, and the CR of SA is .842. Then, the DVI of NoP with SA with respect to the CR of NoP is .160 with 95% bootstrap confidence interval [.088, .245] and $p < .001$. The DVI with respect to the CR of SA is .138 with 95% bootstrap confidence interval [.064, .223] and $p < .001$. The DVIs are positive, the lower bounds of the confidence intervals are quite different from zero, and the *p*-values are small, which means that the CR criterion is met for both factors.

R *function*

The required code to use the *R* function is much simpler (not reproduced here for brevity; see *OSF*). Figure WA6 has output. The results are identical to the results from the *R* code reported in Figure WA5. The difference is that the output is reported for each pair separately and the results for each step are formatted and annotated. The summary of results can therefore be readily copied into analysis reports.

Shiny *app*

The *Shiny* app enables discriminant validation based on summary data, but under the assumption that the items are standardized and parallel. As visualized in Figure WA7, the output features a hypothetical example for illustration purposes in which two constructs F_1 and F_2 are measured with three items with scale reliabilities of .70 and .90, respectively. The multi-item scale scores are correlated .60 at $n = 250$ observations. Figure WA8 has annotated output. Again, the formatting of the output allows the results to be readily copied into analysis reports.

Figure WA5:
Selected Output from *R* Code

Panel A: Step 1 of the proposed procedure

	Estimate	ci.lower	ci.upper	*p*-value
Phi_NoPSA	0.780	0.693	0.852	0.000
Phi_NoPMCC	0.582	0.481	0.674	0.000
Phi_SAMCC	0.449	0.337	0.555	0.000
DVI_1_NoPSA	0.220	0.148	0.307	0.000
DVI_1_NoPMCC	0.418	0.326	0.519	0.000
DVI_1_SAMCC	0.551	0.445	0.663	0.000

Panel B: Step 2 of the proposed procedure

	Estimate	ci.lower	ci.upper	*p*-value
CR_NoP	0.884	0.859	0.905	0.000
CR_SA	0.842	0.807	0.870	0.000
CR_MCC	0.884	0.859	0.905	0.000
DVI_CR_NoP_NoPSA	0.160	0.088	0.245	0.000
DVI_CR_SA_NoPSA	0.138	0.064	0.223	0.000
DVI_CR_NoP_NoPMCC	0.358	0.267	0.459	0.000
DVI_CR_MCC_NoPMCC	0.358	0.269	0.456	0.000
DVI_CR_SA_SAMCC	0.468	0.362	0.578	0.000
DVI_CR_MCC_SAMCC	0.491	0.388	0.600	0.000

Figure WA6:
Selected Output from *R* Function

```
-------------------------------------------------------
Discriminant validation of 'NoP' with 'SA'
-------------------------------------------------------

# Discriminant validation components

  Factor correlation (Phi):

                 Estimate ci.lower ci.upper p-value
  Phi               0.780    0.693    0.852   0.000

  Congeneric reliability (CR):

                 Estimate ci.lower ci.upper p-value
  CR_NoP            0.884    0.859    0.905   0.000
  CR_SA             0.842    0.807    0.870   0.000

# Discriminant validation

  Step 1: Phi criterion

                 Estimate ci.lower ci.upper p-value
  DVI_1             0.220    0.148    0.307   0.000

  Step 2: CR criterion

                 Estimate ci.lower ci.upper p-value
  DVI_CR_NoP        0.160    0.088    0.245   0.000
  DVI_CR_SA         0.138    0.064    0.223   0.000

-------------------------------------------------------
```

Figure WA7
Annotated *Shiny* App Input

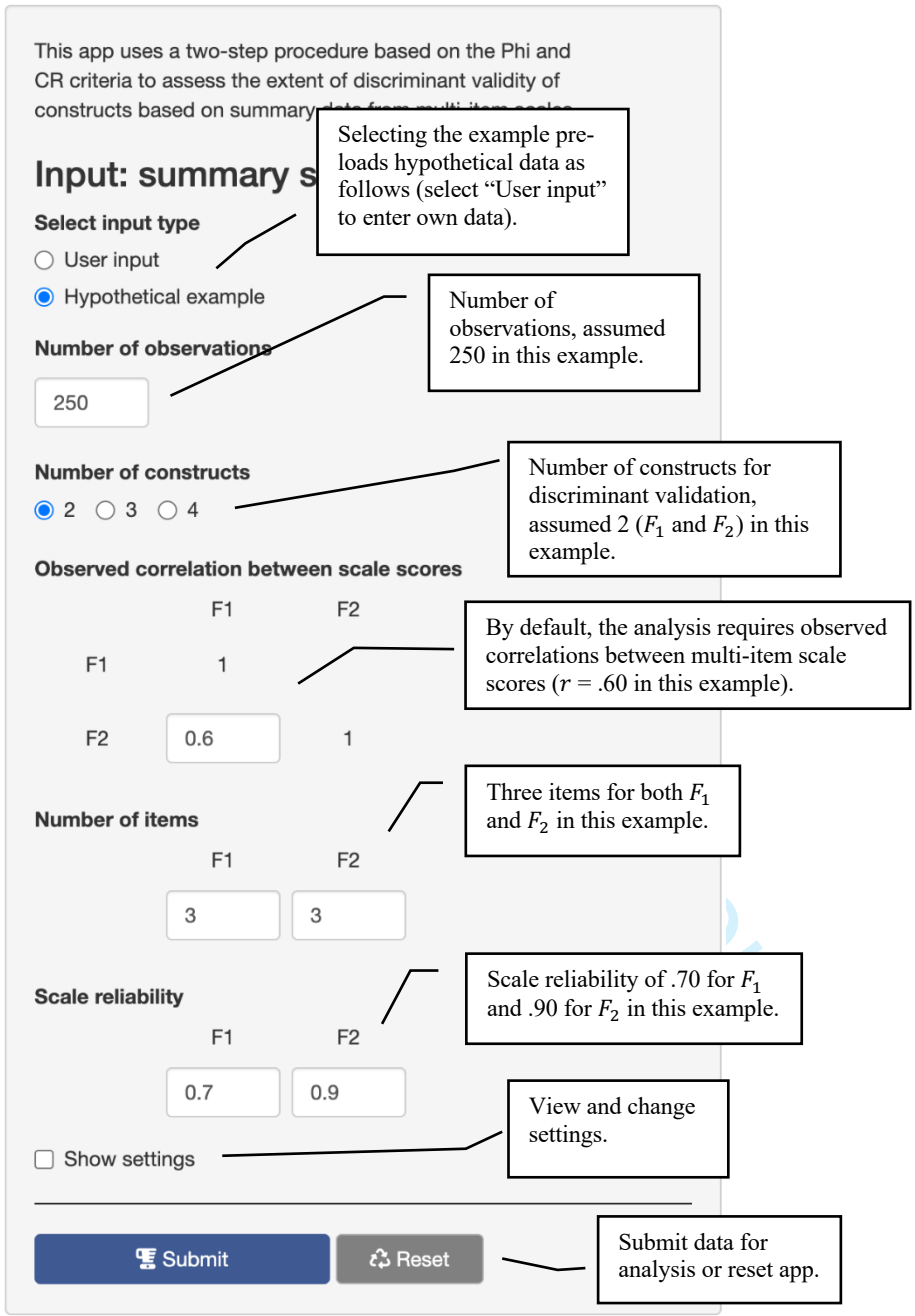


Figure WA8
Annotated *Shiny* App Output

```
============================================================
Discriminant validation based on Phi & CR criteria
============================================================

# Summary statistics data (number of observations = 250)

  Construct   # items     CR     F1     F2
         F1         3   0.70      1   0.76
         F2         3   0.90   0.60      1

------------------------------------------------------------
Discriminant validation of 'F1' with 'F2'
------------------------------------------------------------

# Discriminant validation components

  Factor correlation (Phi):

                    Est.           95% CI   p-value
    Phi             0.76    [0.67, 0.84]      0.00

  Congeneric reliability (CR):

                    Est.           95% CI   p-value
    CR_F1           0.70    [0.64, 0.75]      0.00
    CR_F2           0.90    [0.88, 0.91]      0.00

# Discriminant validation

  Step 1: Phi criterion

                    Est.           95% CI   p-value
    DVI_1           0.24    [0.16, 0.33]      0.00

  Step 2: CR criterion

                    Est.           95% CI   p-value
    DVI_CR_F1       0.08   [-0.02, 0.18]      0.06
    DVI_CR_F2       0.19    [0.11, 0.28]      0.00

============================================================
```

Summary of hypothetical input. The upper diagonal of the correlation matrix refers to the estimated factor correlation of F_1 with F_2, which is $\varphi = .76$.

Estimated factor correlation of F_1 with F_2 is $\varphi = .76$, with 95% bootstrap confidence interval and p-value reported.

Estimated multi-item scale (congeneric) reliability of F_1 is .70. Estimated scale reliability of F_2 is .90. Moreover, 95% bootstrapped confidence intervals and p-values are reported.

Step 1: DVI for the Phi criterion is .24 with 95% bootstrapped confidence interval [.16, .33] quite a bit larger than zero, and the p-value is very small, thus meeting the Phi criterion.

Step 2: DVIs for both F_1 and F_2 are positive. The 95% bootstrapped confidence interval of the DVI with respect to the square root of the CR of F_1 overlaps zero slightly, and the p-value is relatively large. For F_2, DVI is larger. Thus, these results suggest that the CR criterion is only partially met.

Overall, there is strong support for the Phi criterion, and evidence that the CR criterion is partially met (Conclusion 3).

Web Appendix G: Detailed Empirical Results

This Web Appendix reports detailed results of the empirical applications reported in the main text.

Case Study

Table WA3:
Discriminant Validation of Number of Possessions (NoP) and Sparse Aesthetics (SA) Based on Wilson and Bellezza (2022, Study 3)

Discriminant validity index (DVI)	Est.	90% CI	95% CI	99% CI	*p*-value
Phi criterion					
DVI_1	.220	[.158, .292]	[.148, .307]	[.130, .337]	< .001
CR criterion					
$DVI_{CR_{NoP}}$	.160	[.099, .230]	[.088, .245]	[.069, .275]	< .001
$DVI_{CR_{SA}}$	.138	[.076, .209]	[.064, .223]	[.043, .254]	< .001
HTMT criterion with .85 standard					
$DVI_{HTMT_{.85}}$	.052	[-.0003, .114]	[-.009, .126]	[-.026, .155]	.051
FL criterion					
$DVI_{AVE_{NoP}}$	.030	[-.033, .101]	[-.044, .115]	[-.065, .144]	.227
$DVI_{AVE_{SA}}$	-.022	[-.090, .052]	[-.101, .068]	[-.127, .095]	.299
Other fixed comparison standards					
$DVI_{.95}$	.170	[.108, .242]	[.098, .257]	[.080, .287]	< .001
$DVI_{.90}$	.120	[.058, .192]	[.048, .207]	[.030, .237]	< .001
$DVI_{.85}$	.070	[.008, .142]	[-.002, .157]	[-.020, .187]	.029
$DVI_{.80}$	.020	[-.042, .092]	[-.052, .107]	[-.070, .137]	.316
$DVI_{.71}$	-.070	[-.132, .002]	[-.142, .017]	[-.160, .047]	.055
$DVI_{.70}$	-.080	[-.142, -.008]	[-.152, .007]	[-.170, .037]	.035

Notes: Discriminant validation results are based on raw data from Wilson and Bellezza (2022, Study 3). *OSF* has data and analysis code. The focal subscales of the Minimalist Consumer Scale are Number of Possessions (NoP) and Sparse Aesthetic (SA). The factor correlation is an estimated φ = .780 with 90% CI [.708, .842], 95% CI [.693, .852] and .99% CI [.663, .870]. The estimated HTMT = .798 with 90% CI [.736, .850], 95% CI [.724, .859], and 99% CI [.695, .876]. Average variance extracted (AVE) estimates are .656 (NoP) and .574 (SA), and congeneric reliability (CR) estimates are .884 (NoP) and .842 (SA). The sample size is n = 394. Est. is the estimate of the discriminant validity index (DVI) and the next columns denote 90%, 95%, and 99% percentile bootstrap confidence intervals based on 10,000 resamples; the reported *p*-value is the corresponding (one-sided) bootstrap *p*-value of the test that the DVI equals zero, as described in the main text.

Reanalysis of Seven Articles Reporting Discriminant Validity Violations

Table WA4:
Phi, HTMT, FL and CR Criteria Results for the Reanalysis of Seven Articles Reporting Discriminant Validity Violations

Row	Phi criterion: DVI_1 [95% CI] (p)	HTMT criterion: $DVI_{HTMT_{.85}}$ [95% CI] (p)	FL criterion: DVI_{AVE} [95% CI] (p)	CR criterion: DVI_{CR} [95% CI] (p)
1	.070 [-.022, .164] (.065)	-.080 [-.188, .029] (.068)	-.228 [-.335, -.118] (< .001)	-.039 [-.136, .060] (.218)
			-.297 [-.416, -.179] (< .001)	-.078 [-.186, .027] (.071)
2	.120 [-.013, .255] (.038)	-.030 [-.193, .150] (.389)	-.247 [-.397, -.094] (.001)	-.028 [-.171, .114] (.346)
			-.253 [-.420, -.083] (.001)	-.071 [-.229, .085] (.182)
3	.130 [.009, .248] (.017)	-.020 [-.171, .136] (.407)	-.168 [-.301, -.037] (.006)	.021 [-.104, .143] (.376)
			-.243 [-.396, -.087] (.001)	-.061 [-.205, .080] (.199)
4	.083 [.032, .134] (.001)	-.067 [-.121, -.012] (.009)	-.166 [-.229, -.102] (< .001)	-.025 [-.083, .031] (.190)
			-.169 [-.232, -.106] (< .001)	-.027 [-.084, .028] (.173)
5	.110 [.100, .121] (< .001)	-.040 [-.051, -.028] (< .001)	-.103 [-.115, -.090] (< .001)	.063 [.052, .074] (< .001)
			-.192 [-.203, -.181] (< .001)	.085 [.074, .095] (< .001)
6	.117 [.089, .144] (< .001)	-.033 [-.063, -.002] (.019)	-.106 [-.141, -.073] (< .001)	.022 [-.009, .051] (.081)
			-.089 [-.120, -.060] (< .001)	.063 [.034, .090] (< .001)
7	.121 [.099, .142] (< .001)	-.029 [-.052, -.007] (.005)	-.108 [-.135, -.082] (< .001)	-.016 [-.040, .008] (.104)
			-.118 [-.145, -.091] (< .001)	-.023 [-.047, .002] (.037)
8	.138 [.124, .152] (< .001)	-.012 [-.026, .002] (.049)	-.100 [-.117, -.084] (< .001)	-.005 [-.021, .010] (.259)
			-.110 [-.127, -.093] (< .001)	-.012 [-.028, .004] (.065)
9	.180 [.166, .194] (< .001)	.030 [.015, .047] (< .001)	-.014 [-.029, .001] (.036)	.138 [.124, .152] (< .001)
			-.081 [-.095, -.067] (< .001)	.160 [.146, .174] (< .001)
10	.245 [.203, .287] (< .001)	.095 [.051, .143] (< .001)	.022 [-.024, .067] (.178)	.150 [.107, .193] (< .001)
			-.172 [-.216, -.129] (< .001)	.167 [.125, .209] (< .001)
11	.500 [.407, .593] (< .001)	.350 [.269, .487] (< .001)	.214 [.120, .307] (< .001)	.462 [.369, .554] (< .001)
			-.020 [-.114, .073] (.327)	.426 [.333, .518] (< .001)
12	.510 [.411, .608] (< .001)	.360 [.263, .502] (< .001)	.322 [.223, .421] (< .001)	.489 [.390, .587] (< .001)
			-.043 [-.145, .058] (.214)	.384 [.284, .482] (< .001)
13	.530 [.272, .785] (< .001)	.380 [.099, .387] (.002)	.122 [-.137, .377] (.179)	.431 [.172, .683] (.001)
			-.170 [-.430, .085] (.095)	.395 [.134, .647] (.002)

Note: Row numbers correspond to those in Table 2 in the main text, which this table supplements.

Table WA5:
Fixed Standard Results for the Reanalysis of Seven Articles Reporting Discriminant Validity Violations

Row	Fixed standard of .95: $DVI_{.95}$ [95% CI] (p)	Fixed standard of .90: $DVI_{.90}$ [95% CI] (p)	Fixed standard of .85: $DVI_{.85}$ [95% CI] (p)	Fixed standard of .80: $DVI_{.80}$ [95% CI] (p)	Fixed standard of .71: $DVI_{.71}$ [95% CI] (p)	Fixed standard of .70: $DVI_{.70}$ [95% CI] (p)
1	.020 [-.072, .114] (.329)	-.030 [-.122, .064] (.260)	-.080 [-.172, .014] (.045)	-.130 [-.222, -.036] (.002)	-.220 [-.312, -.126] (< .001)	-.230 [-.322, -.136] (< .001)
2	.070 [-.063, .205] (.151)	.020 [-.113, .155] (.381)	-.030 [-.163, .105] (.331)	-.080 [-.213, .055] (.116)	-.170 [-.303, -.035] (.006)	-.180 [-.313, -.045] (.004)
3	.080 [-.041, .198] (.097)	.030 [-.091, .148] (.313)	-.020 [-.141, .098] (.372)	-.070 [-.191, .048] (.129)	-.160 [-.281, -.042] (.004)	-.170 [-.291, -.052] (.003)
4	.033 [-.018, .084] (.099)	-.017 [-.068, .034] (.263)	-.067 [-.118, -.016] (.005)	-.117 [-.168, -.066] (< .001)	-.207 [-.258, -.156] (< .001)	-.217 [-.268, -.166] (< .001)
5	.060 [.050, .071] (< .001)	.010 [-.0005, .021] (.031)	-.040 [-.050, -.029] (< .001)	-.090 [-.100, -.079] (< .001)	-.180 [-.190, -.169] (< .001)	-.190 [-.200, -.179] (< .001)
6	.067 [.039, .094] (< .001)	.017 [-.011, .044] (.113)	-.033 [-.061, -.006] (.008)	-.083 [-.111, -.056] (< .001)	-.173 [-.201, -.146] (< .001)	-.183 [-.211, -.156] (< .001)
7	.071 [.049, .092] (< .001)	.021 [-.001, .042] (.028)	-.029 [-.051, -.008] (.004)	-.079 [-.101, -.058] (< .001)	-.169 [-.191, -.148] (< .001)	-.179 [-.201, -.158] (< .001)
8	.088 [.074, .102] (< .001)	.038 [.024, .052] (< .001)	-.012 [-.026, .002] (.043)	-.062 [-.076, -.048] (< .001)	-.152 [-.166, -.138] (< .001)	-.162 [-.176, -.148] (< .001)
9	.130 [.116, .144] (< .001)	.080 [.066, .094] (< .001)	.030 [.016, .044] (< .001)	-.020 [-.034, -.006] (.003)	-.110 [-.124, -.096] (< .001)	-.120 [-.134, -.106] (< .001)
10	.195 [.153, .237] (< .001)	.145 [.103, .187] (< .001)	.095 [.053, .137] (< .001)	.045 [.003, .087] (.017)	-.045 [-.087, -.003] (.018)	-.055 [-.097, -.013] (.005)
11	.450 [.357, .543] (< .001)	.400 [.307, .493] (< .001)	.350 [.257, .443] (< .001)	.300 [.207, .393] (< .001)	.210 [.117, .303] (< .001)	.200 [.107, .293] (< .001)
12	.460 [.361, .558] (< .001)	.410 [.311, .508] (< .001)	.360 [.261, .458] (< .001)	.310 [.211,.408] (< .001)	.220 [.121, .318] (< .001)	.210 [.111, .308] (< .001)
13	.480 [.222, .735] (< .001)	.430 [.172, .685] (.001)	.380 [.122, .635] (.002)	.330 [.072, .585] (.006)	.240 [-.018, .495] (.032)	.230 [-.028, .485] (.038)

Note: Row numbers correspond to those in Table 2 in the main text, which this table supplements.

References of the Web Appendix

Baumgartner, Hans and Bert Weijters (2019), "Measurement in Marketing," *Foundations and Trends® in Marketing*, 12 (4), 278-400.

Bollen, Kenneth A. (1989), Structural Equations with Latent Variables. New York: Wiley.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36 (3), 153-57.

Cain, Meghan K., Zhiyong Zhang, and Ke-Hai Yuan (2017), "Univariate and Multivariate Skewness and Kurtosis for Measuring Nonnormality: Prevalence, Influence and Estimation," *Behavior Research Methods*, 49 (5), 1716-35.

Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (2), 81-105.

Cheung, Mike W. L. (2009), "Constructing Approximate Confidence Intervals for Parameters with Structural Equation Models," *Structural Equation Modeling: A Multidisciplinary Journal*, 16 (2), 267-94.

Cheung, Shu Fai and Ivan Jacob Agaloos Pesigan (2023), "semlbci: An R Package for Forming Likelihood-Based Confidence Intervals for Parameter Estimates, Correlations, Indirect Effects, and Other Derived Parameters," *Structural Equation Modeling: A Multidisciplinary Journal*, 30 (6), 985-99.

Churchill, Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (1), 64-73.

Curran, Patrick J., Stephen G. West, and John F. Finch (1996), "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis," *Psychological Methods*, 1 (1), 16-29.

Efron, Bradley and Robert J. Tibshirani (1994), An Introduction to the Bootstrap. NY: Chapman and Hall.

Falk, Carl F. (2018), "Are Robust Standard Errors the Best Approach for Interval Estimation with Nonnormal Data in Structural Equation Modeling?," *Structural Equation Modeling: A Multidisciplinary Journal*, 25 (2), 244-66.

Fornell, Claes and David F. Larcker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (1), 39-50.

Franke, George R., Marko Sarstedt, and Nicholas P. Danks (2021), "Assessing Measure Congruence in Nomological Networks," *Journal of Business Research*, 130, 318-34.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," *Journal of Marketing*

Research, 25 (2), 186-92.

Haws, Kelly L., Kevin L. Sample, and John Hulland (2023), "Scale Use and Abuse: Towards Best Practices in the Deployment of Scales," *Journal of Consumer Psychology*, 33 (1), 226-43.

Henseler, Jörg, Christian M. Ringle, and Marko Sarstedt (2015), "A New Criterion for Assessing Discriminant Validity in Variance-Based Structural Equation Modeling," *Journal of the Academy of Marketing Science*, 43 (1), 115-35.

Jöreskog, Karl G. (1971), "Statistical Analysis of Sets of Congeneric Tests," *Psychometrika*, 36 (2), 109-33.

Jorgensen, Terrence D., Sunthud Pornprasertmanit, Alexander M. Schoemann, and Yves Rosseel (2020), "semTools: Useful Tools for Structural Equation Modeling," [available at https://CRAN.R-project.org/package=semTools].

Klein, Richard and Arun Rai (2009), "Interfirm Strategic Information Flows in Logistics Supply Chain Relationships," *MIS Quarterly*, 33 (4), 735-62.

Lance, Charles E., Bryan Dawson, David Birkelbach, and Brian J. Hoffman (2010), "Method Effects, Measurement Error, and Substantive Conclusions," *Organizational Research Methods*, 13 (3), 435-55.

Muthén, Linda K. and Bengt O. Muthén (2002), "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power," *Structural Equation Modeling: A Multidisciplinary Journal*, 9 (4), 599-620.

Nevitt, Jonathan and Gregory R. Hancock (2001), "Performance of Bootstrapping Approaches to Model Test Statistics and Parameter Standard Error Estimation in Structural Equation Modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, 8 (3), 353-77.

Padilla, Miguel A. and Anna Veprinsky (2012), "Correlation Attenuation Due to Measurement Error: A New Approach Using the Bootstrap Procedure," *Educational and Psychological Measurement*, 72 (5), 827-46.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," *Journal of Consumer Research*, 21 (2), 381-91.

Pieters, Constant, Rik Pieters, and Aurélie Lemmens (2022), "Six Methods for Latent Moderation Analysis in Marketing Research: A Comparison and Guidelines," *Journal of Marketing Research*, 59 (5), 941-62.

Pieters, Rik (2017), "Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication," *Journal of Consumer Research*, 44 (3), 692-716.

Preacher, Kristopher J. and James P. Selig (2012), "Advantages of Monte Carlo Confidence Intervals for Indirect Effects," *Communication Methods and Measures*, 6 (2), 77-98.

R Core Team (2025), "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Raykov, Tenko (2002), "Analytic Estimation of Standard Error and Confidence Interval for Scale Reliability," *Multivariate Behavioral Research*, 37 (1), 89-103.

Raykov, Tenko and George A. Marcoulides (2004), "Using the Delta Method for Approximate Interval Estimation of Parameter Functions in SEM," *Structural Equation Modeling*, 11 (4), 621-37.

Ringle, Christian M., Marko Sarstedt, Noemi Sinkovics, and Rudolf R. Sinkovics (2023), "A Perspective on Using Partial Least Squares Structural Equation Modelling in Data Articles," *Data in Brief*, 48, 1-21.

Roemer, Ellen, Florian Schuberth, and Jörg Henseler (2021), "HTMT2–An Improved Criterion for Assessing Discriminant Validity in Structural Equation Modeling," *Industrial Management & Data Systems*, 121 (12), 2637-50.

Rönkkö, Mikko and Eunseong Cho (2022), "An Updated Guideline for Assessing Discriminant Validity," *Organizational Research Methods*, 25 (1), 6-47.

Rosseel, Yves (2012), "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, 48 (2), 1-36.

Sechrest, Lee (1963), "Incremental Validity: A Recommendation," *Educational and Psychological Measurement*, 23 (1), 153-58.

Stoel, Reinoud D., Francisca Galindo Garre, Conor Dolan, and Godfried Van Den Wittenboer (2006), "On the Likelihood Ratio Test in Structural Equation Modeling When Parameters Are Subject to Boundary Constraints," *Psychological Methods*, 11 (4), 439.

Tesser, Abraham and Herbert Krauss (1976), "On Validating a Relationship between Constructs," *Educational and Psychological Measurement*, 36 (1), 111-21.

Vale, C. David and Vincent A. Maurelli (1983), "Simulating Multivariate Nonnormal Distributions," *Psychometrika*, 48 (3), 465-71.

Voorhees, Clay M., Michael K. Brady, Roger Calantone, and Edward Ramirez (2016), "Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies," *Journal of the Academy of Marketing Science*, 44 (1), 119-34.

Wilson, Anne V. and Silvia Bellezza (2022), "Consumer Minimalism," *Journal of Consumer Research*, 48 (5), 796-816.

Yuan, Ke-Hai, Ying Cheng, and Wei Zhang (2010), "Determinants of Standard Errors of MLEs in Confirmatory Factor Analysis," *Psychometrika*, 75 (4), 633-48.