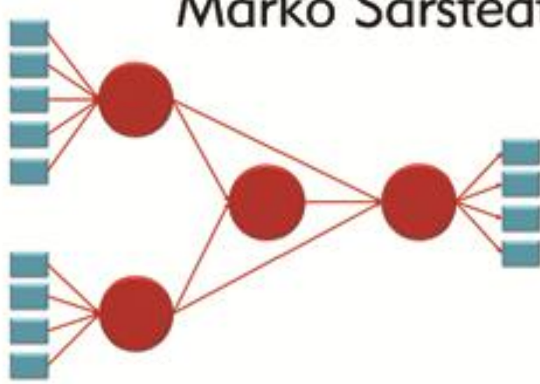
A PRIMER ON PARTIAL LEAST SQUARES STRUCTURAL EQUATION MODELING (PLS-SEM)

2e

Joseph F. Hair, Jr.
G. Tomas M. Hult
Christian M. Ringle
Marko Sarstedt



Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc. This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

CHAPTER 4

Assessing PLS-SEM Results Part I

Evaluation of Reflective Measurement Models

Do not copy, post, or distribute

LEARNING OUTCOMES

1. Gain an overview of Stage 5 of the process for using PLS-SEM, which deals with the evaluation of measurement models.
2. Describe Stage 5a: evaluating reflectively measured constructs.
3. Use the SmartPLS 3 software to assess reflectively measured constructs in the corporate reputation example.

CHAPTER PREVIEW

Having learned how to create and estimate a PLS path model, we now focus on understanding how to assess the quality of the results. Initially, we summarize the primary criteria that are used for PLS path model evaluation and their systematic application. Then, we focus on the evaluation of reflective measurement models. The PLS path model of corporate reputation is a practical application enabling you to review the relevant measurement model **evaluation criteria** and the appropriate reporting of results. This provides a foundation for the overview of formative measurement models in Chapter 5 and how to evaluate structural model results, which is covered in Chapter 6.

OVERVIEW OF STAGE 5: EVALUATION OF MEASUREMENT MODELS

Model estimation delivers empirical measures of the relationships between the indicators and the constructs (measurement models), as well as between the constructs (structural model). The empirical measures enable us to compare the theoretically established measurement and structural models with reality, as represented by the sample data. In other words, we can determine how well the theory fits the data.

PLS-SEM results are reviewed and evaluated using a systematic process. The goal of PLS-SEM is maximizing the **explained variance** (i.e., the **R^2 value**) of the endogenous latent variables in the PLS path model. For this reason, the evaluation of the quality of the PLS-SEM measurement and structural models focuses on metrics indicating the model's predictive capabilities. As with CB-SEM, the most important measurement model metrics for PLS-SEM are reliability, convergent validity, and discriminant validity. For the structural model, the most important evaluation metrics are R^2 (explained variance), f^2 (**effect size**), Q^2 (predictive relevance), and the size and statistical significance of the structural path coefficients. CB-SEM also relies on several of these metrics but in addition provides goodness-of-fit measures based on the discrepancy between the empirical and the model-implied (theoretical) covariance matrix. Since PLS-SEM relies on variances instead of covariances to determine an optimum solution, covariance-based goodness-of-fit measures are not fully transferrable to the PLS-SEM context. Fit measures in PLS-SEM are generally variance based and focus on the discrepancy between the observed (in the case of manifest variables) or approximated (in the case of latent variables) values of the dependent variables and the values predicted by the model in question. Nevertheless, research has proposed several PLS-SEM–based model fit measures, which are, however, in their early stages of development (see Chapter 6 for more details).

The systematic evaluation of these criteria follows a two-step process, as shown in Exhibit 4.1. The process involves separate assessments of the measurement models (Stage 5 of the procedure for using PLS-SEM) and the structural model (Stage 6).

PLS-SEM model assessment initially focuses on the measurement models. Examination of PLS-SEM estimates enables the researcher to

Do not copy, post, or distribute

Exhibit 4.1 Systematic Evaluation of PLS-SEM Results

Stage 5: Evaluation of the Measurement Models	
Stage 5a: Reflective Measurement Models	**Stage 5b: Formative Measurement Models**
• Internal consistency (Cronbach's alpha, composite reliability) • Convergent validity (indicator reliability, average variance extracted) • Discriminant validity	• Convergent validity • Collinearity between indicators • Significance and relevance of outer weights
Stage 6: Evaluation of the Structural Model	
• Coefficients of determination (R^2) • Predictive relevance (Q^2) • Size and significance of path coefficients • f^2 effect sizes • q^2 effect sizes	

evaluate the **reliability** and **validity** of the construct measures. Specifically, multivariate measurement involves using several variables (i.e., multi-items) to measure a construct. An example is the customer loyalty (*CUSL*) construct described in the PLS-SEM corporate reputation model, which we discussed earlier.

The logic of using multiple items as opposed to single items for construct measurement is that the measure will be more accurate. The anticipated improved accuracy is based on the assumption that using several indicators to measure a single concept is more likely to represent all the different aspects of the concept. However, even when using multiple items, the measurement is very likely to contain some degree of measurement error. There are many sources of measurement error in social sciences research, including poorly worded questions in a survey, misunderstanding of the scaling approach, and incorrect application of a statistical method, all of which lead to random and/or systematic errors. The objective is to reduce the measurement error as much as possible. Multivariate measurement enables researchers to more precisely identify measurement error and therefore account for it in research findings.

Do not copy, post, or distribute

Measurement error is the difference between the true value of a variable and the value obtained by a measurement. Specifically, the measured value x_m equals the true value x_t plus a measurement error. The measurement error ($e = \varepsilon_r + \varepsilon_s$) can have a random source (random error ε_r), which threatens reliability, or a systematic source (systematic error ε_s), which threatens validity. This relationship can be expressed as follows:

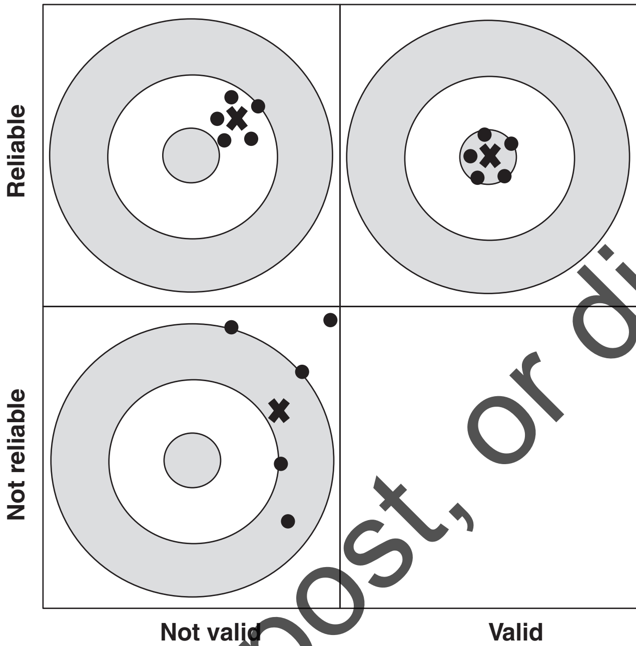
$$x_m = x_t + \varepsilon_r + \varepsilon_s.$$

In Exhibit 4.2, we explain the difference between reliability and validity by comparing a set of three targets. In this analogy, repeated measurements (e.g., of a customer's satisfaction with a specific service) are compared to arrows shot at a target. To measure each true score, we have five measurements (indicated by the black dots). The average value of the dots is indicated by a cross. Validity is indicated when the cross is close to the bull's-eye at the target center. The closer the average value (black cross in Exhibit 4.2) to the true score, the higher the validity. If several arrows are fired, reliability is the distances between the dots showing where the arrows hit the target. If all the dots are close together, the measure is reliable, even though the dots are not necessarily near the bull's-eye. This corresponds to the upper left box, where we have a scenario in which the measure is reliable but not valid. In the upper right box, both reliability and validity are shown. In the lower left box, though, we have a situation in which the measure is neither reliable nor valid. That is, the repeated measurements (dots) are scattered quite widely and the average value (cross) is not close to the bull's-eye. Even if the average value would match the true score (i.e., if the cross were in the bull's-eye), we would still not consider the measure valid. The reason is that an unreliable measure can never be valid, because there is no way we can distinguish the systematic error from the random error (Sarstedt & Mooi, 2014). If we repeat the measurement, say, five more times, the random error would likely shift the cross to a different position. Thus, reliability is a necessary condition for validity. This is also why the not reliable/valid scenario in the lower right box is not possible.

When evaluating the measurement models, we must distinguish between reflectively and formatively measured constructs (Chapter 2). The two approaches are based on different concepts and therefore require consideration of different evaluative measures. Reflective

Do not copy, post, or distribute

Exhibit 4.2 Comparing Reliability and Validity (Sarstedt & Mooi, 2014)



Source: Sarstedt, M., & Mooi, E. A. (2014). *A concise guide to market research* (2nd ed., p. 35). New York: Springer. With kind permission of Springer Science + Business Media.

Do not copy, post, or distribute

measurement models are assessed on their **internal consistency reliability** and validity. The specific measures include the composite reliability (as a means to assess the internal consistency reliability), convergent validity, and discriminant validity. The criteria for reflective measurement models cannot be universally applied to formative measurement models. With formative measures, the first step is to ensure **content validity** before collecting the data and estimating the PLS path model. After model estimation, different metrics are used to assess formative measures for **convergent validity,** the significance and relevance of indicator weights, and the presence of **collinearity** among indicators (Exhibit 4.1).

As implied by its name, a single-item construct (Chapter 2) is not represented by a multi-item measurement model. The relationship (i.e., the correlation) between the single indicator and the latent variable is always 1. Put differently, the single indicator and the latent variable have identical values. Thus, the criteria for the assessment of measurement models are not applicable to single-item constructs. To evaluate the reliability and validity of single-item measures, researchers must rely on proxies or different forms of validity assessment. For example, researchers can assess a single-item variable by means of criterion validity. This is done by correlating the single-item measure with an established criterion variable and comparing the resulting correlation with the correlation that results if the predictor construct is measured by a multi-item scale (e.g., Diamantopoulos et al., 2012). In terms of reliability, researchers often assume that one cannot estimate the reliability of single-item measures based on such techniques as common factor analysis and the correction for attenuation formula (e.g., Sarstedt & Wilczynski, 2009). These procedures require that both the multi-item measure and the single-item measure are included in the same survey. Thus, these analyses are of primary interest when researchers want to assess in a pretest or pilot study whether in the main study, a multi-item scale can be replaced with a single-item measure of the same construct. Still, recent research suggests that the reliability and validity of single items are highly context specific, which renders their assessment in pretests or pilot studies problematic (Sarstedt et al., in press).

The structural model estimates are not examined until the reliability and validity of the constructs have been established. If assessment of reflective (i.e., Stage 5a) and formative (i.e., Stage 5b) measurement models provides evidence of the measures' quality, the structural model estimates are evaluated in Stage 6 (Chapter 6). PLS-SEM assessment of the structural model involves the model's ability to predict the variance in the dependent variables. Hence, after reliability and validity are established, the primary evaluation criteria for PLS-SEM results are the **coefficients of determination** (**R^2** values) as well as the size and significance of the path coefficients. The **f^2 effect sizes**, **predictive relevance** (**Q^2**), and the **q^2 effect sizes** give additional insights about quality of the PLS path model estimations (Exhibit 4.1).

Assessment of PLS-SEM outcomes can be extended to more advanced analyses such as examining mediating or moderating effects, which we discuss in Chapter 7. Similarly, advanced analyses may involve estimating nonlinear effects (e.g., Rigdon, Ringle, & Sarstedt, 2010), conducting an importance-performance matrix analysis (PLS-IPMA; e.g., Rigdon, Ringle, Sarstedt, & Gudergan, 2011;

Do not copy, post, or distribute

Schloderer et al., 2014), assessing the mode of measurement model by using the confirmatory tetrad analysis (CTA-PLS; Gudergan et al., 2008), analyzing hierarchical component models (e.g., Becker, Klein, & Wetzels, 2012; Ringle et al., 2012), considering heterogeneity (e.g., Becker, Rai, Ringle, & Völckner, 2013; Sarstedt & Ringle, 2010), executing multigroup analyses (Sarstedt, Henseler, & Ringle, 2011), and assessing measurement model invariance (Henseler, Ringle, & Sarstedt, in press). In Chapter 8, we discuss several of these aspects in greater detail. The objective of these additional analyses is to extend and further differentiate the findings from the basic PLS path model estimation. Some of these advanced analyses are necessary to obtain a complete understanding of PLS-SEM results (e.g., checking for the presence of unobserved heterogeneity and significantly different subgroups), while others are optional.

The primary rules of thumb on how to evaluate PLS-SEM results are shown in Exhibit 4.3. In the following sections, we provide an overview of the process for assessing reflective measurement models (Stage 5a). Chapter 5 addresses the evaluation of formative measurement models (Stage 5b), while Chapter 6 deals with structural model evaluation.

Exhibit 4.3 Rules of Thumb for Evaluating PLS-SEM Results

- Model assessment in PLS-SEM primarily builds on nonparametric evaluation criteria based on bootstrapping and blindfolding. Goodness-of-fit measures used in CB-SEM are not universally transferrable to PLS-SEM, but recent research has brought forward various model fit criteria.
- Begin the evaluation process by assessing the quality of the reflective and formative measurement models (specific rules of thumb for reflective measurement models follow later in this chapter and in Chapter 5 for formative measurement models).
- If the measurement characteristics of constructs are acceptable, continue with the assessment of the structural model results. Path estimates should be statistically significant and meaningful. Moreover, endogenous constructs in the structural model should have high levels of explained variance as expressed in high R^2 values (Chapter 6 presents specific guidelines).
- Advanced analyses that extend and differentiate initial PLS-SEM findings may be necessary to obtain a correct picture of the results (Chapters 7 and 8).

Do not copy, post, or distribute

STAGE 5A: ASSESSING RESULTS OF REFLECTIVE MEASUREMENT MODELS

Assessment of reflective measurement models includes composite reliability to evaluate internal consistency, individual **indicator reliability,** and **average variance extracted** (**AVE**) to evaluate convergent validity. Assessment of reflective measurement models also includes discriminant validity. The **Fornell-Larcker criterion, cross-loadings,** and especially the heterotrait-monotrait (HTMT) ratio of correlations can be used to examine **discriminant validity**. In the following sections, we address each criterion for the evaluation of reflective measurement models.

Internal Consistency Reliability

The first criterion to be evaluated is typically **internal consistency reliability.** The traditional criterion for internal consistency is **Cronbach's alpha,** which provides an estimate of the reliability based on the intercorrelations of the observed indicator variables. This statistic is defined as follows:

$$Cronbach's\,\alpha = \left(\frac{M}{M-1}\right)\cdot\left(1-\frac{\sum_{i=1}^{M} s_i^2}{s_t^2}\right).$$

In this formula, s_i^2 represents the variance of the indicator variable i of a specific construct, measured with M indicators (i = 1, . . . , M), and s_t^2 is the variance of the sum of all M indicators of that construct. Cronbach's alpha assumes that all indicators are equally reliable (i.e., all the indicators have equal outer loadings on the construct). But PLS-SEM prioritizes the indicators according to their individual reliability. Moreover, Cronbach's alpha is sensitive to the number of items in the scale and generally tends to underestimate the internal consistency reliability. As such, it may be used as a more conservative measure of internal consistency reliability. Due to Cronbach's alpha's limitations, it is technically more appropriate to apply a different measure of internal consistency reliability, which is referred to as **composite reliability.** This measure of reliability takes into account the different outer loadings of the indicator variables and is calculated using the following formula:

Do not copy, post, or distribute

$$\rho_c = \frac{\left(\sum_{i=1}^{M} l_i\right)^2}{\left(\sum_{i=1}^{M} l_i\right)^2 + \sum_{i=1}^{M} var(e_i)},$$

where l_i symbolizes the standardized outer loading of the indicator variable *i* of a specific construct measured with *M* indicators, e_i is the measurement error of indicator variable *i,* and $var(e_i)$ denotes the variance of the measurement error, which is defined as $1 - l_i^2$.

The composite reliability varies between 0 and 1, with higher values indicating higher levels of reliability. It is generally interpreted in the same way as Cronbach's alpha. Specifically, composite reliability values of 0.60 to 0.70 are acceptable in exploratory research, while in more advanced stages of research, values between 0.70 and 0.90 can be regarded as satisfactory. Values above 0.90 (and definitely above 0.95) are not desirable because they indicate that all the indicator variables are measuring the same phenomenon and are therefore not likely to be a valid measure of the construct. Specifically, such composite reliability values occur if one uses semantically redundant items by slightly rephrasing the very same question. As the use of redundant items has adverse consequences for the measures' content validity (e.g., Rossiter, 2002) and may boost error term correlations (Drolet & Morrison, 2001; Hayduk & Littvay, 2012), researchers are advised to minimize the number of redundant indicators. Finally, composite reliability values below 0.60 indicate a lack of internal consistency reliability.

Cronbach's alpha is a conservative measure of reliability (i.e., it results in relatively low reliability values). In contrast, composite reliability tends to overestimate the internal consistency reliability, thereby resulting in comparatively higher reliability estimates. Therefore, it is reasonable to consider and report both criteria. When analyzing and assessing the measures' internal consistency reliability, the true reliability usually lies between Cronbach's alpha (representing the lower bound) and the composite reliability (representing the upper bound).

Convergent Validity

Convergent validity is the extent to which a measure correlates positively with alternative measures of the same construct. Using the

Do not copy, post, or distribute

domain sampling model, indicators of a reflective construct are treated as different (alternative) approaches to measure the same construct. Therefore, the items that are indicators (measures) of a specific reflective construct should converge or share a high proportion of variance. To evaluate convergent validity of reflective constructs, researchers consider the outer loadings of the indicators and the average variance extracted (AVE).

High outer loadings on a construct indicate the associated indicators have much in common, which is captured by the construct. The size of the outer loading is also commonly called **indicator reliability.** At a minimum, the outer loadings of all indicators should be statistically significant. Because a significant outer loading could still be fairly weak, a common rule of thumb is that the standardized outer loadings should be 0.708 or higher. The rationale behind this rule can be understood in the context of the square of a standardized indicator's outer loading, referred to as the **communality** of an item. The square of a standardized indicator's outer loading represents how much of the variation in an item is explained by the construct and is described as the variance extracted from the item. An established rule of thumb is that a latent variable should explain a substantial part of each indicator's variance, usually at least 50%. This also implies that the variance shared between the construct and its indicator is larger than the measurement error variance. This means that an indicator's outer loading should be above 0.708 since that number squared (0.708^2) equals 0.50. Note that in most instances, 0.70 is considered close enough to 0.708 to be acceptable.

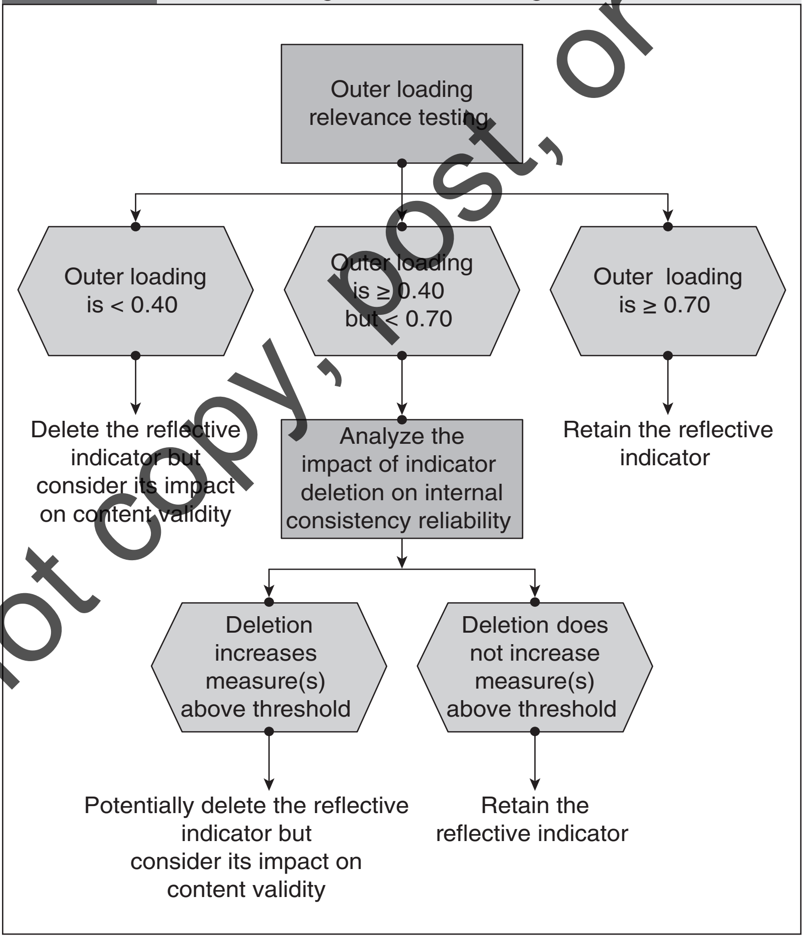
Researchers frequently obtain weaker outer loadings (<0.70) in social science studies, especially when newly developed scales are used (Hulland, 1999). Rather than automatically eliminating indicators when their outer loading is below 0.70, researchers should carefully examine the effects of item removal on the composite reliability, as well as on the content validity of the construct. Generally, indicators with outer loadings between 0.40 and 0.70 should be considered for removal from the scale only when deleting the indicator leads to an increase in the composite reliability (or the average variance extracted; see next section) above the suggested threshold value. Another consideration in the decision of whether to delete an indicator is the extent to which its removal affects content validity. Indicators with weaker outer loadings are sometimes retained on the basis of their

Do not copy, post, or distribute

contribution to content validity. Indicators with very low outer loadings (below 0.40) should, however, always be eliminated from the construct (Bagozzi, Yi, & Philipps, 1991; Hair et al., 2011). Exhibit 4.4 illustrates the recommendations regarding indicator deletion based on outer loadings.

A common measure to establish convergent validity on the construct level is the **average variance extracted** (**AVE**). This criterion is defined as the grand mean value of the squared loadings of the indicators associated with the construct (i.e., the sum of the squared loadings divided by the number of indicators). Therefore, the AVE is

Exhibit 4.4 Outer Loading Relevance Testing



Do not copy, post, or distribute

equivalent to the **communality** of a construct. The AVE is calculated using the following formula:

$$\text{AVE} = \left(\frac{\sum_{i=1}^{M} l_i^2}{M} \right).$$

Using the same logic as that used with the individual indicators, an AVE value of 0.50 or higher indicates that, on average, the construct explains more than half of the variance of its indicators. Conversely, an AVE of less than 0.50 indicates that, on average, more variance remains in the error of the items than in the variance explained by the construct.

The AVE of each reflectively measured construct should be evaluated. In the example introduced in Chapter 2, an AVE estimate is needed only for constructs *COMP, CUSL,* and *LIKE.* For the single-item construct *CUSA,* the AVE is not an appropriate measure since the indicator's outer loading is fixed at 1.00.

Discriminant Validity

Discriminant validity is the extent to which a construct is truly distinct from other constructs by empirical standards. Thus, establishing discriminant validity implies that a construct is unique and captures phenomena not represented by other constructs in the model. Traditionally, researchers have relied on two measures of discriminant validity. The **cross-loadings** are typically the first approach to assess the discriminant validity of the indicators. Specifically, an indicator's outer loading on the associated construct should be greater than any of its cross-loadings (i.e., its correlation) on other constructs. The best way to assess and report cross-loadings is in a table with rows for the indicators and columns for the latent variable. Exhibit 4.5 illustrates this analysis in an example with three latent variables (Y_1, Y_2, and Y_3), each measured with two indicators. As can be seen, the loadings always exceed the cross-loadings. For example, x_{11} loads high on its corresponding construct Y_1 (0.75) but much lower on constructs Y_2 (0.49) and Y_3 (0.41). In this example, the analysis of cross-loadings suggests that discriminant validity has been established. On the contrary, the presence of cross-loadings that exceed the indicators' outer loadings would represent a discriminant validity problem.

The **Fornell-Larcker criterion** is the second approach to assessing discriminant validity. It compares the square root of the AVE values

Do not copy, post, or distribute

Exhibit 4.5 Cross-Loadings Analysis

	Y_1	Y_2	Y_3
x_{11}	**0.75**	0.49	0.41
x_{12}	**0.83**	0.27	0.35
x_{21}	0.55	**0.82**	0.60
x_{22}	0.45	**0.82**	0.42
x_{31}	0.43	0.53	**0.87**
x_{32}	0.42	0.55	**0.84**

Note: One expects that an indicator has the highest loading value (in bold) with the construct to which it has been assigned to.

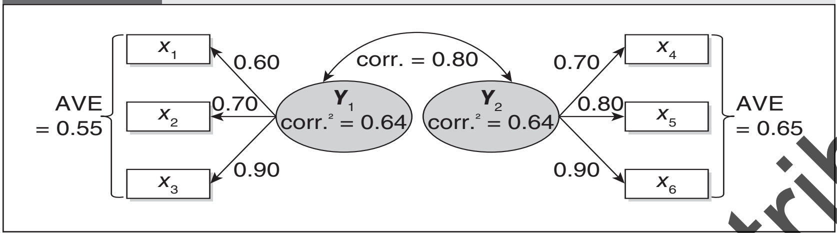
with the latent variable correlations. Specifically, the square root of each construct's AVE should be greater than its highest correlation with any other construct. An alternative approach to evaluating the results of the Fornell-Larcker criterion is to determine whether the AVE is larger than the squared correlation with any other construct. The logic of the Fornell-Larcker method is based on the idea that a construct shares more variance with its associated indicators than with any other construct.

Exhibit 4.6 illustrates this concept. In the example, the AVE values of the constructs Y_1 and Y_2 are 0.55 and 0.65, respectively. The AVE values are obtained by squaring each outer loading, obtaining the sum of the three squared outer loadings, and then calculating the average value. For example, with respect to construct Y_1, 0.60, 0.70, and 0.90 squared are 0.36, 0.49, and 0.81, respectively. The sum of these three numbers is 1.66, and the average value is therefore 0.55 (i.e., 1.66/3). The correlation between constructs Y_1 and Y_2 (as indicated by the double-headed arrow linking the two constructs) is 0.80. Squaring the correlation of 0.80 indicates that 64% (i.e., the squared correlation; $0.80^2 = 0.64$) of each construct's variation is explained by the other construct. Therefore, Y_1 explains less variance in its indicator measures x_1 to x_3 than it shares with Y_2, which implies that the two constructs (Y_1 and Y_2), which are conceptually different, are not sufficiently different in terms of their empirical standards. Thus, in this example, discriminant validity is not established.

The analysis and presentation of the results of the Fornell-Larcker criterion are illustrated in Exhibit 4.7—for a PLS path model with

Do not copy, post, or distribute

Exhibit 4.6 Visual Representation of the Fornell-Larcker Criterion



two reflective constructs (i.e., Y_1 and Y_2), one formative construct (i.e., Y_3), and a single-item construct (i.e., Y_4). The first consideration is that only reflective multi-item constructs are evaluated using the Fornell-Larcker criterion. Therefore, constructs Y_3 and Y_4 are exceptions to this type of evaluation since the AVE value is not a meaningful criterion for formative and single-item measures. Looking only at constructs Y_1 and Y_2, note that the square root of each construct's AVE is on the diagonal. The nondiagonal elements represent the correlations between the latent variables. To establish discriminant validity, the square root of each construct's AVE must be larger than its correlation with other constructs. To evaluate the reflective construct Y_2 in Exhibit 4.7, one would compare all correlations in the row of Y_2 and the column of Y_2 with its square root of the AVE. In the case study illustration of the corporate reputation path model later in this chapter, the actual estimated values for this type of analysis are provided.

Recent research that critically examined the performance of cross-loadings and the Fornell-Larcker criterion for discriminant validity assessment has found that neither approach reliably detects

Exhibit 4.7 Example of Fornell-Larcker Criterion Analysis

	Y_1	Y_2	Y_3	Y_4
Y_1	$\sqrt{AVE_{Y_1}}$			
Y_2	$CORR_{Y_1Y_2}$	$\sqrt{AVE_{Y_2}}$		
Y_3	$CORR_{Y_1Y_3}$	$CORR_{Y_2Y_3}$	*Formative measurement model*	
Y_4	$CORR_{Y_1Y_4}$	$CORR_{Y_2Y_4}$	$CORR_{Y_3Y_4}$	*Single-item construct*

Do not copy, post, or distribute

discriminant validity issues (Henseler et al., 2015). Specifically, cross-loadings fail to indicate a lack of discriminant validity when two constructs are perfectly correlated, which renders this criterion ineffective for empirical research. Similarly, the Fornell-Larcker criterion performs very poorly, especially when indicator loadings of the constructs under consideration differ only slightly (e.g., all indicator loadings vary between 0.60 and 0.80). When indicator loadings vary more strongly, the Fornell-Larcker criterion's performance in detecting discriminant validity issues improves but is still rather poor overall. (also see Voorhees, Brady, Calantone, & Ramirez, 2016).

As a remedy, Henseler et al. (2015) propose assessing the **heterotrait-monotrait ratio (HTMT)** of the correlations. In short, **HTMT** is the ratio of the between-trait correlations to the within-trait correlations. HTMT is the mean of all correlations of indicators across constructs measuring different constructs (i.e., the **heterotrait-heteromethod correlations**) relative to the (geometric) mean of the average correlations of indicators measuring the same construct (i.e., the monotrait-heteromethod correlations; for a formal definition of the HTMT statistic, see Henseler et al., 2015). Technically, the HTMT approach is an estimate of what the true correlation between two constructs would be, if they were perfectly measured (i.e., if they were perfectly reliable). This true correlation is also referred to as **disattenuated correlation.** A disattenuated correlation between two constructs close to 1 indicates a lack of discriminant validity.

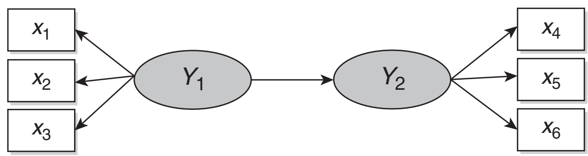
Exhibit 4.8 illustrates the HTMT approach. The average **heterotrait-heteromethod correlations** equal all pairwise correlations between variables x_1, x_2, and x_3 and x_4, x_5, and x_6 (gray-shaded area in the correlation matrix in Exhibit 4.8). In the example, the average heterotrait-heteromethod correlation is 0.341. The average **monotrait-heteromethod correlations** of Y_1 equal the mean of all pairwise correlations between x_1, x_2, and x_3 (i.e., 0.712). Similarly, the mean of all pairwise correlations between x_4, x_5, and x_6 (i.e., 0.409) defines the average monotrait-heteromethod correlations of Y_2. The HTMT statistic for the relationship between Y_1 and Y_2 therefore equals

$$HTMT(Y_1, Y_2) = \frac{0.341}{\sqrt{0.712 \cdot 0.409}} = 0.632.$$

The exact threshold level of the HTMT is debatable; after all, "when is a correlation close to 1?" Based on prior research and their study results, Henseler et al. (2015) suggest a threshold value of 0.90

Do not copy, post, or distribute

Exhibit 4.8 Visual Representation of the HTMT Approach



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1					
x_2	0.770	1				
x_3	0.701	0.665	1			
x_4	0.426	0.339	0.393	1		
x_5	0.423	0.345	0.385	0.574	1	
x_6	0.274	0.235	0.250	0.318	0.335	1

if the path model includes constructs that are conceptually very similar (e.g., affective satisfaction, cognitive satisfaction, and loyalty). In other words, an HTMT value above 0.90 suggests a lack of discriminant validity. When the constructs in the path model are conceptually more distinct, a lower and thus more conservative threshold value of 0.85 seems warranted (Henseler et al., 2015). Furthermore, the HTMT can serve as the basis of a statistical discriminant validity test. However, as PLS-SEM does not rely on any distributional assumptions, standard parametric significance tests cannot be applied to test whether the HTMT statistic is significantly different from 1. Instead, researchers have to rely on a procedure called **bootstrapping** to derive a distribution of the HTMT statistic (see Chapter 5 for more details on the bootstrapping procedure).

In bootstrapping, subsamples are randomly drawn (with replacement) from the original set of data. Each subsample is then used to estimate the model. This process is repeated until a large number of random subsamples have been created, typically about 5,000. The estimated parameters from the subsamples (in this case, the HTMT statistic) are used to derive standard errors for the estimates. With this information, it is possible to derive a **bootstrap confidence interval.** The confidence interval is the range into which the true HTMT population value will fall, assuming a certain level of confidence (e.g., 95%).

Do not copy, post, or distribute

A confidence interval containing the value 1 indicates a lack of discriminant validity. Conversely, if the value 1 falls outside the interval's range, this suggests that the two constructs are empirically distinct. Since the HTMT-based assessment using a confidence interval relies on inferential statistics, one should primarily rely on this criterion, especially in light of the limitations of cross-loadings and the Fornell-Larcker criterion. However, the latter two measures still constitute standard means for discriminant validity assessment.

What should researchers do if any of the criteria signal a lack of discriminant validity? There are different ways to handle discriminant validity problems (Exhibit 4.9). The first approach retains the constructs that cause discriminant validity problems in the model and aims at increasing the average monotrait-heteromethod correlations and/or decreasing the average heteromethod-heterotrait correlations of the constructs measures.

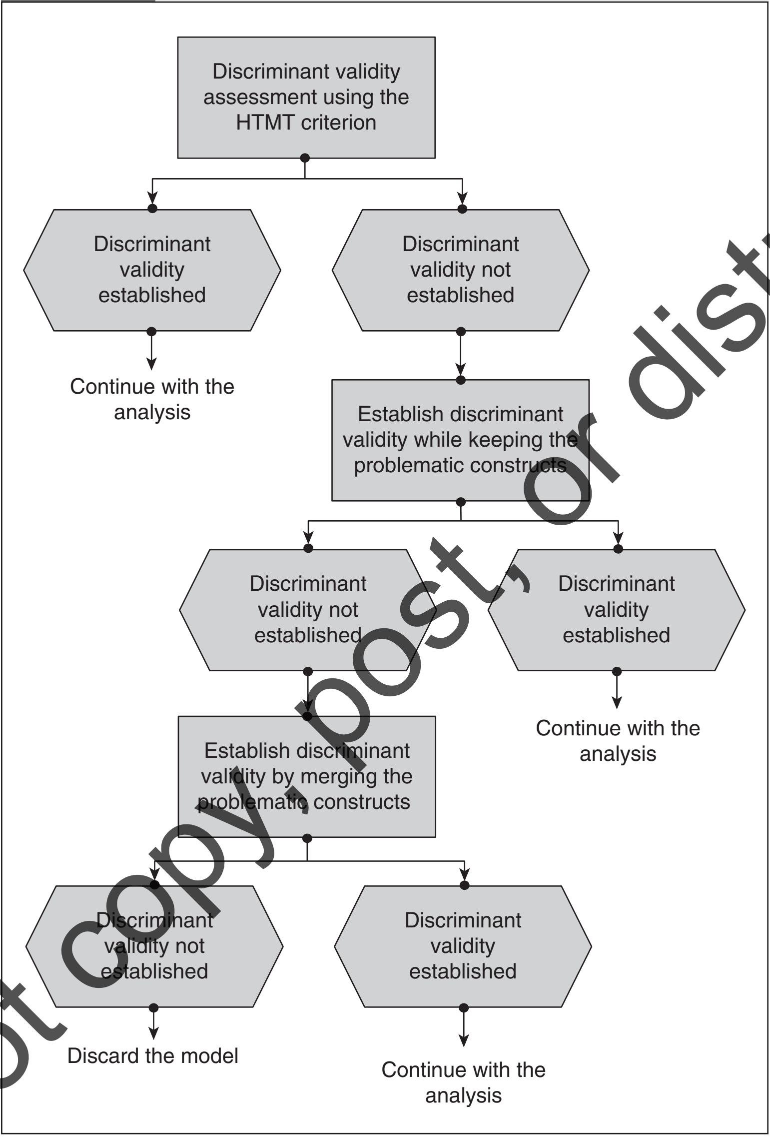
To decrease the HTMT by increasing a construct's average monotrait-heteromethod correlations, one can eliminate items that have low correlations with other items measuring the same construct. Likewise, heterogeneous subdimensions in the construct's set of items could also deflate the average monotrait-heteromethod correlations. In this case, the construct (e.g., quality) can be split into homogeneous subconstructs (e.g., product quality and service quality), perhaps using a higher-order construct, if the measurement theory supports this step (e.g., Kocyigit & Ringle, 2011). These subconstructs then replace the more general construct in the model. When following this approach, however, the discriminant validity of the newly generated constructs with all the other constructs in the model needs to be reevaluated.

To decrease the average heteromethod-heterotrait correlations, one can (1) eliminate items that are strongly correlated with items in the opposing construct, or (2) reassign these indicators to the other construct, if theoretically plausible. It is important to note that the elimination of items purely on statistical grounds can have adverse consequences for the content validity of the constructs. Therefore, this step entails carefully examining the scales (based on prior research results or on a pretest when newly developed measures are involved) to determine whether all the construct domain facets have been captured. At least two expert coders should conduct this judgment independently to ensure a high degree of objectivity.

Another approach to treating discriminant validity problems involves merging the constructs that cause the problems into a more general construct. Again, measurement theory must support this step.

Do not copy, post, or distribute

Exhibit 4.9 Handling Discriminant Validity Problems



Do not copy, post, or distribute

In this case, the more general construct replaces the problematic constructs in the model. This step may entail modifications to increase a construct's average monotrait-heteromethod correlations and/or to decrease the average heteromethod-heterotrait correlations.

In Exhibit 4.10, we summarize the criteria used to assess the reliability and validity of reflective construct measures. If the criteria are

Exhibit 4.10 Rules of Thumb for Evaluating Reflective Measurement Models

- Internal consistency reliability: composite reliability should be higher than 0.70 (in exploratory research, 0.60 to 0.70 is considered acceptable). Consider Cronbach's alpha as the lower bound and composite reliability as the upper bound of internal consistency reliability.
- Indicator reliability: the indicator's outer loadings should be higher than 0.70. Indicators with outer loadings between 0.40 and 0.70 should be considered for removal only if the deletion leads to an increase in composite reliability and AVE above the suggested threshold value.
- Convergent validity: the AVE should be higher than 0.50.
- Discriminant validity:
  - Use the HTMT criterion to assess discriminant validity in PLS-SEM.
  - The confidence interval of the HTMT statistic should not include the value 1 for all combinations of constructs.
  - According to the traditional discriminant validity assessment methods, an indicator's outer loadings on a construct should be higher than all its cross-loadings with other constructs. Furthermore, the square root of the AVE of each construct should be higher than its highest correlation with any other construct (Fornell-Larcker criterion).

not met, the researcher may decide to remove single indicators from a specific construct in an attempt to more closely meet the criteria. However, removing indicators should be carried out with care since the elimination of one or more indicators may improve the reliability or discriminant validity but at the same time may decrease the measurement's content validity.

Do not copy, post, or distribute

CASE STUDY ILLUSTRATION—REFLECTIVE MEASUREMENT MODELS

Running the PLS-SEM Algorithm

We continue working with our PLS-SEM example on corporate reputation. In Chapter 3, we explained how to estimate the PLS path model and how to obtain the results by opening the default report in the SmartPLS 3 software. Recall that to do so, you must first load the simple corporate reputation model and then run the model by clicking on the icon at the top right or by using the pull-down menu by going to **Calculate → PLS Algorithm.** After running the **PLS Algorithm,** the SmartPLS results report automatically opens; if not, go to the **Calculation Results** tab on the bottom left of the screen and click on **Report**.

Before analyzing the results, you need to quickly check if the algorithm converged (i.e., the stop criterion of the algorithm was reached and not the maximum number of iterations). To do so, go to **Interim Results → Stop Criterion Changes** in the results report. You will then see the table shown in Exhibit 4.11, which shows the number of iterations of the PLS-SEM algorithm. This number should be lower than the maximum number of iterations (e.g., 300) that you defined in the PLS-SEM algorithm parameter settings (Chapter 2). At the bottom left side of the table, you will see that the algorithm converged after Iteration 5.

If the PLS-SEM algorithm does not converge in fewer than 300 iterations (the default setting in the software), the algorithm could not find a stable solution. This kind of situation almost never occurs. But if it does occur, there are two possible causes of the problem: (1) the selected stop criterion is at a very small level (e.g., 1.0E-10) so that little changes in the coefficients of the measurement models prevent the PLS-SEM algorithm from stopping, or (2) there are problems with the data and they need to be checked carefully. For example, data problems may occur if the sample size is too small or if an indicator has many identical values (i.e., the same data points, which results in insufficient variability).

When your PLS path model estimation converges, which it practically always does, you need to examine the following PLS-SEM calculation results tables from the results report for reflective measurement model assessment: **Outer Loadings, Composite Reliability, Cronbach's Alpha, Average Variance Extracted (AVE),** and **Discriminant Validity.** We examine other information in the report in Chapters 5 and 6, when we extend the simple path model by including formative measures and examine the structural model results.

Exhibit 4.11 Stop Criterion Table in SmartPLS

Stop Criterion Changes

Matrix

	comp_1	comp_2	comp_3	cusa	cusl_1	cusl_2	cusl_3	like_1	like_2	like_3
Iteration 0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Iteration 1	0.536	0.341	0.328	1.000	0.368	0.421	0.365	0.419	0.378	0.359
Iteration 2	0.536	0.340	0.328	1.000	0.369	0.420	0.365	0.418	0.378	0.360
Iteration 3	0.536	0.340	0.328	1.000	0.369	0.420	0.365	0.418	0.378	0.360
Iteration 4	0.536	0.340	0.328	1.000	0.369	0.420	0.365	0.418	0.378	0.360
Iteration 5	0.536	0.340	0.328	1.000	0.369	0.420	0.365	0.418	0.378	0.360

Do not copy, post, or distribute

Reflective Measurement Model Evaluation

The simple corporate reputation model has three latent variables with reflective measurement models (i.e., *COMP, CUSL,* and *LIKE*) as well as a single-item construct (*CUSA*). For the reflective measurement models, we need the estimates for the relationships between the reflective latent variables and their indicators (i.e., outer loadings). Exhibit 4.12 displays the results table for the outer loadings, which can be found under **Final Results → Outer Loadings.** By default, the outer loadings are also displayed in the modeling window after running the PLS-SEM algorithm. All outer loadings of the reflective constructs *COMP, CUSL,* and *LIKE* are well above the threshold value of 0.70, which suggests sufficient levels of indicator reliability. The indicator *comp_2* (outer loading: 0.798) has the smallest indicator reliability with a value of 0.637 (0.798^2), while the indicator *cusl_2* (outer loading: 0.917) has the highest indicator reliability, with a value of 0.841 (0.917^2).

To evaluate the composite reliability of the construct measures, left-click on the **Construct Reliability and Validity** tab under **Quality Criteria** in the results report. Here, you have the option of displaying the composite reliability values using a bar chart or in a matrix format. Exhibit 4.13 shows the bar chart of the constructs' composite reliability values. The horizontal blue line indicates the common minimum threshold level for composite reliability (i.e., 0.70). If a composite reliability value is above this threshold value, the corresponding bar is colored green. If the composite reliability value is lower than 0.70, the bar is colored red. In our example, all composite reliability values exceed the threshold. Clicking on the **Matrix** tab shows the specific composite reliability values. With values of 0.865 (*COMP*), 0.899 (*CUSL*), and 0.899 (*LIKE*), all three reflective constructs have high

Exhibit 4.12 Outer Loadings

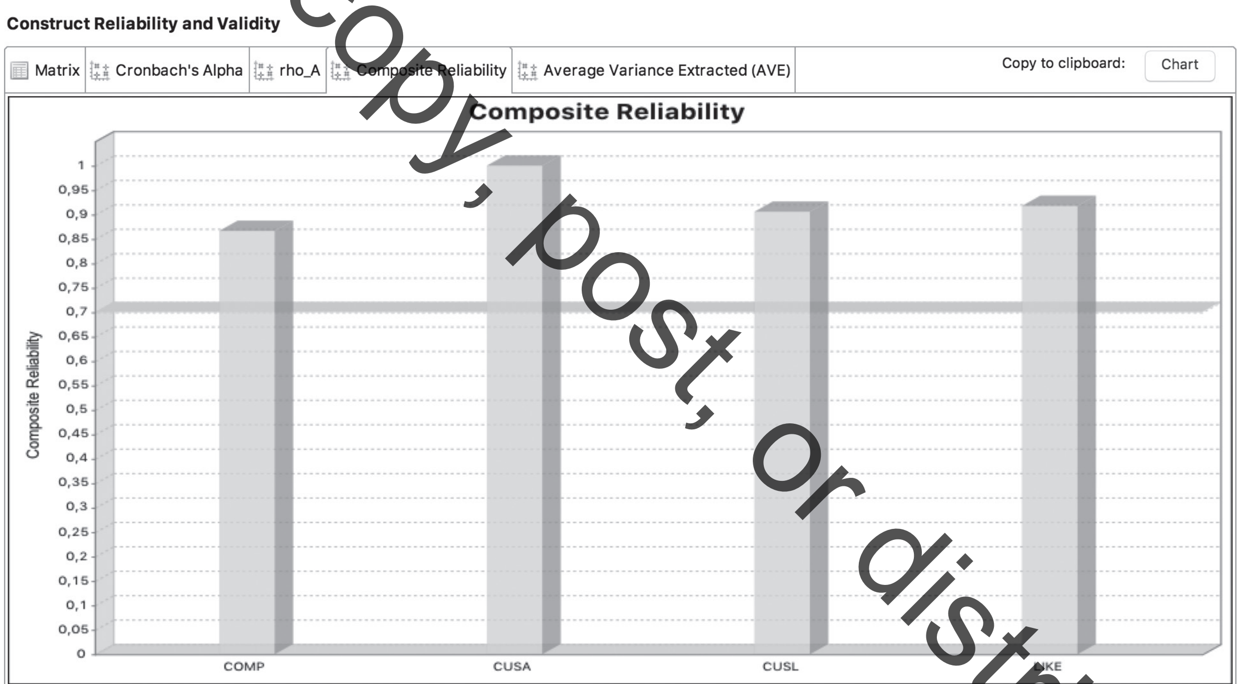
Outer Loadings

Matrix

	COMP	CUSA	CUSL	LIKE
comp_1	0.858			
comp_2	0.798			
comp_3	0.818			
cusa		1.000		
cusl_1			0.833	
cusl_2			0.917	
cusl_3			0.843	
like_1				0.879
like_2				0.870
like_3				0.843

Do not copy, post, or distribute

Exhibit 4.13 Composite Reliability



Do not copy, post, or distribute

levels of internal consistency reliability. Note that the composite reliability value of the single-item variable *CUSA* is 1.00. But this cannot be interpreted as evidence that the construct exhibits perfect reliability and should not be reported with other measures of reliability.

Going to **Quality Criteria → Construct Reliability and Validity** gives you the option to show the chart of Cronbach's alpha values for all constructs (Exhibit 4.14). All bars in the chart appear in green, indicating that all construct measures are above the 0.70 threshold. The specific Cronbach's alpha (0.776 for *COMP*, 0.831 for *CUSL*, and 0.831 for *LIKE*) values can be accessed by left-clicking on the **Matrix** tab. Again, as *CUSA* is measured using a single item, interpreting this construct's Cronbach's alpha value is not meaningful.

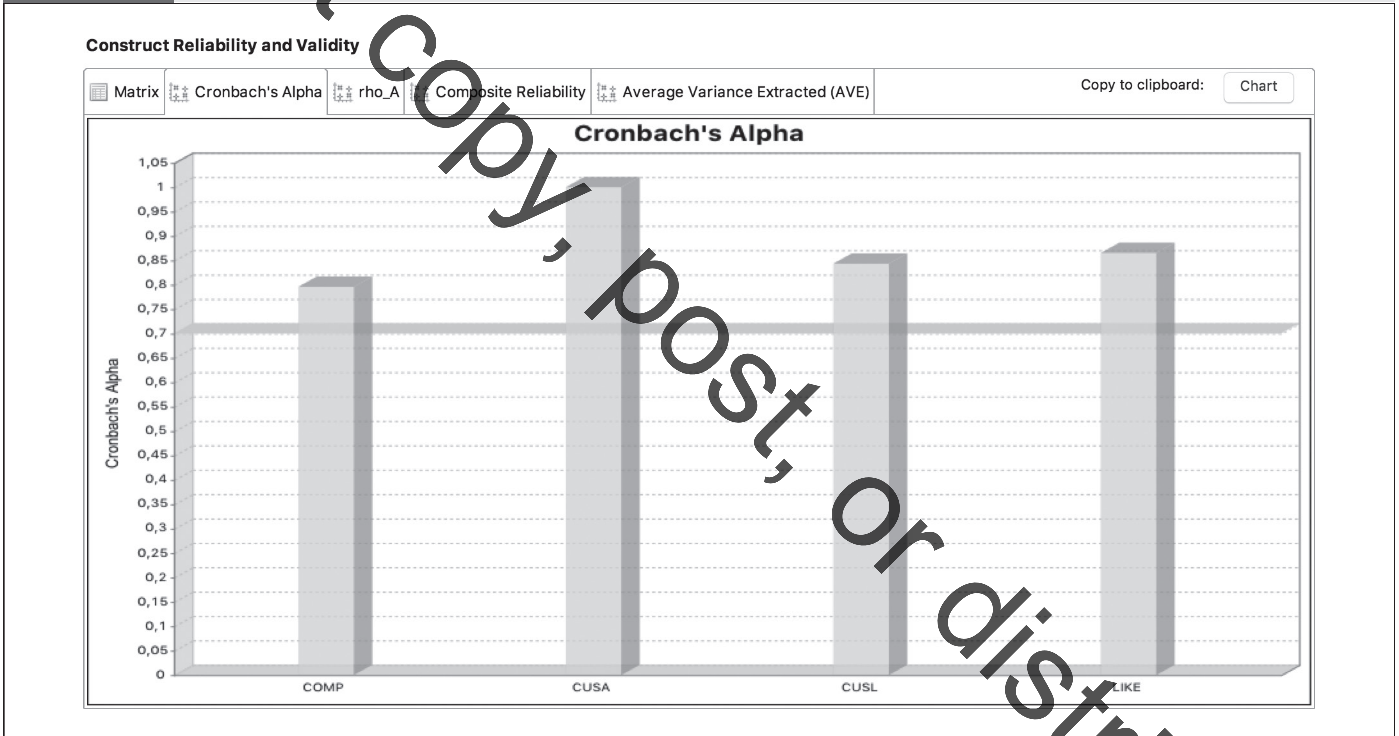
Convergent validity assessment is based on the AVE values, which can be accessed by going to **Quality Criteria → Construct Reliability and Validity** in the results report. As with composite reliability and Cronbach's alpha, SmartPLS offers the option of displaying the results using bar charts (Exhibit 4.15) or in a matrix format. In this example, the AVE values of *COMP* (0.681), *CUSL* (0.748), and *LIKE* (0.747) are well above the required minimum level of 0.50. Thus, the measures of the three reflective constructs have high levels of convergent validity.

Finally, in the **Discriminant Validity** tab under **Quality Criteria,** SmartPLS 3 offers several means to assess whether the construct measures discriminate well empirically. According to the Fornell-Larcker criterion, the square root of the AVE of each construct should be higher than the construct's highest correlation with any other construct in the model (this notion is identical to comparing the AVE with the squared correlations between the constructs). Exhibit 4.16 shows the results of the Fornell-Larcker criterion assessment with the square root of the reflective constructs' AVE on the diagonal and the correlations between the constructs in the off-diagonal position. For example, the reflective construct *COMP* has a value of 0.825 for the square root of its AVE, which needs to be compared with all correlation values in the column of *COMP*. Note that for *CUSL*, you need to consider the correlations in both the row and column. Overall, the square roots of the AVEs for the reflective constructs *COMP* (0.825), *CUSL* (0.865), and *LIKE* (0.864) are all higher than the correlations of these constructs with other latent variables in the path model, thus indicating all constructs are valid measures of unique concepts.

Another alternative to assessing discriminant validity is the cross-loadings. One can check the cross-loadings (click on **Cross Loadings** in the **Discriminant Validity** section of the results report) to make this evaluation. Discriminant validity is established when an indicator's

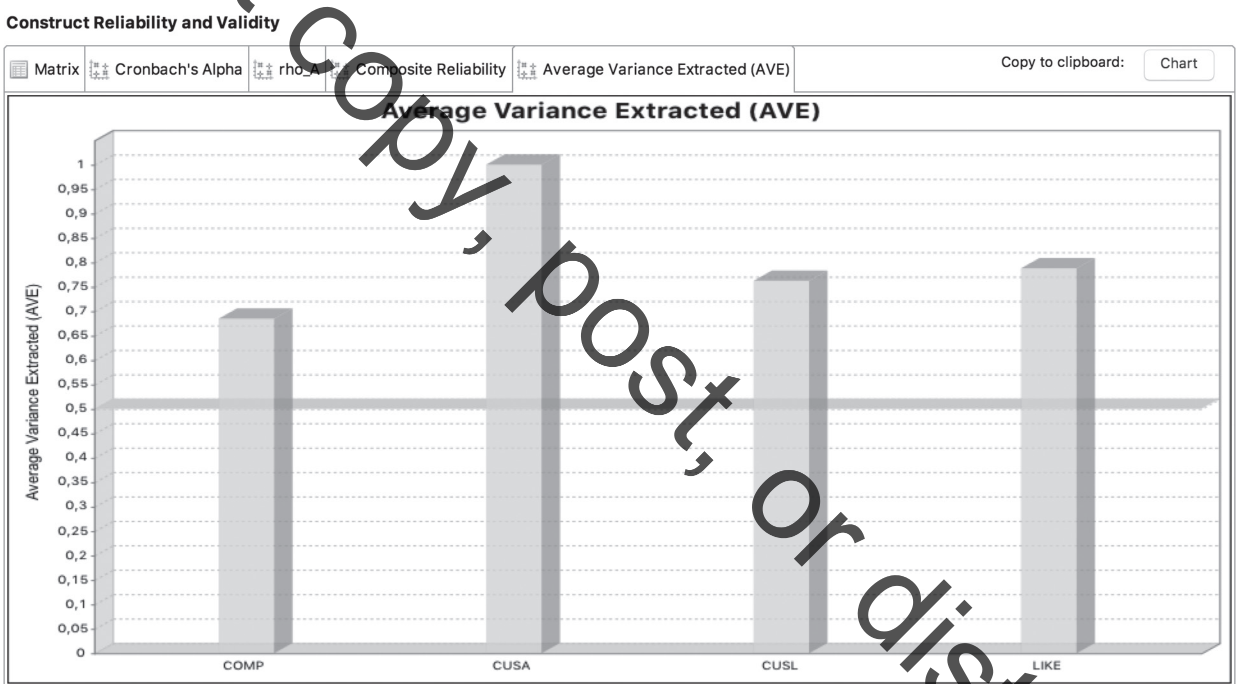
Do not copy, post, or distribute

Exhibit 4.14 Cronbach's Alpha



Do not copy, post, or distribute

Exhibit 4.15 Average Variance Extracted (AVE)



Do not copy, post, or distribute

Exhibit 4.16 Fornell-Larcker Criterion

Discriminant Validity

Fornell-Larcker Criterium | Cross Loadings | Heterotrait-Monotrait Ratio (HTMT)

	COMP	CUSA	CUSL	LIKE
COMP	0.825			
CUSA	0.436	1.000		
CUSL	0.450	0.689	0.865	
LIKE	0.645	0.528	0.615	0.864

loading on its assigned construct is higher than all of its cross-loadings with other constructs. Exhibit 4.17 shows the loadings and cross-loadings for every indicator. For example, the indicator *comp_1* has the highest value for the loading with its corresponding construct *COMP* (0.858), while all cross-loadings with other constructs are considerably lower (e.g., *comp_1* on *CUSA*: 0.464). The same finding holds for the other indicators of *COMP* as well as the indicators measuring *CUSL* and *LIKE*. Overall, cross-loadings as well as the Fornell-Larcker criterion provide evidence for the constructs' discriminant validity.

However, note that while frequently used in applied research, neither the Fornell-Larcker criterion nor the cross-loadings allow for reliably detecting discriminant validity issues. Therefore, an alternative, more reliable criterion, HTMT, should be applied. The **Discriminant Validity** section of the results report includes the **Heterotrait-Monotrait Ratio (HTMT)**. Exhibit 4.18 shows the HTMT values for all pairs of constructs in a matrix format. The next tab also shows these HTMT values in bar charts, using 0.85 as the relevant threshold level. As can be seen, all HTMT values are

Exhibit 4.17 Cross-Loadings

Discriminant Validity

Fornell-Larcker Criterium | Cross Loadings | Heterotrait-Monotrait Ratio (HTMT)

	COMP	CUSA	CUSL	LIKE
comp_1	0.858	0.464	0.465	0.607
comp_2	0.798	0.286	0.304	0.460
comp_3	0.818	0.272	0.296	0.497
cusa	0.436	1.000	0.689	0.528
cusl_1	0.430	0.536	0.833	0.557
cusl_2	0.396	0.655	0.917	0.573
cusl_3	0.341	0.593	0.843	0.461
like_1	0.602	0.510	0.561	0.879
like_2	0.523	0.434	0.530	0.870
like_3	0.544	0.420	0.499	0.843

Do not copy, post, or distribute

Exhibit 4.18 HTMT

Discriminant Validity

Fornell-Larcker Criterion | Cross Loadings | Heterotrait-Monotrait Ratio (HTMT) | Heterotrait-Monotrait Ratio (HTMT)

	COMP	CUSA	CUSL	LIKE
COMP				
CUSA	0.465			
CUSL	0.532	0.755		
LIKE	0.780	0.577	0.737	

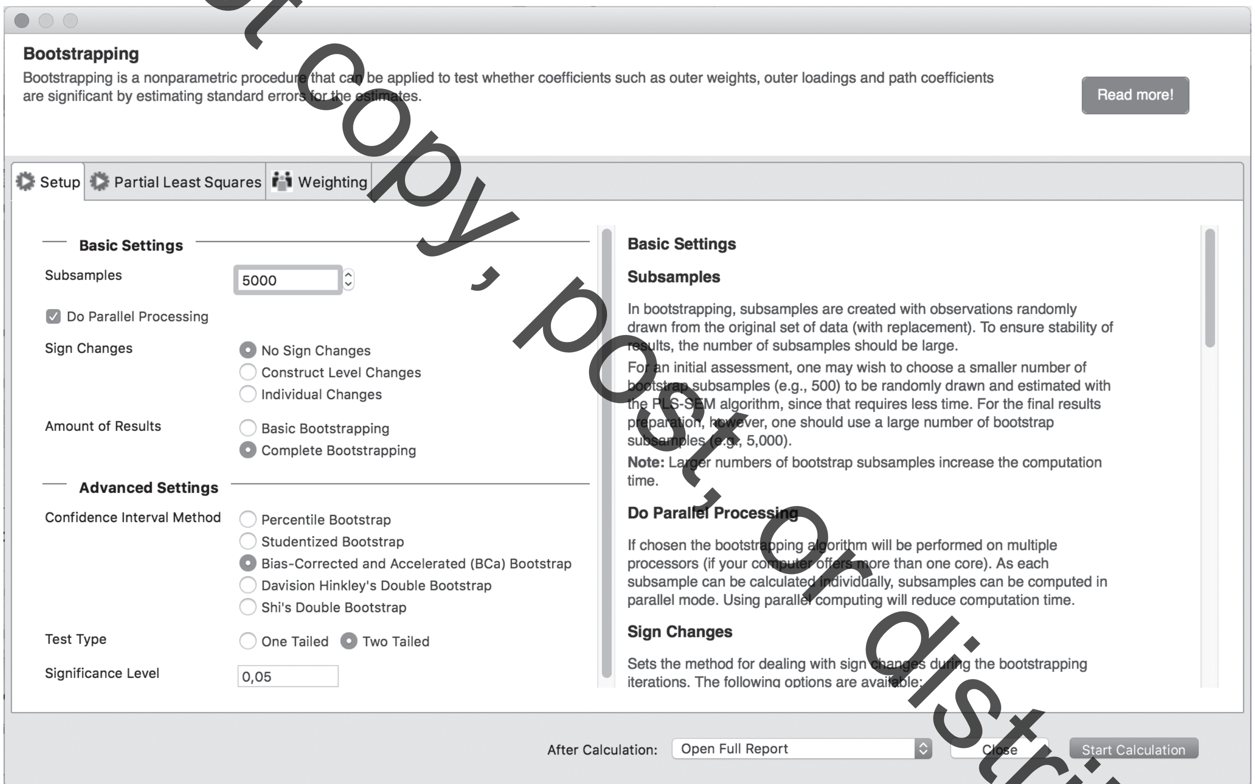
clearly lower than the more conservative threshold value of 0.85, even for *CUSA* and *CUSL*, which, from a conceptual viewpoint, are very similar. Recall that the threshold value for conceptually similar constructs is 0.90.

In addition to examining the HTMT ratios, you should test whether the HTMT values are significantly different from 1. This requires computing bootstrap confidence intervals obtained by running the bootstrapping option. To run the bootstrapping procedure, go back to the modeling window and left-click on **Calculate → Bootstrapping** in the pull-down menu. In the dialog box that opens, choose the bootstrapping options as displayed in Exhibit 4.19 (Chapter 5 includes a more detailed introduction to the bootstrapping procedure and the parameter settings). Make sure to select the **Complete Bootstrapping** option, which, unlike the **Basic Bootstrapping** option, includes the results for HTMT. Finally, click on **Start Calculation**.

After running bootstrapping, open the results report. Go to **Quality Criteria → Heterotrait-Monotrait (HTMT)** and left-click on the tab **Confidence Intervals Bias Corrected.** The menu that opens up (Exhibit 4.20) shows the original HTMT values (column **Original Sample (O)**) for each combination of constructs in the model, along with the average HTMT values computed from the 5,000 bootstrap samples (column **Sample Mean (M)**). Note that the results in Exhibit 4.20 will differ from your results and will change when rerunning the bootstrapping procedure. The reason is that bootstrapping builds on randomly drawn bootstrap samples, which will differ every time the procedure is run. The differences in the overall bootstrapping results are marginal, however, provided that a sufficiently large number of bootstrap samples have been drawn (e.g., 5,000). The columns labeled **2.5%** and **97.5%** show the lower and upper bounds of the 95% (bias-corrected and accelerated) confidence interval. As can be seen, neither of the confidence

Do not copy, post, or distribute

Exhibit 4.19 Bootstrapping Options in SmartPLS



Do not copy, post, or distribute

Exhibit 4.20 Confidence Intervals for HTMT

Heterotrait-Monotrait Ratio (HTMT)

Mean, STDEV, T-Values, P-Values | Confidence Intervals | Confidence Intervals Bias Corrected | Samples

	Original Sample (O)	Sample Mean (M)	Bias	2.5%	97.5%
CUSA -> COMP	0.465	0.465	0.000	0.364	0.565
CUSL -> COMP	0.532	0.533	0.000	0.421	0.638
CUSL -> CUSA	0.755	0.754	-0.000	0.684	0.814
LIKE -> COMP	0.780	0.778	-0.001	0.690	0.853
LIKE -> CUSA	0.577	0.577	-0.000	0.489	0.661
LIKE -> CUSL	0.737	0.736	-0.001	0.653	0.816

intervals includes the value 1. For example, the lower and upper bounds of the confidence interval of HTMT for the relationship between *CUSA* and *COMP* are 0.364 and 0.565, respectively (again, your values will likely look slightly different because bootstrapping is a random process). As expected, since the conservative HTMT threshold of 0.85 already supports discriminant validity (Exhibit 4.18), the bootstrap confidence interval results of the HTMT criterion also clearly speak in favor of the discriminant validity of the constructs.

Exhibit 4.21 summarizes the results of the reflective measurement model assessment. As can be seen, all model evaluation criteria have been met, providing support for the measures' reliability and validity.

Exhibit 4.21 Results Summary for Reflective Measurement Models

Latent Variable	Indicators	Convergent Validity			Internal Consistency Reliability		Discriminant Validity
		Loadings	Indicator Reliability	AVE	Composite Reliability	Cronbach's Alpha	
		>0.70	>0.50	>0.50	0.60–0.90	0.60–0.90	HTMT confidence interval does not include 1
COMP	*comp_1*	0.858	0.736	0.681	0.865	0.776	Yes
	comp_2	0.798	0.637				
	comp_3	0.818	0.669				
CUSL	*cusl_1*	0.833	0.694	0.748	0.899	0.831	Yes
	cusl_2	0.917	0.841				
	cusl_3	0.843	0.711				
LIKE	*like_1*	0.879	0.773	0.747	0.899	0.831	Yes
	like_2	0.870	0.757				
	like_3	0.843	0.711				

Do not copy, post, or distribute

SUMMARY

- **Gain an overview of Stage 5 of the process for using PLS-SEM, which deals with the evaluation of measurement models.** PLS-SEM results are reviewed and evaluated using a systematic process. The goal of PLS-SEM is maximizing the explained variance (i.e., the R^2 value) of the endogenous latent variables in the PLS path model. For this reason, the evaluation of the quality of the PLS-SEM measurement and structural models focuses on metrics indicating the model's predictive capabilities. Evaluation of PLS-SEM results is a two-step approach (Stages 5 and 6) that starts with evaluating the quality of the measurement models (Stage 5). Each type of measurement model (i.e., reflective or formative) has specific evaluation criteria. With reflective measurement models, reliability and validity must be assessed (Stage 5a). In contrast, evaluation of formative measurement models (Stage 5b) involves testing the measures' convergent validity and the significance and relevance of the indicator weights as well as collinearity. Satisfactory outcomes for the measurement model are a prerequisite for evaluating the relationships in the structural model (Stage 6), which includes testing the significance of path coefficients and the **coefficient of determination** (R^2 value). Depending on the specific model and the goal of the study, researchers may want to use additional advanced analyses such as mediation or moderation, which we discuss in Chapters 7 and 8.

- **Describe Stage 5a: Evaluating reflectively measured constructs.** The goal of reflective measurement model assessment is to ensure the reliability and validity of the construct measures and therefore provide support for the suitability of their inclusion in the path model. The key criteria include indicator reliability, composite reliability, convergent validity, and discriminant validity. Convergent validity means the construct includes more than 50% of the indicator's variance. Discriminant validity means that every reflective construct must share more variance with its own indicators than with other constructs in the path model. Reflective constructs are appropriate for PLS-SEM analyses if they meet all these requirements.

- **Use the SmartPLS 3 software to assess reflectively measured constructs in the corporate reputation example.** The case study illustration uses the corporate reputation path model and the data set introduced in Chapter 2. The SmartPLS 3 software provides all relevant results for the evaluation of the measurement models.

Do not copy, post, or distribute

Tables and figures for this example demonstrate how to correctly report and interpret the PLS-SEM results. This hands-on example not only summarizes the concepts that have been introduced before but also provides additional insights for their practical application.

REVIEW QUESTIONS

1. What is indicator reliability and what is the minimum threshold value for this criterion?
2. What is composite reliability and what is the minimum threshold value for this criterion?
3. What is average variance extracted and what is the minimum threshold value for this criterion?
4. Explain the idea behind discriminant validity and how it can be established.

CRITICAL THINKING QUESTIONS

1. Why are the criteria for reflective measurement model assessment not applicable to formative measures?
2. How do you evaluate single-item constructs? Why is internal consistency reliability a meaningless criterion when evaluating single-item constructs?
3. Should researchers rely purely on statistical evaluation criteria to select a final set of indicators to include in the path model? Discuss the trade-off between statistical analyses and content validity.

KEY TERMS

AVE
Average variance extracted (AVE)
Bootstrap confidence interval
Bootstrapping
Coefficient of determination (R^2)
Collinearity
Communality (construct)
Communality (item)
Composite reliability
Content validity

Do not copy, post, or distribute

Convergent validity
Cronbach's alpha
Cross-loadings
Disattenuated correlation
Discriminant validity
Evaluation criteria
Explained variance
f^2 effect size
Formative measurement models
Fornell-Larcker criterion
Heterotrait-heteromethod correlations
Heterotrait-monotrait ratio (HTMT)
HTMT
Indicator reliability
Internal consistency reliability
Monotrait-heteromethod correlations
Predictive relevance (Q^2)
q^2 effect size
Q^2 value
R^2 value
Reliability
Validity

SUGGESTED READINGS

Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly, 25,* 1–16.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–358). Mahwah, NJ: Erlbaum.

Chin, W. W. (2010). How to write up and report PLS analyses. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications in marketing and related fields* (Springer Handbooks of Computational Statistics Series, Vol. II, pp. 655–690). Berlin: Springer.

Do Valle, P. O., & Assaker, G. (in press). Using partial least squares structural equation modeling in tourism research: A review of past research and recommendations for future applications. *Journal of Travel Research*.

Gefen, D., Rigdon, E. E., & Straub, D. W. (2011). Editor's comment: An update and extension to SEM guidelines for administrative and social science research. *MIS Quarterly, 35,* iii–xiv.

Götz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer Handbooks of Computational Statistics Series, Vol. II, pp. 691–711). Berlin: Springer.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19,* 139–151.

Do not copy, post, or distribute

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Planning, 46,* 1–12.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science, 40,* 414–433.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43,* 115–135.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing, 20,* 277–320.

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). A critical look at the use of PLS-SEM in *MIS Quarterly. MIS Quarterly, 36,* iii–xiv.

Roldán, J. L., & Sánchez-Franco, M. J. (2012). Variance-based structural equation modeling: Guidelines for using partial least squares in information systems research. In M. Mora, O. Gelman, A. L. Steenkamp, & M. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp. 193–221). Hershey, PA: IGI Global.

Straub, D., Boudreau, M. C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems, 13,* 380–427.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48,* 159–205.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44,* 119–134.

Do not copy, post, or distribute