A Tutorial on the Meta-Analytic Structural Equation Modeling of Reliability Coefficients

Ronny Scherer

University of Oslo, Norway

Timothy Teo

Murdoch University, Australia

Author Note

Ronny Scherer, Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo; Timothy Teo, School of Education, Murdoch University, Australia.

Correspondence concerning this article should be addressed to Ronny Scherer, University of Oslo, Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), Postbox 1161 Blindern, 0318 Oslo; E-Mail: ronny.scherer@cemo.uio.no

Timothy Teo, Murdoch University, 90 South Street, Murdoch, Western Australia 6150, Australia; Email: timothy.teo@murdoch.edu.au

This paper was accepted for publication in the journal *Psychological Methods*:

Scherer, R., & Teo, T. (2020). A Tutorial on the Meta-Analytic Structural Equation Modeling of Reliability Coefficients. *Psychological Methods*, http://dx.doi.org/10.1037/met0000261

Abstract

Reporting the reliability of the scores obtained from a scale or test is part of the standard repertoire of empirical studies in psychology. With reliability being a key concept in psychometrics, researchers have become more and more interested in evaluating reliability coefficients across studies and, ultimately, quantify and explain possible between-study variation. This approach—commonly known as 'reliability generalization'—can be specified within the framework of meta-analysis. The existing procedures of reliability generalization, however, have several methodological issues: (a) Unrealistic and often untested assumptions on the measurement model underlying the reliability coefficients (e.g., essential τ-equivalence for Cronbach's α), (b) the use of univariate approaches to synthesizing reliabilities of total and subscale scores, (c) the lack of comparability across different types of reliability coefficients. However, these issues can be addressed directly through meta-analytic structural equation modeling (MASEM)—a method that combines meta-analysis with structural equation modeling through synthesizing either correlation matrices or model parameters across studies. The primary objective of this paper is to present the potential MASEM has for the meta-analysis of reliability coefficients. We review the extant body of literature on the use of reliability generalization, discuss and illustrate two MASEM approaches (i.e., correlation-based and parameter-based MASEM), and propose some practical guidelines. Future directions for utilizing MASEM for reliability generalization are discussed.

Keywords: Meta-analysis; meta-analytic structural equation modeling (MASEM); omega coefficient; reliability generalization; scale reliability

Translational Abstract

To evaluate whether scores obtained from psychological tests or scales are reliable is key to psychological assessment. With growing numbers of tests and scales being administered to different samples, reliability indicators can vary considerably between studies. As a consequence, researchers developed meta-analytic techniques in order to synthesize these reliability coefficients to an overall coefficient, quantify its variation between studies, and explain this variation across studies. These techniques, however, have several methodological issues, one of which refers to their reliance on reliability indicators that may not appropriately represent the data (e.g., Cronbach's α). However, a relatively new meta-analytic technique—meta-analytic structural equation modeling (MASEM)—can overcome these issues. In this paper, we two different forms of MASEM and review their potential for the meta-analysis of reliabilities. Two examples illustrate their application to empirical data. We further provide some guidelines on how to perform MASEM to synthesize reliability coefficients. Detailed supplementary material (including R code) is provided so that researchers can replicate our findings and transfer the syntax to their research contexts. We hope to stimulate the use of MASEM for the meta-analysis of reliability coefficients.

A Tutorial on the Meta-Analytic Structural Equation Modeling of Reliability Coefficients

Introduction

With growing numbers of empirical studies administering psychological assessments, evaluating the psychometric quality of tests and scales has become a standard for psychological testing (AERA, APA, & NCME, 2014). Part of this evaluation is the estimation of reliability coefficients that describe the consistency of the test or scale scores across contexts (e.g., items, test parts, measurement occasions; e.g., Cronbach, 1947). As prominent measures of psychological constructs, such as standardized tests of cognitive skills or scales capturing personality traits, have generated numerous studies (e.g., J. Deng et al., 2019; Gnambs, 2014; Wheeler, Vassar, Worley, & Barnes, 2011), researchers have become more and more interested in evaluating reliability coefficients *across* studies (López-López, Botella, Sánchez-Meca, & Marin-Martínez, 2013; M. C. Rodriguez & Maeda, 2006). This approach is known as "reliability generalization" and represents a form of psychometric meta-analysis (Hunter & Schmidt, 2014). More specifically, reliability generalization synthesizes reliability coefficients across studies, mainly Cronbach's α, to an overall reliability estimate, quantifies possible between-study variation through random-effects models, and examines moderator effects through mixed-effects models (Botella, Suero, & Gambara, 2010).

The existing, primarily univariate approaches to reliability generalization, however, have several limitations, such as the reporting of different types of reliability coefficients across studies, the ignorance of the dependencies among reliability coefficients of total and subscale scores, and the possible violations of key assumptions associated with the well-established reliability coefficient of Cronbach's α. Especially the latter has raised concerns in the psychometric community, resulting in a call for selecting model-based reliability coefficients that circumvent the oftentimes unrealistic assumptions behind Cronbach's α (Bentler, 2016; McNeish, 2018; Yang & Green, 2011). Considering this and the limitations

univariate approaches to reliability generalization have, we argue that meta-analytic structural equation modeling (MASEM)—a relatively new approach to meta-analysis that brings together the strengths of meta-analyses and structural equation models to test theories and hypotheses based on models (Cheung, 2015; Jak & Cheung, 2018b)—can overcome these issues and provides a flexible tool for meta-analyzing reliability coefficients across studies. Specifically, MASEM allows researchers to synthesize correlation matrices and/or model parameters across studies, taking into account the dependencies among correlations and quantifying between-study heterogeneity (Cheung & Cheung, 2016). Given its multivariate nature, MASEM circumvents possible erroneous inferences that may be drawn from univariate meta-analytic approaches (Jak & Cheung, 2019). Moreover, MASEM allows researchers to specify and compare measurement models underlying the reliability estimation and, ultimately, test whether the assumptions behind reliability coefficients are met—in other words, through MASEM, researchers can select a suitable reliability coefficient based on the factor structure of the test or scale.

The main goals of this tutorial paper are therefore to (a) review the existing approaches to meta-analyzing reliability coefficients across studies and (b) illustrate and discuss the potential of MASEM for reliability generalization. We present two illustrative data examples and show how different MASEM approaches (i.e., correlation- and parameter-based MASEM) can be performed to obtain an overall estimate of the score reliability. Furthermore, we propose steps researchers can take to evaluate the reliability coefficients meta-analytically. These steps depend on the features of the meta-analytic data and the researchers' goals, and include (a) specifying a measurement model that represents the factor structure of the data, (b) pooling correlation matrices or model parameters across studies with fixed or random effects, and (c) deriving an overall reliability estimate from the model parameters.

Scale Reliability

Reliability is one of the key concepts in the evaluation of scale scores in psychology, education, and other fields. In the context of Classical Test Theory, scale scores X are usually conceptualized as the sum of item scores $X = \sum_{j=1}^{J} X_j$ ($j = 1, \ldots, J$ items) and are decomposed into the scale true score T_X and the scale error score E, $X = T_X + E$ (Novick & Lewis, 1967). Under the assumption of the independence between the true and the error scores, the scale reliability ρ_X is defined as the ratio between the true scale score variance and the observed scale score variance, $\rho_X = \sigma^2(T_X)/\sigma^2(X)$. By far, the most common estimate of reliability, in the form of an internal consistency coefficient, is Cronbach's α (McNeish, 2018). For a scale comprising J items with variances s_j^2 and the variance of the observed sum score s_X^2, the estimated coefficient of Cronbach's α can be defined as (Cronbach, 1951):

$$\alpha = \left(\frac{J}{J-1}\right)\left(1 - \frac{\sum_{j=1}^{J} s_j^2}{s_X^2}\right) \qquad (1)$$

As Kelley and Pornprasertmanit (2016) noted, Cronbach's α is a consistent estimate of the population internal consistency if certain assumptions hold. One of the key assumptions is that a one-factor model with uncorrelated residuals and equal factor loadings across items is a suitable representation of the data—an assumption that has been criticized as being unrealistic in most situations (e.g., Bentler, 2017; Yang & Green, 2011). More specifically, McNeish (2018) summarized the assumptions behind Cronbach's α as follows: (1) τ-equivalence (i.e., items contribute equally to the scale score, they have the same factor loadings); (2) normally distributed, continuous items within the scale; (3) independent errors (i.e., item error scores are not correlated); (4) unidimensionality (i.e., a one-factor model represents that data). Violations of these assumptions can bias Cronbach's α positively or negatively (Yang & Green, 2011).

Suppose that J indicator variables $X_1, \ldots, X_J$ (i.e., items) follow a one-factor model (Figure 1). For the *i*th person ($i = 1, \ldots, N$) and the *j*th indicator variable ($j = 1, \ldots, J$), the factor model is defined as follows (Bollen, 1989; McDonald, 1999)

$$X_{ij} = \mu_j + \lambda_j \eta_i + e_{ij} \qquad (2)$$

where μ_j is the population mean of the indicator X_j, λ_j the factor loading of the indicator X_j, η_i the factor score of the *i*th person, and e_{ij} the residual for the *i*th person and the *j*th indicator with variance $Var(e_{ij}) = \theta_{jj}$. Under the assumptions that the factor score variance is fixed to 1 and that item residuals are uncorrelated, the reliability coefficient omega total ω_T is calculated as:

$$\omega_T = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} \theta_{jj}} \qquad (3)$$

If indeed a congeneric factor model—that is a one-factor model with freely estimated factor loadings, freely estimated residual variances, and uncorrelated residuals—holds, omega total is an appropriate coefficient of the scale reliability, that is, the reliability of the scale sum score $X = X_1 + \cdots + X_J$ (Kelley & Pornprasertmanit, 2016; McDonald, 1999). In the psychometric literature, this reliability coefficient is often referred to as "scale reliability", "scale score reliability", or "composite reliability" (McDonald, 1999; Raykov, 2004). Notice that these terms refer to the reliability of a scale *score* derived from a scale or test—they represent the degree to which these scores are free of measurement error.

Geldhof, Preacher, and Zyphur (2014) noted that the formula for omega total is identical to that of Cronbach's α under essential τ-equivalence, that is, the assumption of equal factor loadings across items in a one-factor model, and perfect model fit (see also Graham, 2006; Raykov, 1997). Cronbach's α and the composite reliability coefficients share conceptual similarities as they both represent the ratio between the variance of the true score and the total variance (Geldhof et al., 2014). From a factor-analytic perspective, the

equivalence assumption can be tested explicitly by comparing the one-factor model with the equal factor loadings to the one with freely estimated loadings (Yang & Green, 2011). If indeed the equivalence assumption holds and the model itself fits the data, researchers may calculate Cronbach's α from the parameters of the corresponding factor model (Raykov & Marcoulides, 2014). Nevertheless, if the assumption is violated, Cronbach's α underestimates the score reliability (Gu, Little, & Kingston, 2013; Sijtsma, 2008). Although the omega coefficient may be more appropriate to describe the score reliability of a scale and test (e.g., $\alpha \leq \omega_T$) (L. Deng & Chan, 2016), Bentler (2017) warns against its use if the congeneric one-factor model does not fit the data. As a consequence, several versions of the omega coefficient have been developed to account for possible deviations from the unidimensional congeneric model (Padilla & Divers, 2015; Teo & Fan, 2013; Zinbarg, Revelle, Yovel, & Li, 2005).

Indeed, in real-data situations, the assumption of a one-factor measurement model representing the factor structure of a scale may be compromised by, for instance, covariances among item residuals or nested factors. If residuals are correlated, the calculation of ω_T can be adjusted by accounting for the error covariances (see McNeish, 2018; Raykov, 2004):

$$\omega_{TCov} = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \sum_{j=1}^{J} \theta_{jj} + 2\sum_{j=2}^{J}\sum_{l=1}^{j} \theta_{jl}} \qquad (4)$$

where θ_{jl} represents the covariance between the residuals of items j and l. In cases where nested factors $s_1, \ldots, s_p$ exist next to a general factor g, the omega coefficient can be modified to represent the reliability of the overall scale, yet accounting for the loadings on the specific factors. Assuming this bifactor model structure with uncorrelated factors, Gignac (2014) formulated the resultant reliability coefficient, omega hierarchical ω_H, as follows:

$$\omega_H = \frac{\left(\sum_{j=1}^{J} \lambda_j^{(g)}\right)^2}{\left(\sum_{j=1}^{J} \lambda_j^{(g)}\right)^2 + \sum_{j=1}^{J} \theta_{jj} + \left(\sum_{j=1}^{J_1} \lambda_j^{(s_1)}\right)^2 + \cdots + \left(\sum_{j=1}^{J_p} \lambda_j^{(s_p)}\right)^2} \qquad (5)$$

where $\lambda_j^{(g)}$ represents the loading of item j on the general factor g, and $\lambda_j^{(s_1)}, \ldots, \lambda_j^{(s_p)}$ represent the loadings of item j on the specific factors $s_1, \ldots, s_p$. The specific factors comprise $J_1, \ldots, J_p$ items. This omega coefficient represents the ratio between the squared sum of factor loadings of the general factor and the model-estimated total variance (A. Rodriguez, Reise, & Haviland, 2016), and the factor loadings can be obtained from a bifactor model or a Schmid-Leiman transformed, second-order factor model (Gignac, Reynolds, & Kovacs, 2017; Reise, 2012). Green and Yang (2015) noted that ω_H "allows researchers to state the degree that summed item scores are saturated by the general factor"—in this sense, higher values of ω_H point to the "scale scores as due to the general factor" (p. 16).

These coefficients have been developed further, and many alternative coefficients exist that take into account deviations from the essentially τ-equivalent model (Bentler, 2017). Overall, the psychometric literature points to the usefulness of structural equation modeling approaches (primarily confirmatory factor analysis) to obtaining model-based reliability coefficients of test and scale scores and the need for testing the assumptions underlying the famous Cronbach's α coefficient.

Approaches to Meta-Analyzing Scale Reliabilities

This section presents the state-of-the-art of approaches to synthesizing reliability coefficients meta-analytically and reviews their potential and limitations. We begin by reviewing univariate approaches to synthesizing scale reliabilities and their underlying transformations of the reliability coefficients. Next, we present a multivariate approach proposed by Raykov and Marcoulides (2013).

Univariate Meta-Analysis of Scale Reliability

The process of synthesizing reliability coefficients through meta-analytic procedures, quantifying between-study heterogeneity, and possibly explaining this heterogeneity with study features (i.e., moderators) is referred to as 'reliability generalization' (Vacha-Haase,

1998). With much of the psychological literature reporting Cronbach's α as a reliability coefficient, the extant literature has focused on developing reliability generalization procedures for this coefficient using univariate meta-analysis (Holland, 2015; M. C. Rodriguez & Maeda, 2006; Sánchez-Meca, López-López, & López-Pina, 2013). Among other features, these procedures vary with respect to the transformation of the α-coefficient to normalize the distribution or stabilize the corresponding variance (Botella et al., 2010; López-López et al., 2013). For instance, several transformations were proposed, including Fisher's Z-transformation, which is based on the assumption that Cronbach's α can be considered a correlation coefficient, the Hakstian and Whalen (1976) T-transformation, and the Bonett (2002) transformation. Table 1 shows the details of these transformations, their back-transformation, and the estimate of the sampling variances. Some simulations studies suggested that these transformations perform similarly in mixed-effects model, especially when estimating the regression coefficients of moderators (López-López et al., 2013). Other studies showed the preference of the Hakstian-Whalen and Bonett transformations as opposed to Fisher's Z-transformation and the use of raw reliability coefficients (Sánchez-Meca et al., 2013). Comparing thirteen statistical models for synthesizing Cronbach's α across studies (i.e., models specified for different estimators, fixed vs. random effects, and with or without a transformation), Sánchez-Meca et al. (2013) concluded that "there is (…) no evidence base upon which to rule out or advocate the use of transformation(s)" (p. 405).

Despite the variety of reliability generalization approaches, the current developments to improve them (e.g., Brannick & Zhang, 2013), and the impressive body of literature, the dominant approaches have at least two issues: First, they mainly rely on Cronbach's α as an appropriate coefficient to describe the scale reliability, although alternative coefficients may be more appropriate. Second, when multiple reliability coefficients of, for instance, subscale scores are reported, they do not consider the dependencies among them. Nevertheless, the key

strength of the univariate approaches lies in the straightforward estimation of the between-study variance and the inclusion of possible moderators (López-López et al., 2013).

Synthesizing Correlations in Separate Univariate Meta-Analyses

When the correlations among test items are made available by the authors of primary studies, an overall correlation matrix can be obtained (Carpenter, Son, Harris, Alexander, & Horner, 2016). This overall correlation matrix is then used for structural equation modeling—essentially, researchers can specify factor models and estimate the reliability of the scale or test across studies from the factor loadings, factor variances, and residual variances. In the early works of meta-analytic structural equation modeling, several independent, univariate meta-analyses of the correlations were conducted to populate the overall correlation matrix (Sheng, Kong, Cortina, & Hou, 2016; Viswesvaran & Ones, 1995). This approach has several limitations: First, multiple yet separate meta-analyses of single correlation coefficients are oftentimes based on different sample sizes, because some primary studies may only provide some correlation coefficients, yet the full set of correlations among all relevant variables. The resultant, pooled correlation matrix is then submitted to structural equation modeling with some arbitrary sample size (e.g., the arithmetic or harmonic mean of all sample sizes) that may impact the precision of model parameters (Cheung, 2015). Second, separate meta-analyses may result in a non-positive definite pooled correlation or covariance matrix—such a matrix cannot be submitted to structural equation modeling (Kline, 2016). Third, separate meta-analyses are based on the assumption that the correlations within a correlation matrix are independent—violations of this independence can bias especially the standard errors of the structural equation model parameters (Cheung, 2015). Considering these issues, more and more evidence has surfaced that indicates the clear preference of multivariate rather than univariate approaches to populating the item-item correlation matrix (e.g., Cheung & Chan, 2005, 2009; Sheng et al., 2016). Cheung (2015) noted that some authors identified conditions

under which the univariate and multivariate approaches may provide similar results (see Hafdahl, 2007; Ishak et al., 2008). Nevertheless, Tang and Cheung (2016) illustrated how separate univariate meta-analyses of correlations among constructs can lead to entirely different results than the more precise multivariate approaches. Overall, populating an overall correlation matrix through multiple, independent meta-analyses in the first step and submitting the resultant matrix to structural equation modeling in the second step to retrieve reliability coefficients has several issues and may bias the structural equation model parameters from which reliability coefficients are derived.

Raykov's and Marcoulides' (2013) Multivariate Approach

Raykov and Marcoulides (2013) proposed a multivariate approach to the meta-analysis of scale reliability. The authors specified a multi-group confirmatory factor-analytic model to the covariance matrices of the primary studies, treating studies as groups. Under scalar invariance constraints (i.e., the same configuration of the measurement model, equal factor loadings and item intercepts across studies), an overall scale reliability estimate, its standard error, and confidence interval can be obtained. However, if these invariance constraints are not met, comparisons of reliability coefficients across groups may be compromised (Raykov, 2004). Besides the invariance of the measurement model that underlies the reliability estimation, Raykov and Marcoulides (2013) also considered reliability invariance a prerequisite for synthesizing scale reliabilities to an overall reliability estimate.

From our perspective, Raykov's and Marcoulides' (2013) approach has several strengths: First, it is based on a latent variable model and therefore circumvents issues related to the comparability of different reliability coefficients: Different reliability coefficients are not meta-analyzed at the same time, but a uniform coefficient is obtained from the existing covariance matrices. Second, the approach is based on invariance constraints, ensuring the comparability of the measurement model and thus the meaningful interpretation of scale

reliabilities across the primary studies. At the same time, Raykov's and Marcoulides' (2013) approach has several weaknesses: First, the measurement invariance testing is based on covariance matrices and requires that the primary study reports include a correlation matrix and the descriptive statistics of the underlying variables. This however also requires that the same measurement instrument has to be used without any modifications of, for instance, item formulations or response categories in all studies (Cheung, 2015). In some instances, however, partially invariant item indicators (e.g., with adapted item wordings or response categories) may be sufficient to achieve sufficient degrees of comparability. Second, although scalar invariance across primary studies is a desirable characteristic of the measurement model, it is hardly achieved in meta-analytic datasets with many studies (see also Marsh, Guo, et al., 2018). To our best knowledge, however, the consequences of deviations from measurement invariance are still to be examined for meta-analytic data that are based on correlation rather than covariance matrices. As we will explain in one of the empirical examples, metric invariance can imply reliability invariance, depending on the model specification. Third, the approach is based on the assumption of fixed effects in the pooled covariance matrices—an assumption that may not be met in many meta-analytic situations (Cheung & Cheung, 2016). Due to the invariance constraints, reasons for possible misfit of the measurement model cannot be clearly attributed to the variation of the covariance or correlation matrices between studies, model misspecification, or both (Cheung, 2015).

Current Practices of Reliability Generalization and Related Issues

To provide an overview of the current practices of meta-analyzing reliability coefficients and back the claim that univariate approaches to reliability generalization have dominated this field, we performed a review of the extant literature, focusing on the meta-analytic approaches taken to synthesize reliabilities (e.g., type of reliability coefficient, aggregation method, transformation of reliability coefficients, measurement model check,

testing further assumptions underlying the selected reliability coefficient). Supplementary Material S3 provides a detailed summary of these approaches, and Supplementary Material S4 contains the list of publications, the search and screening history, and the coding of these variables. Overall, our review of extant literature indicated (a) the dominance of univariate approaches to synthesizing reliability coefficients, even in generalizing multiple reliability coefficients (e.g., of overall scores and subscale scores), (b) the lack of testing the assumptions underlying these coefficients, and (c) the preference of reporting Cronbach's α as a reliability coefficient.

The dominance of univariate approaches to meta-analyzing mainly Cronbach's α seems problematic because the extant psychometric literature suggests that the assumptions behind the most commonly used measure of internal consistency, Cronbach's α, are oftentimes not met (McNeish, 2018; Sijtsma, 2008). Model-based approaches to deriving overall reliability estimates could circumvent this issue by exploring the factor structure of a scale or test and selecting an appropriate reliability coefficient based on this structure—however, these approaches were hardly taken into consideration. Ignoring the dependencies among multiple covariances or correlations within studies is another issue associated with univariate meta-analysis. Raykov's and Marcoulides' (2013) multivariate approach circumvents the problematic univariate assumption and provides a model-based reliability estimate—however, this approach includes only fixed effects and requires covariance matrices. MASEM offers feasible solutions to these issues (Cheung, 2015).

Meta-Analytic Structural Equation Modeling of Reliability Coefficients

In their review of the extant literature on random-effects models for meta-analytic structural equation modeling (MASEM), Cheung and Cheung (2016) noted that two forms of MASEM have dominated the field: correlation-based and parameter-based MASEM. The former synthesizes the correlation or covariance matrices across all primary studies and

ultimately results in an overall correlation or covariance matrix which is submitted to structural equation modeling. The latter performs structural equation modeling to retrieve the model parameters researchers are interested in (e.g., indirect effects, factor loadings, scale reliabilities) and, subsequently, meta-analyzes these model parameters. Correlation-based MASEM includes several approaches, such as one- and two-stage approaches. In the following, we will describe the two forms of MASEM and review their potential for the meta-analysis of reliability coefficients. We will however not discuss the recently developed full-information MASEM approach (Yu, Downes, Carter, & O'Boyle, 2016), due to its performance issues with specifying structural equation models to a pooled correlation matrix—although this bootstrapping approach estimates credibility intervals of model parameters reasonably well, test statistics and goodness-of-fit indices may be inaccurate (Cheung, 2018a).

In a nutshell, the following sections convey at least two key advantages of correlation- and parameter-based MASEM over the univariate meta-analysis of Cronbach's α: First, both MASEM approaches allow and require researchers to specify an appropriate measurement model and to select the corresponding reliability coefficient rather than relying on Cronbach's α or other, possibly inappropriate coefficients reported in the primary studies. Second, taking into account the multivariate nature of the data (i.e., multiple correlations or model parameters per study) reduces the risk of bias in reliability estimates.

Correlation-Based MASEM

Correlation-based MASEM has experienced many developments, and several approaches exist to populate an overall correlation matrix. As noted earlier, Viswesvaran and Ones (1995) proposed a two-step approach that consisted of several, separate univariate meta-analyses for each correlation within a correlation matrix in the first step, and the use of the resultant correlation matrix to specify structural equation models. Although this procedure has

several limitations (e.g., ignoring the dependence among correlation coefficients, handling inefficient handling of missing data), it is still applied in practice (Sheng et al., 2016). In her overview of correlation-based MASEM, Jak (2015) argued that the generalized least squares (GLS) method proposed by Becker (1992) addresses some of these limitations, in particular the dependencies between correlation coefficients. However, this method does not allow researchers to specify and estimate latent variable models (Jak & Cheung, 2019). As a response, Cheung and Chan (2005) developed a two-stage structural equation modeling approach which circumvents these issues and opens a broader range of structural equation models for meta-analysis.

Two-stage meta-analytic structural equation modeling (TSMASEM). The first stage of this approach combines the correlation matrices from the primary studies using fixed- or random-effects multivariate meta-analysis (Cheung, 2015). In the following, we will focus on the *random-effects model* as an oftentimes more realistic representation of the meta-analytic data (Cheung & Cheung, 2016). Vectorizing the correlation matrix from the *i*th primary study $\boldsymbol{\rho}_i = vechs(\boldsymbol{P}_i)$, where $\boldsymbol{P}_i$ represents each study's population correlation matrix, Cheung (2015) describes this first stage for random-effects model assuming that $\boldsymbol{e}_i \sim N\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{e}_i}\right)$ and $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{u}})$ as follows:

Level 1: $$\boldsymbol{r}_i = \boldsymbol{\rho}_i + \boldsymbol{e}_i \quad (6)$$

Level 2: $$\boldsymbol{\rho}_i = \boldsymbol{\rho}_R + \boldsymbol{u}_i \quad (7)$$

In this model, $\boldsymbol{V}_{\boldsymbol{e}_i}$ denotes the sampling covariance matrix and $\boldsymbol{V}_{\boldsymbol{u}}$ the heterogeneity covariance matrix obtained from the random effects $\boldsymbol{u}_i$. $\boldsymbol{\rho}_R = \boldsymbol{\rho}(\theta)$ represents the pooled population correlation matrix under the random-effects model which forms the basis for the structural model with parameters $\boldsymbol{\theta}$ in the second stage. Through maximum-likelihood estimation, this model provides the average correlation matrix $\widehat{\boldsymbol{\rho}}_R$, the asymptotic sampling covariance matrix $\widehat{\boldsymbol{V}}_R$, and the covariance matrix of between-study heterogeneity $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$. Notice

that the covariance matrices $\widehat{\boldsymbol{V}}_R$ and $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$ can become very large the more variables are included. For p variables, the average correlation matrix is a symmetric $p \times p$-matrix with $p(p-1)/2$ off-diagonal correlations, $\widehat{\boldsymbol{V}}_R$ is a symmetric $p \times p$-matrix, and $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$ represents a symmetric $[p(p-1)/2] \times [p(p-1)/2]$-matrix. As a consequence, Cheung (2015) suggested adding some constraints, for instance, to the matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$.

Some features of the model estimation are worth noting: This estimation is based on the assumption of multivariate normality of the data. TSMASEM can accommodate missing correlations in correlation matrices efficiently through maximum-likelihood estimation procedures if they are completely at random or at random (Cheung, 2015; Jak & Cheung, 2018a). Moreover, the model equations (6) and (7) do not per se ensure that all correlation matrices obtained from the primary studies are positive definite—instead, an initial check for positive definiteness must be performed before the pooling of correlation matrices. Furthermore, the stage-1 model uses the correct sample sizes, that is, the sample sizes of the primary studies rather than their arithmetic or harmonic means. Cheung and Cheung (2016) note that the stage-1 model incorporates the between-study heterogeneity, thus allowing for random effects in the correlation coefficients. In other words, the log-likelihood that is used during the maximum-likelihood estimation is a function of both $\boldsymbol{\rho}_R$ and $\boldsymbol{V}_{\boldsymbol{u}}$, $\ell(\boldsymbol{\rho}_R, \boldsymbol{V}_{\boldsymbol{u}}) = \log \mathcal{L}(\boldsymbol{\rho}_R, \boldsymbol{V}_{\boldsymbol{u}})$ (Cheung, 2014). Alternatively, researchers may consider using restricted maximum-likelihood (REML) estimation to minimize the possible negative bias in the variance components in the maximum-likelihood estimation. As REML removes the fixed effects before estimating the variance components, ad-hoc procedures are necessary to estimate them (Cheung, 2015). This procedure limits comparisons between models with the likelihood-ratio test to those that differ only in their variance components. At the time of writing, REML still requires a stable implementation for its application in correlation-based MASEM (see also Cheung, 2013). In practice, the covariance matrix of random effects can

become large (especially when many variables are included in the model). To circumvent convergence issues and to reduce the number of model parameters, researchers may impose constraints on this covariance matrix by, for instance, allowing only diagonal elements (Cheung, 2015). These constraints however may not necessarily reflect on the real nature of the random effects, as they may indeed covary (Jak, 2015). Researchers must be aware of this limitation when specifying the stage-1 model.

The second stage of the TSMASEM approach comprises the fitting of the structural equation model to the average correlation matrix $\widehat{\boldsymbol{\rho}}_R$ and the asymptotic sampling covariance matrix $\widehat{\boldsymbol{V}}_R$. Cheung (2015) suggested using weighted least squares (WLS) estimation with the discrepancy function $F_{WLS}(\theta) = (\boldsymbol{r}_R - \widehat{\boldsymbol{\rho}}_R)^T \widehat{\boldsymbol{V}}_R^{-1} (\boldsymbol{r}_R - \widehat{\boldsymbol{\rho}}_R)$ to specify and estimate the structural equation models. This estimation procedure weighs the correlations by the inverse of the asymptotic covariance matrix $\widehat{\boldsymbol{V}}_R^{-1}$. WLS estimation treats the pooled correlation matrix correctly, that is, as a matrix of correlation rather than variances and covariances to circumvent incorrect standard errors and inferences drawn from significant tests. To ensure the correct estimation, several nonlinear constraints can be imposed on the model-implied correlation matrix, such as constraining its diagonal elements and the factor variance to 1 (Cheung, 2015; Steiger, 2002). In contrast to the stage-1 model estimation, the stage-2 model does not rely on the assumption of multivariate normality. However, Oort and Jak (2016) found that the performance of maximum-likelihood estimation in stage 2 under this assumption is similar to that of the WLS estimation. Given the size of the weight matrices, WLS estimation may also require larger sample sizes than maximum-likelihood estimation (Olsson, Foss, Troye, & Howell, 2000).

It should also be noted that the covariance matrix of random effects $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$ is not considered in the stage-2 model estimation; yet, it is a part of the stage-1 pooling of the correlation matrices (Cheung & Chan, 2005; Cheung, 2015). It may therefore not have a

major impact on the estimates of $\widehat{\boldsymbol{\rho}}_R$. To compare competing models, several likelihood-based goodness-of-fit indices and the likelihood-ratio test are available. In case researchers want to specify a *fixed-effects model* to pool the correlation matrices across studies, the matrix of random effects is a null matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{u}} = \mathbf{0}$. Similar to the random-effects model, the stage-1 pooling is based on maximum-likelihood estimation and results in a correlation matrix $\widehat{\boldsymbol{\rho}}_F$ and an asymptotic sampling covariance matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{F}}$, both of which are then submitted to the stage-2 specification of the structural equation model. The matrix $\widehat{\boldsymbol{V}}_F^{-1}$ serves as the weighting matrix in the WLS estimation.

Reviewing the TSMASEM approach, we see its potential for the meta-analysis of reliability coefficients (Table 3): TSMASEM results in an overall correlation matrix that is aggregated across the primary studies, accounting for dependencies among correlations and accommodating for missing data and heterogeneity. This correlation matrix can form the basis for selecting a suitable factor model that represents the structure of a test or a scale. On the basis of this factor model, researchers can then extract the relevant model parameters (i.e., factor loadings, factor variance, and residual variances) to estimate an overall reliability coefficient. However, this overall estimate does not include the between-study heterogeneity, because the random effects are part of the stage-1 model only—as a consequence, the exploration of moderator effects on the reliability coefficient is currently limited to subgroup analyses in stage 2 (Table 2). To summarize, TSMASEM for reliability generalization requires researchers to (a) synthesize correlation matrices based on a fixed- or random-effects model in the first stage (including the checking for positive definiteness), (b) decide for a final stage-1 model (fixed- vs. random-effects model), (c) specify, estimate, and evaluate the congeneric factor model or alternative measurement models in stage 2, (d) decide for a final stage-2 model, and (e) estimate the reliability coefficient based on the final stage-2 model. These steps are described in more detail in Table 3.

It is important to emphasize that the correlations are the sources of heterogeneity in TSMASEM. Considering this, TSMASEM does not (yet) allow researchers to quantify between-study heterogeneity of the model parameters (i.e., factor loadings and reliability coefficients) directly, because the random effects of correlations are not part of the stage-2 model. Moreover, the relationship between random effects of correlations and model parameters is nonlinear (Cheung, 2015). The current specification of TSMASEM is therefore based on the assumption that the measurement model is an adequate representation of the pooled correlation matrix (i.e., fixed effects) and its asymptotic covariance matrix. Besides, the configuration of the measurement model and the factor loadings are assumed to be invariant across studies. Nevertheless, researchers can perform metric invariance testing across subgroups of primary studies and ultimately compare reliability coefficients (Jak, 2015; Jak & Cheung, 2018b).

One-stage meta-analytic structural equation modeling (OSMASEM). Recently, Jak and Cheung (2019) developed the one-stage analogue to TSMASEM—namely one-stage MASEM—which performs the pooling of correlation matrices and the estimation of the structural equation model in one step. Based on the same multivariate random-effects model as in the first stage of TSMASEM (see equations [6] and [7]), correlations are decomposed into (Jak & Cheung, 2019): $\boldsymbol{r}_i = \boldsymbol{\rho}_R + \boldsymbol{e}_i + \boldsymbol{u}_i$, with $\boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{e}_i})$, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{u}})$, sampling covariance matrix $\boldsymbol{V}_{\boldsymbol{e}_i}$ and the covariance matrix of random effects $\boldsymbol{V}_{\boldsymbol{u}}$. Through maximum-likelihood estimation, this model provides the average correlation matrix $\widehat{\boldsymbol{\rho}}_R$, the asymptotic sampling covariance matrix $\widehat{\boldsymbol{V}}_R$, and the covariance matrix of between-study heterogeneity $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$. Once again, as in the TSMASEM approach, the resultant, pooled correlation matrix is used to specify and estimate the structural equation model of interest. Using the RAM-formulation, Jak and Cheung (2019) represent this restriction as follows: With $\boldsymbol{A}$ representing the matrix of factor loadings and path coefficients (so-called "asymmetric paths"), $\boldsymbol{S}$

representing the matrix of variances and covariances, $\boldsymbol{F}$ representing a matrix selecting between latent and manifest variables, and $\boldsymbol{I}$ representing an identity matrix, the pooled correlation matrix $\boldsymbol{\rho}_R$ is expressed as

$$\boldsymbol{\rho}_R = \text{vechs}(\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{A})^{-1T}\boldsymbol{F}^T) \qquad (8).$$

Both the $\boldsymbol{A}$ and $\boldsymbol{S}$ matrices can be regressed on study-level moderator variables. Jak and Cheung (2019) noted that this approach follows the logic of moderated nonlinear factor analysis, in which parameters in a factor model are expressed as functions of categorical or continuous grouping variables (Bauer, 2017). Ultimately, including moderator variables is aimed at explaining between-study heterogeneity in the correlation coefficients. If indeed equation (8) is extended by regression models of $\boldsymbol{A}$ and $\boldsymbol{S}$, the resultant covariance matrix of random effects becomes a covariance matrix of residuals. Examining the reduction of between-study variation after adding the moderators provides information about the variance explanation (through a pseudo-R^2 value). OSMASEM includes the random effects of the correlations and handles missing data through maximum-likelihood estimation (see Table 2). The pooling of the correlation matrices and the estimation of the structural equation model are based on maximum-likelihood estimation and the assumption of multivariate normality. For more in-depth explanations and hands-on examples of OSMASEM, we kindly refer readers to Jak and Cheung (2019).

For the meta-analysis of reliability coefficients, OSMASEM offers possibilities to obtain an overall reliability estimate (see Table 3). More specifically, a factor model can be specified and estimated based on the pooled correlation matrices, and a suitable reliability coefficient extracted from the resultant model parameters. In contrast to TSMASEM, OSMASEM additionally allows researchers to regress the $\boldsymbol{A}$-matrix of factor loadings and the **S**-matrix of factor and residual variances on study features. In this sense, OSMASEM can incorporate moderator effects on the key elements that are used to estimate the overall

reliability—yet not on scale reliabilities as model-derived parameters (Cheung, 2018b). To summarize, OSMASEM for reliability generalization requires researchers to (a) synthesize correlation matrices based on a fixed- or random-effects model (including the checking for positive definiteness), specify, estimate, and evaluate the congeneric factor model or alternative measurement models, (b) decide for a final measurement model, and (c) estimate the reliability coefficient based on the measurement model. These steps are described in more detail in Table 3.

Parameter-Based MASEM

Parameter-based MASEM specifies the structural equation model to the correlation or covariance matrices of the primary studies in the first stage, assuming that the model fits the data of all studies i (Cheung & Cheung, 2016). Ultimately, a vector $\boldsymbol{\rho}_i = \boldsymbol{\rho}(\boldsymbol{\theta}_i)$ of true model parameters (e.g., structural coefficients, factor loadings) results from this stage. Although the correlation matrices of the primary studies may vary, the same model is specified and estimated to all studies. As a consequence, a set of model parameters $\boldsymbol{t}_i = \widehat{\boldsymbol{\theta}}_i$ along with an asymptotic sampling covariance matrix $\widehat{\boldsymbol{V}}_{ti}$ is obtained. In the second stage, these parameters are meta-analyzed using, for instance, multivariate meta-analysis. A random-effects model for parameters of the ith primary study can be specified as:

$$\boldsymbol{t}_i = \boldsymbol{\theta}_R + \boldsymbol{u}_{\theta i} + \boldsymbol{e}_{\theta i} \qquad (9)$$

with $\boldsymbol{\theta}_R$ as the population vector of model parameters, random effects $\boldsymbol{u}_{\theta i}$ with the heterogeneity covariance matrix $\boldsymbol{V}_{\boldsymbol{u}_\theta}$, and the sampling covariance matrix $\boldsymbol{V}_{ti}$ (Cheung, 2015).

In contrast to correlation-based MASEM, parameter-based MASEM synthesizes model parameters across studies, allowing for between-study heterogeneity in these parameters and possible moderator effects. However, the handling of missing data is less efficient in parameter-based MASEM, given that the structural equation model is specified

already in the first step—in other words, primary studies with missing data must be excluded (Jak & Cheung, 2019). Moreover, the selection of estimators that can handle correlation matrices as input data to estimate structural equation models is limited in many software packages. For instance, to our best knowledge, researchers can choose among maximum-likelihood (`ML`), generalized least squares (`GLS`), and the unweighted least squares (`ULS`) estimation the R package lavaan; only the latter does not require multivariate normality.

Besides, the structural equation model has to fit the primary study data in the first place to achieve at least configural invariance of the measurement model across studies. As the model parameters rather than the correlations are the sources of heterogeneity in parameter-based MASEM, the between-study variances of factor loadings can be quantified and considered to be indicators of metric invariance. If indeed heterogeneity in factor loadings exist, metric invariance is not given, and the comparability of reliability coefficients across studies may be compromised (Raykov & Marcoulides, 2013). However, this approach is not without limitations: Assuming metric invariance across samples, given the previously described model constraints in MASEM (i.e., factor variance constrained to 1 and correlation matrices as input data) and the direct relation between factor loadings and the reliability coefficients (see equations [3]-[5]), implies the invariance of composite reliabilities. Nevertheless, this does not necessarily imply that strong or even strict measurement invariance are given, because the intercepts and residuals variances are not constrained.

For the meta-analysis of reliability coefficients, we see the potential of parameter-based MASEM (Table 3): Researchers can specify a suitable factor model to the primary study data and estimate the corresponding, model-based reliability coefficients. These coefficients are then aggregated using univariate (e.g., when only one coefficient is extracted) or multivariate meta-analysis (e.g., when multiple coefficients are extracted, for instance, of several scales or subscales). This direct aggregation of reliability coefficients enables

researchers to quantify between-study heterogeneity as the variance component of random effects (τ^2). Ultimately, this variation can be explained by moderators (Table 2). Moreover, extending the meta-analytic models by additional levels of analysis allows researchers to account for possible, additional hierarchies in the data (e.g., studies nested in countries), resulting in further variance components (Cheung, 2015).

To summarize, parameter-based MASEM for reliability generalization requires researchers to (a) specify, estimate, and evaluate the congeneric factor model or alternative measurement models for each primary study in stage 1 (including the checking for positive definiteness if correlation matrices are available), (b) decide for a final stage-1 model, (c) estimate the reliability coefficients based on the final stage-1 model, (d) meta-analyze these coefficients across studies (through univariate meta-analysis if only one coefficient per study is extracted or multivariate meta-analysis of multiple coefficients of scales and subscales are extracted), and (e) quantify (and possibly explain) between-study heterogeneity of the reliability coefficient. These steps are described in more detail in Table 3.

Some Notes on Missing Correlations and Non-Positive Definite Correlation Matrices

The MASEM approaches presented in this paper are based on the reporting of correlation coefficients in primary studies. These coefficients are either pooled across all studies (TSMASEM and OSMASEM) or used directly to derive parameters of a structural equation model (parameter-based MASEM). In practice, however, primary studies may not report all correlation coefficients needed to test a certain structural equation modeling, possibly due to missing values in the primary study data (e.g., participant drop-outs, study designs with planned missing data) or due to fact that not all variables were assessed. The following correlation matrices extracted from two hypothetical primary studies illustrate this situation:

Study 1—Complete correlation matrix with all 6 variables

$$\begin{pmatrix} 1 & & & & & \\ r_{21} & 1 & & & & \\ r_{31} & r_{32} & 1 & & & \\ r_{41} & r_{42} & r_{43} & 1 & & \\ r_{51} & r_{52} & r_{53} & r_{54} & 1 & \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & 1 \end{pmatrix}$$

Study 2—Incomplete correlation matrix with only 3 variables

$$\begin{pmatrix} 1 & & & & & \\ r_{21} & 1 & & & & \\ r_{31} & r_{32} & 1 & & & \\ NA & NA & NA & NA & & \\ NA & NA & NA & NA & NA & \\ NA & NA & NA & NA & NA & NA \end{pmatrix}$$

While study 1 contributes 15 correlations, study 2 contributes only 3 correlations. While the primary study authors may have reported a 3×3 correlation matrix in their publication, the MASEM model contains six variables and requires a 6×6 correlation matrix. Study 2 may therefore exhibit a non-positive definite correlation matrix due to its missing correlations (Cheung, 2015). Ultimately, this situation results in correlation matrices of different sizes. Ideally, if researchers are interested in testing a factor model to compute the reliability coefficient of a scale comprising p variables, each primary study would provide $p(p-1)/2$ correlations. For instance, for a scale comprising six variables, some primary studies may contribute with correlation matrices comprising six correlations that were based on four variables, although 15 correlations would be needed for a complete matrix. Both the stage-1 TSMASEM and OSMASEM approaches treat correlation matrices as matrices with missing values on the remaining nine correlations. Through maximum-likelihood estimation, complete and incomplete correlation matrices can be pooled to an average correlation matrix (see Cheung, 2015). In the case of parameter-based MASEM, in which the factor model with a latent variable measured by six indicator variables is fit to the correlation matrices of the primary studies, only complete correlation matrices can be considered to obtain the relevant

model parameters (i.e., factor loadings and residuals variances; see Cheung & Cheung, 2016). In other words, the same factor model must be specified and estimated in all primary studies.

The input correlation matrices extracted from the primary studies may be non-positive definite and would therefore neither be submitted to structural equation modeling nor the pooling of correlation matrices (Cheung, 2015). Possible reasons that can lead to such correlation matrices may include, but are not limited to correlation coefficients close to zero, small negative correlations next to positive correlation coefficients in the same matrix (for example, due to rounding or reporting errors), missing correlations in the correlation matrix needed for the pooling stage, the mismatch between reported correlation matrices and model parameters in the primary studies, or the discrepancy between the sample sizes the reported correlation matrices and model parameters are based on. Hypothesizing about the specific, substantive reasons for such occurrences is, however, beyond the scope of this tutorial. At this point, we notice that excluding the non-positive definite correlation matrices represents a current limitation of MASEM as it reduces the meta-analytic sample. To circumvent this reduction, researchers may consider approaches that either do not require correlation matrices to be positive definite (e.g., Bayesian estimation with certain distributional assumptions on priors; Chung et al., 2015) or substitute the non-positive definite matrices by the nearest positive definite matrices (e.g., through the `nearPD()` function in the R package Matrix; Bates et al., 2019). Although the latter seems especially appealing to sustaining the meta-analytic sample, the effects of substituting non-positive definite matrices on the SEM parameters are yet to be examined.

Step-by-Step Tutorial: Overview of the Empirical Examples

To illustrate the usefulness and strengths of these MASEM approaches, we present two examples of how scale reliabilities can be meta-analyzed following the proposed analytic steps (Table 3). The first example is based on a meta-analysis of the correlations among

subscale scores of technology acceptance measures. This meta-analysis contains a large sample of primary studies and independent samples, resulting in 142 correlation matrices. Given that few primary studies included all technology acceptance measures and assessed only a selection of these variable, the resultant correlation matrices vary in their sizes. In other words, for most studies, the overall 6x6-correlation matrix with 15 correlations among the six variables was incomplete (i.e., correlations within this correlation matrix were missing). Besides, some correlation matrices were non-positive definite. As a consequence, only TSMASEM and OSMASEM are available to synthesize the scale reliabilities. A one-factor model without any additional modifications forms the basis for the reliability generalization of ω_T. We further use this example to illustrate how researchers can perform subgroup analyses to examine possible differences in reliability coefficients under invariance constraints. The R code, the corresponding output, and detailed comments are part of the Supplementary Material S1.

The second example is based on a meta-analytic dataset of the Rosenberg Self-Esteem Scale which contains only complete and positive definite correlation matrices. In contrast to the first example, correlations among items rather than subscales are reported, and the overall sample size is small (37 independent samples). Moreover, a nested-factor model forms the basis for estimating the reliability coefficient ω_H. Besides TSMASEM and OSMASEM, parameter-based MASEM can be applied to synthesize the reliability coefficients due to the availability of complete correlation matrices. The R code, the corresponding output, and detailed comments are shown in the Supplementary Material S2.

Modeling the random effects of the extracted correlations or the estimated model parameters, we quantify the between-study variance (τ^2), test the homogeneity of using the Q statistic, and report the heterogeneity statistic I^2. The I^2 statistic is represented as $I^2 = 100\% \times \frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{v}}$ with the within-study sampling variance of $\tilde{v} = \frac{(k-1)\sum_{i=1}^{I} 1/v_i}{\left(\sum_{i=1}^{k} 1/v_i\right)^2 - \sum_{i=1}^{k} 1/v_i^2}$ and

sampling variances v_i for each of the k primary studies (Cheung, 2014; Higgins & Thompson, 2002). For most random-effects models, we obtain the likelihood-based confidence intervals (LBCIs) from the likelihood-ratio statistic (Neale & Miller, 1997). These confidence intervals outperform the Wald CIs, especially for variance components and especially when the number of studies or independent samples is small (Cheung, 2009). To evaluate the fit of the factor models, we refer to common fit indices (Comparative Fit Index [CFI], Root Mean Squared Error of Approximation [RMSEA], Standardized Root Mean Squared Residual [SRMR]) and recommended guidelines (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004).

Empirical Example 1: Technology Acceptance Measures

Data

Scherer, Siddiq, and Tondeur (2019) recently conducted a meta-analysis of pre- and in-service teachers' technology acceptance—a construct measured by six variables: perceived ease of use (PEOU), perceived usefulness (PEU), attitudes toward technology (ATT), subjective norms (SN), technological self-efficacy (TSE), and facilitating conditions (FC). These variables were represented by continuous subscale scores, which were assumed to be approximately normally distributed. Factor analyses suggested that a one-factor model assuming that the latent variable technology acceptance is indicated by these six variables represented the data best (Scherer et al., 2020). The authors extracted 142 correlation matrices from 132 research papers. Overall, the teacher sample in the primary studies contained 37211 teachers (47.7 % pre-service and 52.3 % in-service teachers), with sample sizes between 29 and 1981 teachers (64.7 % female). The technology acceptance measures referred to technology in general (51.6 %) or specific technologies (48.4 %).

Analysis

We first took the TSMASEM approach and synthesized the correlation matrices under the assumptions of fixed or random effects. In the second stage, we specified a one-factor model, evaluated its fit to the data, and extracted the factor loadings and residual variances for scale reliability estimation. Standard errors and confidence intervals were obtained from the likelihood estimation (LBCI). Second, we applied OSMASEM and examined the possible moderation effects on the factor loadings by the type of technology—a variable used as a reference in the technology acceptance measures. Scale reliability was then derived from the model parameters directly. The fact that not all primary studies provided complete 6x6-correlation matrices did not allow us to perform parameter-based MASEM, because fitting the same factor model to primary study data based on correlation matrices of different sizes (i.e., with missing correlations) was not possible.

Some papers reported multiple correlation matrices, and single correlations within these matrices exhibited variation within and between studies. Such dependencies could be handled by a multilevel or robust standard error MASEM approach (e.g., Wilson, Polanin, & Lipsey, 2016)—however, at the time of writing, the performance of these approaches has not yet been evaluated in simulation studies for the different MASEM approaches and for structural equation models beyond meta-regression. Moreover, while these approaches account for the existence of multiple correlations among the same variables in one study, they do not necessarily account for the dependencies among multiple correlations among different variables in one study at the same time (Jak & Cheung, 2019). Researchers are therefore encouraged to monitor the methodological developments concerning this issue.

As the authors of the initial meta-analysis argued that dependencies among correlation matrices due to this structure may indeed exist, we applied some decision rules to select one correlation matrix from the studies reporting multiple correlation matrices (see Supplementary

Material S1). However, in some instances, the authors of the primary studies reported an overall correlation matrix across multiple groups after testing for measurement invariance. In these cases, applying selection rules was not necessary.

The full data set contained correlation matrices, sample sizes, and study names and was stored in the object `TAM1`. We used the R package metaSEM (Cheung, 2018b) for TSMASEM and OSMASEM. Please find the data, the R code, and output in the Supplementary Material S1.

Results

TSMASEM.

Step 1: Checking correlation matrices for positive definiteness. In the first step, we checked one of the key prerequisites of the TSMASEM approach, that is, the positive definiteness of the correlation matrices. The metaSEM package offers the `is.pd()` function to which all correlation matrices (stored in the object `TAM1$data`) are submitted, `is.pd(TAM1$data)`. The output of this function is `TRUE` if the correlation matrix is positive definite and `FALSE` otherwise. Of the 142 correlation matrices, 128 were positive definite and formed the reduced dataset `TAM2`. The excluded correlation matrices contained either a mixture between positive, small negative, close-to-zero correlations or contributed only few correlations (i.e., had many missing values). After applying the decision rules for selecting one correlation matrix per study for studies reporting multiple correlation matrices, the final meta-analytic dataset (`TAM3`) comprised 117 correlation matrices with an overall sample size of $N = 36619$ participants. This dataset was submitted to the first stage of TSMASEM.

Steps 2 and 3: Pooling correlation matrices under fixed- and random-effects models. In the first TSMASEM stage, we pooled the correlation matrices assuming fixed or random effects of the correlations. This step is conveyed through the `tssem1()` function to which

the positive definite correlation matrices (`TAM3$data`) and the sample sizes (`TAM3$n`) are submitted. Defining the object `TSMASEM.rem`, we estimated all parameters of the stage-1 random-effects model (`method="REM"`) as follows:

```
TSMASEM.rem <- tssem1(TAM3$data, TAM3$n, method="REM",
                      RE.type="Diag", I2="I2q")
summary(TSMASEM.rem)
```

This object contains both the pooled correlations (and, ultimately, the pooled correlation matrix) as fixed effects and the between-study variances (τ^2) as random effects. Besides, it provides the heterogeneity indices I^2 for each correlation based on the Q statistics (`I2="I2q"`). As the data set contained six variables and 15 correlations off the diagonal, a total of 120 entries in the covariance matrix of the random effects (i.e., 15 variances and 105 covariances) would have to be estimated from the 117 correlation matrices. To reduce the number of random-effects parameters (i.e., the number of entries in the 15×15-matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$), the option `RE.type="Diag"` specifies the covariance matrix of random effects as a diagonal matrix, and random effects are considered independent. For a larger sample of primary studies, researchers may choose the alternative option `RE.type="Symm"` to estimate the full, symmetric covariance matrix. We notice that the diagonal constraint we imposed on the stage-1 model may not reflect the true nature of the random-effects covariance structure, because covariances among random effects are likely to occur (Cheung & Cheung, 2016). Constraining these covariances may, however, impact the standard errors of the pooled correlations. For this illustrative example, however, estimating the full random-effects covariance matrix did not converge due to the limited sample size. In general, researchers have to weigh and be aware of the practical limitations associated with small meta-analytic samples and the assumptions made in the stage-1 model (Cheung, 2015).

Under the random-effects model, considerable heterogeneity was indicated by high heterogeneity indices between I^2 = 82.5 % and 92.7 % and between-study variance estimates ranging from τ^2 = 0.015 to 0.035 with LBCIs that did not contain zero and (see Supplementary Material S1). At this point, reasons for this heterogeneity may be manifold and include aspects of the methodological diversity of the primary studies (e.g., differences in assessment administration, scale composition, statistical modeling). The overall homogeneity test supported the existence of random effects, $Q(617) = 6582.6$, $p < .001$. The resultant correlation matrix is shown in Table 4.

To further substantiate the appropriateness of the random-effects model in the stage-1 TSMASEM analyses, we specified and estimated a fixed-effects model (`TSMASEM.fem`):

```
TSMASEM.fem <- tssem1FEM(TAM3$data, TAM3$n)
summary(TSMASEM.fem)
```

This model did not fit the data well ($\chi^2[617] = 10740.5$, $p < .001$, RMSEA = 0.229, SRMR = 0.167, CFI = 0.745, AIC = 9506, BIC = 4257), suggesting that between-study variation in the correlation matrices may exist. The resultant pooled correlation matrix is also shown in Table 4. Despite the poor fit of the fixed-effects model, these pooled correlations correlated highly with those obtained from the random-effects model, $r = .977$ (see Figure 2). Nevertheless, the pooling of the correlation matrices with random effects seemed more appropriate.

Step 4: Specifying, estimating, and evaluating the congeneric model. In the next stage, we specified the congeneric factor model (`OneFactorModel`) on the basis of the stage-1 random-effects model. The latent variable was identified by fixing the factor variance to 1. Given this and the fact that a correlation matrix with variances of the manifest indicators constrained to 1 was the basis for the structural equation modeling step, factor loadings and residual variances were dependent with $\theta_{jj} = 1 - \lambda_j^2$ for each manifest indicator j (Jak & Cheung, 2018b). We first specified the model using the lavaan language and secondly

transformed it into the RAM-formulation (McArdle, 2005)—the `lavaan2RAM()` function performs this transformation. For the congeneric model, this formulation contains the matrix of factor loadings ***A***, the matrix of the factor variance and residual variances ***S***, and the ***F***-matrix indicating whether or not a variable is latent (coded as 0) or manifest (coded as 1).

```
OneFactorModel <- " # General factor (gTA)
                    # Factor loadings labeled L1-l6
                    gTA =~ L1*PEOU + L2*PU + L3*ATT +
                    L4*SN + L5*FC + L6*TSE
                    # Residual variances labeled R1-R6
                    PEOU ~~ R1*PEOU
                    PU ~~ R2*PU
                    ATT ~~ R3*ATT
                    SN ~~ R4*SN
                    FC ~~ R5*FC
                    TSE ~~ R6*TSE
                    # Factor variance constrained to 1
                    gTA ~~ 1*gTA "
RAM <- lavaan2RAM(OneFactorModel, obs.variables = c("PU",
                    "PEOU", "ATT", "SN", "TSE", "FC"))
# Matrices in the RAM framework
A <- RAM$A
S <- RAM$S
F <- RAM$F
```

Along with the pooled correlation matrix (`TSMASEM.rem`), these model specification matrices (***A***, ***S***, and ***F***) are then submitted to the `tssem2()` function:

```
TSMASEM.cfa1 <- tssem2(TSMASEM.rem, Amatrix=A, Smatrix=S,
```

```
                    Fmatrix=F,
                    intervals.type="LB",
                    diag.constraints=TRUE,
                    model.name="One factor model REM",
                    mx.algebras=list(SREL=
                    mxAlgebra(((L1+L2+L3+L4+L5+L6)^2)/
                    ((L1+L2+L3+L4+L5+L6)^2+
                    R1+R2+R3+R4+R5+R6), name="SREL")))
TSMASEM.cfa1 <- rerun(TSMASEM.cfa1)
summary(TSMASEM.cfa1)
```

In this function, we requested LBCIs (`intervals.type="LB"`) and ensured that the diagonal elements of the model-implied correlation matrix are all 1 (`diag.constraints=TRUE`). Researchers may also request the Wald CIs and relax the diagonal constraints as there are no mediators in the measurement model (see Cheung, 2015, for more detailed explanations). To circumvent estimation errors, this model was rerun until a parameter solution had been found (`rerun()`). The object `TSMASEM.cfa1` contains all relevant parameters to evaluate the fit of the congeneric model and the overall scale reliability. The stage-2 factor model based on the stage-1 random-effects model resulted in a good fit, $\chi^2(9) = 29.6$, $p < .001$, RMSEA = 0.008, SRMR = 0.038, CFI = 0.994, AIC = 12, BIC = -65. Figure 3 shows the corresponding factor loadings and residual variances. Already in this step, we estimated the reliability coefficient ω_T (`SREL`) from the model parameters. Notice that, although the model estimation was based on the stage-1 random-effects model, stage-2 random effects of model parameters are not estimated. As a consequence, this approach assumes that factor loadings are invariant across the primary studies (i.e., metric invariance holds).

Step 5: Specifying, estimating, and evaluating alternative measurement models. Instead of synthesizing the reliability coefficient ω_T, researchers may be interested in synthesizing Cronbach's α. To test whether the τ-equivalence assumption was met and to obtain this reliability coefficient from the measurement model, we constrained the factor loadings (`L1-L6`) to be equal for all manifest indicators (`L1`). Using the same functions as presented in step 4, we based the model estimation on the outcomes of the stage-1 random-effects TSMASEM. The model assuming the τ-equivalence (`TSMASEM.cfa1.eq`) showed a reasonable fit to the data, $\chi^2(14) = 228.9$, $p < .001$, RMSEA = 0.021, SRMR = 0.101, CFI = 0.936, AIC = 201, BIC = 82.

Step 6: Deciding on a final measurement model. Comparing the models with and without equality constraints on the factor loadings by means of chi-square difference testing (in R: `anova(TSMASEM.cfa1,TSMASEM.cfa1.eq)`) suggested that the model with freely estimated factor loadings showed a significantly better fit to the data, $\Delta\chi^2(5) = 199.4$, $p < .001$. Along similar lines, the information criteria were smaller for this model. These findings indicated that the reliability coefficient ω_T is a more suitable coefficient for reliability generalization. We accepted the one-factor model without any equality constraints of factor loadings and residual variances as the final measurement model.

The estimation of the measurement model was based on the pooled correlation matrix and its asymptotic covariance matrix; random effects of the model parameters (i.e., factor loadings and residual variances) are not included. Recently, some attempts were made to quantify the heterogeneity of these parameters by parametric bootstrapping and the delta method (Cheung, 2018a; Yu, Downes, Carter, & O'Boyle, 2018). In the R package metaSEM, researchers can ultilize the `tssemParaVar()` function to perform both methods on the TSMASEM stage-2 parameters (see Supplementary Material S1).

Step 7: Estimating the overall scale reliability. The measurement model based on the stage-1 random-effects model resulted in an overall reliability coefficient of ω_T = 0.790, 95 % LBCI [0.779, 0.800].

Step 8: Subgroup analyses. The data set contained additional information about the teacher sample (i.e., pre- vs. in-service teacher samples) and the type of technology referred to in the technology acceptance measures (i.e., specific technology vs. technology in general). Using these two grouping variables, we performed subgroup analyses to examine whether the group-specific reliability coefficients may differ between these groups. We present these analyses focusing on the type of teacher sample here, while the analyses for the type of technology are described in the Supplementary Material S1.

First, we subset the data into studies of in- and pre-service teachers. Second, we performed stage-1 TSMASEM with random effects for each of these subgroups:

```
# In-Service teachers
stage1_ins <- tssem1(TAM3_ins, n_ins, method="REM",
RE.type="Diag")
# Pre-Service teachers
stage1_pre <- tssem1(TAM3_pre, n_pre, method="REM",
RE.type="Diag")
```

Next, we specified and estimated the congeneric model (see step 4) for each of these subsamples and found that this model exhibited good fit to the samples of in-service teachers (χ^2[9, N = 17987] = 24.8, p = .003, RMSEA = 0.010, SRMR = 0.051, CFI = 0.991, AIC = 7, BIC = -63) and pre-service teachers (χ^2[9, N = 18632] = 16.1, p = .067, RMSEA = 0.007, SRMR = 0.038, CFI = 0.996, AIC = -2, BIC = -72). The resultant reliability coefficients were ω_T = 0.799, 95 % LBCI [0.784, 0.812], and, respectively, ω_T = 0.778, 95 % LBCI [0.763, 0.792]. In order to ensure that these coefficients are in fact comparable, the invariance of

factor loadings across the two subsamples must be given. Hence, we specified a two-group model with these invariance constraints and compared it to a model with freely estimated factor loadings. The details of this specification, the estimation, and comparison are part of the Supplementary Material S1. The metric invariance model showed a good fit to the data, $\chi^2(24) = 55.9$, $p < .001$, RMSEA = 0.009, SRMR = 0.052, CFI = 0.991, AIC = 92, BIC = 245. Nevertheless, the chi-square difference test indicated that model fit differed significantly between the two models, $\Delta\chi^2(6) = 15.1$, $p = .020$. These finding suggested that the invariance of factor loadings across the two subsamples may not hold. Researchers should consequently be cautious with comparing the two reliability coefficients.

OSMASEM.

Step 1: Checking correlation matrices for positive definiteness. We already performed the test for positive definite correlation matrices under the TSMASEM approach and excluded non-positive definite matrices. The resultant data were stored in the object `TAM3`.

Step 2: Pooling correlation matrices and specifying, estimating, and evaluating the congeneric factor model. We pooled the correlation matrices under a random-effects model and estimated the congeneric model using the model specification matrices (***A***, ***S***, and ***F***; see TSMASEM step 4). The corresponding R code is as follows:

```
# Combine the data
TAM3 <- Cor2DataFrame(TAM3$data, TAM3$n)
# Create the M-matrix with the relevant model specification
# The matrices Ax and Sx are empty as no moderators are
included
M0 <- create.vechsR(A0=A, S0=S, F0=F, Ax=NULL, Sx=NULL)
# Create heterogeneity variances (random-effects model)
T0 <- create.Tau2(RAM=RAM, RE.type="Diag")
```

```
# Define the reliability coefficient as a new parameter SREL
SREL <- mxAlgebra(((L1+L2+L3+L4+L5+L6)^2)/
                ((L1+L2+L3+L4+L5+L6)^2+((1-L1^2)+(1-L2^2)+(1-
                L3^2)+(1-L4^2)+(1-L5^2)+(1-L6^2))),
          name="SREL")
# Model estimation
OneFactorModel.os.fit <- osmasem(model.name="One factor CFA",
                      Mmatrix=M0,
                      Tmatrix=T0,
                      mxModel.Args=list(SREL, mxCI(c("SREL"))),
                      data = TAM3)
```

After combining the data to a new data frame (`TAM3`) for the `osmasem()` function, the factor model is specified (`M0`) using the specification matrices $\boldsymbol{A}$, $\boldsymbol{S}$, and $\boldsymbol{F}$, yet no matrix with moderator variables (i.e., the moderator matrices $\boldsymbol{Ax}$ and $\boldsymbol{Sx}$ are empty). Next, the matrix of random effects (`T0`) is defined with diagonal constraints comparable to the TSMASEM approach. Finally, these elements, along with the defined reliability coefficient `SREL`, are submitted to the `osmasem()` function in order to perform the pooling of the correlation matrices and the estimation of the congeneric model in one step. Simultaneously, the reliability coefficient and its Wald 95 % CI are estimated. In case researchers want to perform OSMASEM assuming fixed effects, the $\boldsymbol{T}$-matrix command can be modified by `RE.type="Zero"`. Comparing the one-factor congeneric models based on fixed and random effects indicated the preference for the model with random effects of the correlations, $\Delta\chi^2(15) = 7145.9$, $p < .001$. Please find the detailed R code and output in the Supplementary Material S1.

Steps 3 and 4: Pooling correlation matrices, estimating alternative measurement models, and comparing models. Similar to TSMASEM, we specified and estimated the one-factor model with equal factor loadings by adjusting the above-described syntax (see Supplementary Material S1). Once again, the comparison between the congeneric model and the model with equality constraints suggested the preference of the former, $\Delta\chi^2(5) = 145.7$, $p < .001$. The congeneric model may serve as the final measurement model.

Step 5: Estimating the overall scale reliability. The overall scale reliability (`SREL`) based on the random-effects model was $\omega_T = 0.790$, 95 % CI [0.778, 0.801].

Step 6: Evaluating moderator effects. To further substantiate the findings obtained from the TSMASEM subgroup analyses, we introduced the type of teacher sample (variable `InService`) as a possible moderator of the ***A***-matrix (i.e., the matrix of factor loadings). To achieve this, a moderator matrix ***Ax*** is defined and submitted to the model specification:

```
# Create an A-moderation matrix
Ax <- matrix(c(0,0,0,0,0,0,"0*data.InService",
               0,0,0,0,0,0,"0*data.InService ",
               0,0,0,0,0,0,"0*data.InService ",
               0,0,0,0,0,0,"0*data.InService ",
               0,0,0,0,0,0,"0*data.InService ",
               0,0,0,0,0,0,"0*data.InService ",
               0,0,0,0,0,0,0),
             nrow = 7, ncol = 7, byrow = TRUE)
# Create the M-matrix with the relevant model specification
M1 <- create.vechsR(A0=A, S0=S, F0=F, Ax=Ax, Sx=NULL)
```

The resultant variance explanations by the teacher sample, indicated by the reduction of between-study variances of the correlations, ranged from 0.0 % to 23.2 % across the six

variables, and only one moderator effect on the factor loading of the variable FC was significant, $B = 0.109$, $SE = 0.041$, $p = 0.012$. Comparing the models with and without moderator effects indicated an improvement of model fit after introducing the moderator, $\Delta\chi^2(6) = 13.35$, $p = .038$. Hence, in line with the results of the TSMASEM subgroup analyses, the measurement invariance across the two teacher samples may be compromised.

Summary

Performing TSMASEM and OSMASEM, we found that a one-factor model, which was based on random effects in the correlation matrices, showed a good fit to the data. This model formed the basis for synthesizing ω_T as a reliability coefficient across studies. Moderation effects on the factor loadings by the types of teacher sample were evident.

Empirical Example 2: Rosenberg Self-Esteem Scale

Data

Gnambs, Scharl, and Schroeders (2018a) examined the factor structure of the 10-item Rosenberg Self-Esteem Scale (RSES) meta-analytically and found support for a nested-factor structure. The authors extracted the data from 34 studies and 113 independent samples (N = 140671) and performed fixed-effects TSMASEM. The sample sizes ranged between 59 and 22131 participants of the primary studies (55 % female) and were obtained from countries across the world, including the United States of America (18 %), The Netherlands (8 %), and Germany (6 %). Gnambs et al. (2018a) tested the robustness of their findings for a subsample of primary studies that comprised the complete correlation matrices obtained from 37 independent samples. These samples were nested in 15 studies. We use this subsample of the study to illustrate the meta-analysis of reliability coefficients performing TSMASEM, OSMASEM, and parameter-based MASEM (see Gnambs et al., 2018b for the data set).

Analysis

Similar to example 1, we obtain an overall estimate of the scale reliability, performing TSMASEM and OSMASEM under an appropriate factor model of the data. In contrast to the first example, this factor model represents a bifactor model with uncorrelated factors (e.g., A. Rodriguez, Reise, & Haviland, 2016), hereafter referred to as a "nested-factor model". Given that all correlation matrices are complete, we also perform parameter-based MASEM to estimate the reliability heterogeneity (Cheung & Jak, 2016). We use the R package metaSEM (Cheung, 2018b) for TSMASEM and OSMASEM, lavaan (Rosseel, 2018) and metafor (Viechtbauer, 2010) for parameter-based MASEM.

Two issues with analyzing the data are worth noting: First, Gnambs et al. (2018a) either extracted correlations among ordinal responses on the RSES items or computed the model-implied correlations from the parameters of factor models. The authors ensured that each primary study administered the RSES with at least four response categories per item, so that item responses may be treated continuously, although more categories would have been desirable to back this treatment (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Correlations were quantified as Pearson's *r* and pooled across studies under the multivariate normality assumption —an assumption that may not hold for ordinal item responses (Li, 2016).

Second, correlation matrices were nested in studies—this design feature ultimately created a hierarchical data structure. After inspecting the primary study data, Gnambs et al. (2018a) concluded that the 37 samples (and the corresponding correlation matrices) can be considered independent, and the RSES administration was the common element (see also Jak & Cheung, 2019). For the purpose of this illustration, we will also assume independence among study samples and estimate reliabilities as a sample- rather than study-based coefficient. Although the independence assumption may in fact not hold, we have based the TSMASEM and OSMASEM approaches on it, due to the current lack of validated multilevel

MASEM approaches (Jak & Cheung, 2019). Nevertheless, we performed three-level univariate random-effects modeling in the second stage of the parameter-based MASEM approach.

Given that the implementation of TSMASEM and OSMASEM were already presented in detail for example 1, we focus only on the modifications for this example and kindly refer readers to the Supplementary Material S2 for the detailed R code, output, and explanations.

Results

TSMASEM.

Step 1: Checking correlation matrices for positive definiteness. Using the `is.pd()` function to the data (`RSES1`), we found that the 37 correlation matrices were positive definite and could thus be included in the subsequent pooling stage.

Steps 2 and 3: Pooling correlation matrices under fixed- and random-effects models. In the TSMASEM stage-1 analyses, the fixed-effects model (`TSMASEM.fem`) represented the data to a sufficient degree, $\chi^2(1620) = 29892.0$, $p < .001$, RMSEA = 0.076, SRMR = 0.113, CFI = 0.937, AIC = 26652, BIC = 11074. However, the random-effects model (`TSMASEM.rem`) indicated between-study heterogeneity ($Q[1620] = 31677.6$, $p < .001$), the between-study variances of the 45 item-item correlations ranged between $\tau^2 = 0.004$ and 0.028 with confidence intervals that did not include zero. The heterogeneity indices were considerably large (I^2 = 92.9 %–99.1 %), despite the fact that all primary studies used the same set of items to assess self-esteem. Possible reasons for this heterogeneity may be related to differences in sample and study characteristics (Gnambs et al., 2018a). Hence, the stage-1 random-effects model seemed more appropriate for the pooling of correlation matrices. The pooled correlations for the fixed- and random-effects models were highly correlated, $r = .989$ (see Supplementary Material S3). Once again, the diagonal constraints of the covariance

matrix of random effects were implemented due its large parameter space (i.e., symmetric 45×45-matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{u}}$).

Step 4: Specifying, estimating, and evaluating the congeneric model. In the next step, we specified and estimated the congeneric one-factor model, transferring the R code from example 1 to the RSES data. If indeed this model represents the data well, the reliability coefficient ω_T could be reported. This model showed a reasonable yet not good fit to the data, $\chi^2(35) = 1208.3$, $p < .001$, RMSEA = 0.017, SRMR = 0.080, CFI = 0.931, AIC = 1138, BIC = 802. Reporting ω_T may therefore not be appropriate.

Step 5: Specifying, estimating, and evaluating alternative measurement models. As Gnambs et al. (2018a) and Jak and Cheung (2019) noticed, the factor structure of the RSES may not be best represented by a one-factor model—instead, the authors showed that nested-factor models were superior. In the current subsample, the model with a general factor of self-esteem and two uncorrelated specific factors that explained variance in the negatively- and positively-worded RSES items (see Figure 4) may represent the factorial structure significantly better than the one-factor model. We consequently specified and estimated this model as an alternative measurement model (`NestedFactorModel`).

```
NestedFactorModel <- "# General self-esteem factor (gRSES)
                      # Factor loadings labeled Lg1-Lg10
                      gRSES =~ Lg1*Item1 + Lg2*Item2 + Lg3*Item3 +
                           Lg4*Item4 + Lg5*Item5 + Lg6*Item6 +
                           Lg7*Item7 + Lg8*Item8 + Lg9*Item9 +
                           Lg10*Item10
                      # Specific factor s1 (positive wording)
                      s1 =~ Ls1*Item1 + Ls3*Item3 + Ls4*Item4 +
                           Ls7*Item7 + Ls10*Item10
                      # Specific factor s2 (negative wording)
```

```
s2 =~ Ls2*Item2 + Ls5*Item5 + Ls6*Item6 +
      Ls8*Item8 + Ls9*Item9
# Restriction on the factor covariances
gRSES ~~ 0*s1
gRSES ~~ 0*s2
s1 ~~ 0*s2
# Residual variances labeled R1-R10
Item1 ~~ R1*Item1
Item2 ~~ R2*Item2
Item3 ~~ R3*Item3
Item4 ~~ R4*Item4
Item5 ~~ R5*Item5
Item6 ~~ R6*Item6
Item7 ~~ R7*Item7
Item8 ~~ R8*Item8
Item9 ~~ R9*Item9
Item10 ~~ R10*Item10
# Factor variances constrained to 1
gRSES ~~ 1*gRSES
s1 ~~ 1*s1
s2 ~~ 1*s2 "
```

Similar to example 1, this model specification was then translated into the RAM-formulation (with model specification matrices **A**, **S**, and **F**) and subsequently submitted to the `TSSEM2()` function. ω_H was the corresponding coefficient of scale reliability (see equation [3]; `SREL`). The R code for the stage-2 TSMASEM reads:

```
TSMASEM.nfm <- tssem2(TSMASEM.rem, Amatrix=A, Smatrix=S,
                      Fmatrix=F, intervals.type="LB",
```

```
                    diag.constraints = TRUE,
                    model.name = "Nested factor model",
                    mx.algebras=list(SREL=mxAlgebra(((Lg1+Lg2+
                    Lg3+Lg4+Lg5+Lg6+Lg7+Lg8+Lg9+Lg10)^2)/((Lg1
                    +Lg2+Lg3+Lg4+Lg5+Lg6+Lg7+Lg8+Lg9+Lg10)^2+
                    (Ls1+Ls3+Ls4+Ls7+Ls10)^2+
                    (Ls2+Ls5+Ls6+Ls8+Ls9)^2+
                    R1+R2+R3+R4+R5+R6+R7+R8+R9+R10),
                name="SREL")))
```

This model exhibited a very good fit to the data, $\chi^2(25) = 44.9$, $p = .009$, RMSEA = 0.003, SRMR = 0.017, CFI = 0.998, AIC = -5, BIC = -245. The factor loadings of the general factor (`Lg1-Lg10`) were positive and significant, the factor loadings of the specific factors (`Ls1-Ls10`) ranged between negative, close-to-zero, and positive values (Table 5)—an anomaly often encountered in bifactor models (Eid, Geiser, Koch, & Heene, 2017).

Step 6: Deciding on a final measurement model. Comparing the one-factor model with the nested-factor model suggested the preference of the latter over the former, $\Delta\chi^2(10) = 1163.3$, $p < .001$. We consequently based the reliability estimation on the nested-factor model.

Step 7: Estimating the overall scale reliability. The overall reliability coefficient across the 37 samples was $\omega_H = 0.745$, 95 % LBCI [0.735, 0.755].

OSMASEM.

Step 1: Checking correlation matrices for positive definiteness. We already performed this step for the TSMASEM approach; all correlation matrices were positive definite.

Step 2: Pooling correlation matrices and specifying, estimating, and evaluating the congeneric factor model. We pooled the correlation matrices under a random-effects model

and estimated the congeneric model using the model specification matrices. Supplementary Material S2 shows the corresponding model parameters and their confidence intervals.

Steps 3 and 4: Pooling correlation matrices, estimating alternative measurement models, and comparing these models with the congeneric model. Similar to TSMASEM, we compared the one-factor model with the nested-factor model and found support for the latter, $\Delta\chi^2(10) = 639.9$, $p < .001$.

Step 5: Estimating the overall scale reliability. The overall scale reliability based on the nested-factor model with random effects was $\omega_H = 0.746$, 95 % CI [0.735, 0.756].

Step 6: Evaluating moderator effects. In a final step, we examined possible moderation effects on the $\boldsymbol{A}$-matrix by the language of test administration (coded as *1=English*, *0=Language other than English*). Following the same procedure as presented in example 1, we specified the moderation matrix $\boldsymbol{Ax}$, estimated the model parameters, and compared the model to the one without the moderator. The latter comparison suggested that moderation effects on the factor loadings existed, $\Delta\chi^2(30) = 101.3$, $p < .001$. Hence, researchers should be cautious with comparing the reliability coefficient across languages of test administration due to measurement non-invariance. Please find all model parameters in the Supplementary Material S2.

Parameter-based MASEM.

Step 1: Checking correlation matrices for positive definiteness. We already performed this step for the TSMASEM approach; all correlation matrices were positive definite.

Step 2: Specifying, estimating, and evaluating the congeneric factor model for each primary study. We performed this step using model specification of the one-factor model (see TSMASEM and OSMASEM) and the `sem()` function in the R package lavaan. We estimated the congeneric model for each primary study and extracted the resultant model fit

indices (see Supplementary Material S2 and S3). The model showed poor fit to the data in all primary studies and was thus rejected.

Step 3: Specifying, estimating, and evaluating alternative factor models for each primary study. The nested-factor model served as an alternative measurement model and, indeed, represented the data well for 34 of the 37 correlation matrices—three correlation matrices had to be excluded due to non-convergence of the model (see Supplementary Material S2 and S3).

Step 4: Deciding on a final measurement model. The nested-factor model formed the basis for the subsequent reliability estimation as it fit well to the primary studies (i.e., configural invariance across studies held).

Step 5: Estimating the scale reliabilities for each primary study sample. As part of the model estimation in lavaan, the scale reliabilities were estimated for each study sample. Overall, the omega coefficients ranged between $\omega_H = 0.411$ and $\omega_H = 0.861$ with a median of 0.767 and a mean of 0.735 (*SD* = 0.108). The full list of reliabilities is contained in the Supplementary Material S3.

Steps 6-8: Performing univariate meta-analysis under fixed- and random-effects models. In the next step, we submitted the reliability coefficients and the sample sizes to three-level univariate random-effects meta-analyses using the `rma.mv()` function in metafor. This function can treat the reliability coefficients as correlations or as internal consistencies comparable to Cronbach's α and accounts for the hierarchical data structure (with samples [`sampleID`] nested in studies [`studyID`])—we modeled the nesting of reliabilities in study papers next to samples explicitly (level 1: model parameters, level 2: study samples, level 3: study papers; see Supplementary Material S2). Treating reliabilities as correlations, the three-level univariate random-effects model (`REM.opt1.ml`) is specified as

follows (with correlations and sampling variances obtained from the `escalc()` function and stored in the object `opt1`; see Supplementary Material S2):

```
# Three-level random-effects model using REML
REM.opt1.ml <- rma.mv(yi, vi, data=opt1,
                      random = list( ~ 1 | sampleID,
                                     ~ 1 | studyID))
summary(REM.opt1.ml)
```

The random-effects model indicated heterogeneity across samples, yet not across studies (Q[33] = 4633.5, $p < .001$; I_2^2 = 99.5 %, I_3^2 = 0.0 %) with variance estimates of τ_2^2 = 0.010 (level 2) and τ_3^2 = 0.000 (level 3). Although the between-study variance component may not be necessary, the three-level model may capture the hierarchical nature of the data better than the (common) two-level model—the boundary estimate of zero may be due to the confounding of the sample-level and study-level variance components in the model (see also Snijders & Bosker, 2012). The overall reliability coefficient under this model was ω_H = 0.738, 95 % CI [0.703, 0.773]. Notice that the confidence interval of this point estimate is wider than ones observed in TSMASEM and OSMASEM, primarily because the second stage of parameter-based MASEM uses the number of effect sizes (n = 34) rather than the sum of all primary sample sizes (n = 110843; as in TSMASEM and OSMASEM) as the sample size in the meta-analytic model. Moreover, notice that the estimation of the variance components was based on REML estimation; meta-analysts may however use alternative estimation procedures, such as maximum-likelihood estimation in this step (`method="ML"`).

Following the same procedure, the univariate meta-analysis can also be based on transformations of the reliability coefficients. The `rma.mv()` function readily implements these transformations (options: `measure="ARAW"`, `"AHW"`, or `"ABT"` for the raw coefficients, the Hakstian-Whalen, or the Bonett transformations). The resultant scale

reliabilities and variance components are shown in Table 6. Overall, the reliability coefficients ranged between ω_H = 0.738 and 0.754 across the different modeling approaches.

Multivariate meta-analysis of factor loadings. To examine whether comparisons of scale reliabilities across studies are valid, we performed a multivariate random-effects meta-analysis of the factor loadings. As noted earlier (step 4), the configural invariance of the nested-factor model could be assumed across the 34 study samples. The fact that this model fits in the primary studies does, however, not ensure that the factor loadings are invariant, that is, metric invariance holds. To test the invariance of factor loadings across studies, we performed multivariate meta-analysis of the factor loadings using the `meta()` function in the metaSEM package (see Supplementary Material S2). The corresponding random-effects model contained the between-study variances of the 20 factor loadings in the nested-factor model. An inspection of these variances, the I^2 statistic, and the overall homogeneity test (Q[660] = 20133.7, p < .001) suggested that factor loadings varied between studies and were therefore not invariant. Although researchers may be cautious with generalizing reliability coefficients across studies in this situation, the alternative assumption of fixed effects in factor loadings (i.e., metric invariance) may not be realistic, and, in this case, also implies the invariance of reliability coefficients given the model specification (Cheung, 2015). Another alternative may be represented in the assumption of partial metric invariance across studies.

This multivariate model results in pooled factor loadings and, as a by-product, researchers can compute the overall reliability coefficient from them. Applying equation (5), the multivariate random-effects model results in $\omega_H = 0.755$. At the time of writing, however, a solution to quantifying the between-study variance of this estimate, which was derived from model parameters that may have different between-study variances and covariances, was not yet available.

Summary

Overall, the approaches to meta-analyzing the RSES reliabilities (i.e., TSMASEM, OSMASEM, parameter-based MASEM) differed only marginally in the overall reliability estimates. In contrast to example 1, a more complex measurement model formed the basis for the reliability estimation (i.e., a nested-factor model). Unlike TSMASEM and OSMASEM, the parameter-based MASEM approach provided insights into the between-study variation of model parameters, including factor loadings and reliability coefficients.

Take-Home Messages from the Empirical Examples

The empirical examples illustrated both the strengths and the limitations of meta-analytic structural equation modeling for synthesizing scale reliabilities. As MASEM relies on the data based on either the correlations among items or subscales or model parameters instead of reported reliability coefficients, researchers can estimate the same type of reliability coefficient across studies. In order to achieve this, an appropriate measurement model is needed that forms the basis for the reliability estimation. The second example has shown that the congeneric one-factor model for estimating ω_T may not necessarily fit the data well; yet, alternative models with correlated residuals or nested factors may capture the factor structure of a scale better. These deviations from the congeneric model, however, require adjusting the selection and estimation of the reliability coefficient (e.g., ω_H in the first example, ω_T in the second example). Another strength of TSMASEM and OSMASEM is that not all primary studies need to provide full correlation matrices with all correlations among the variables of interest (see Example 1). Although heterogeneity in the correlations can be quantified in stage-1 TSMASEM and OSMASEM, random effects in the estimated reliability coefficients still have to be conceptualized for these approaches. Nevertheless, parameter-based MASEM can address this issue and even account for hierarchical data structures (see Example 2). Overall, the two empirical examples emphasize the following key elements of meta-analyzing

scale reliabilities using MASEM: (a) Specifying an appropriate measurement model, (b) selecting an appropriate reliability coefficient given the measurement model, and (c) obtaining an overall reliability coefficient and, if possible, between-study heterogeneity.

General Discussion

Next to reviewing the current state-of-the-art in the area of reliability generalization, we were aimed at showcasing how MASEM can be utilized to meta-analyze scale reliabilities across studies. In this paper, we presented both correlation-based MASEM and parameter-based MASEM as two vehicles to provide an overall reliability estimate. We illustrated this potential using two examples and argued for the usefulness of *correlation-based MASEM* for reliability generalization in the presence of missing data and deviations from the congeneric factor model. We argued for the usefulness of *parameter-based MASEM* for synthesizing scale reliabilities, quantifying and explaining heterogeneity between studies, and accommodating hierarchical data structures. Finally, we proposed a sequence of steps researchers can take in order to utilize MASEM for reliability generalization.

Implications for the Meta-Analysis of Reliability Coefficients

Overall, we consider MASEM to be useful for the meta-analysis of reliabilities in many ways: First, MASEM allows researchers to specify measurement models that represent the factor structure of the scale and, on the basis of these models, select an appropriate reliability coefficient. Hence, researchers do no longer need to fully rely on the coefficients the authors of the primary studies chose—most commonly, Cronbach's α coefficients are reported (McNeish, 2018). This flexibility opens a broad range of alternative coefficients that do not rely on the assumptions of Cronbach's α (McDonald, 1999). In contrast to the current practices of reliability generalization (see Holland, 2015 and our review), MASEM synthesizes model-based reliabilities, even for complex models such as the bifactor model (Rodriguez et al., 2016). Nevertheless, whether MASEM can in fact be applied to obtain

reliability coefficients from a factor model depends on the availability of item- or subscale-level data (Carpenter et al., 2016)—extracting correlation or covariance matrices is critical to the application of correlation-based MASEM, parameter-based MASEM could rely on the reported model parameters (i.e., factor loadings, residual and factor variances).

Second, MASEM takes into account the multivariate nature of the correlations and covariances underlying the estimation of a factor model and a reliability coefficient (Cheung & Chan, 2005). In their recent paper, Tang and Cheung (2016) illustrated the enormous discrepancies in model parameters and fit statistics between the TSMASEM approach and the univariate-*r* approach which pools correlations in separate meta-analyses. Considering this example and the evidence for the better performance of MASEM through, for instance, TSMASEM (Cheung, 2015), we argue for the multivariate MASEM approaches as opposed to the univariate-*r* approach. Similarly, we argue that multivariate approaches should be considered especially when multiple reliability coefficients are directly extracted from the primary studies (Cheung & Cheung, 2016).

When is the use of MASEM useful for the meta-analysis of scale reliabilities? In this paper, we argued for the potential of MASEM in the following situations: (a) when researchers want to test the factor structure of a scale and, ultimately, select a model-based reliability coefficient, (b) when the assumptions of Cronbach's α are not met and alternative coefficients are more appropriate (based on some evidence from primary studies on deviations from the essential τ-equivalent model), (c) when the authors of primary studies report different types of reliabilities, (d) when multiple reliability coefficients are extracted from multidimensional scales or a set of scales, (e) when sufficient information about the item- or subscale-level correlations are provided.

Finally, researchers may ask 'Which of the MASEM approaches should be used?' The answer to this question depends on the researchers' goals and questions on the one hand, and

the features of the meta-analytic data on the other hand. We would like to highlight several strengths of parameter-based MASEM: First, if correlation matrices are extracted, the factor models are fit to the data of all primary studies, and researchers can ensure that the same model represents the data across studies (i.e., configural invariance; Cheung & Cheung, 2016). Second, factor loadings may be extracted from the factor models in the first stage and synthesized through multivariate meta-analysis in the second stage. The resultant matrix of random effects could provide further insights into the measurement invariance across studies (i.e., metric invariance). Third, when model-based reliability coefficients are synthesized in stage 2, parameter-based MASEM can also provide between-study variances as indicators of heterogeneity. This heterogeneity may then be explained by moderating variables. Fourth, the stage-2 meta-analysis allows researchers to consider the multilevel structure of the data (e.g., study samples nested in studies or countries). Fifth, the stage-1 model estimation can be based on estimators that circumvent the multivariate normality assumption. In sum, parameter-based MASEM is especially useful when researchers intend to quantify and explain between-study variance in reliability coefficients.

Alternatively, researchers may choose a correlation-based MASEM approach. In contrast to parameter-based MASEM, the factor models are fit to the pooled correlation matrix, so that multiple, competing models can be tested and compared efficiently (Cheung, 2015). This allows researchers to synthesize multiple, model-based reliabilities for the entire meta-analytic sample, even if some correlations are missing. However, as random effects are estimated for the correlations, yet not the derived model parameters, quantifying heterogeneity in scale reliabilities across studies is limited to subgroup analyses (Jak & Cheung, 2018b). At the same, the recently developed one-stage MASEM approach may allow researchers to model this heterogeneity explicitly. In sum, correlation-based MASEM is

suitable to research questions surrounding the synthesis of model-based reliability coefficients, especially in the presence of missing data.

Recommendations for the Use of MASEM for Reliability Generalization

In the paper, we argued that MASEM can be utilized for the meta-analysis of reliability coefficients. The choice for using either correlation-based MASEM or parameter-based MASEM and, of correlation-based MASEM is chosen, either the one- or two-stage procedure, results in slightly different sequences of steps researchers need to take to synthesize reliabilities and/or quantify their heterogeneity. Table 3 summarizes these steps and provides some information about suitable software packages for their implementation. Overall, all procedures require the specification of a measurement model that fits the data to an acceptable extent, be it for all the primary studies individually (parameter-based MASEM) or the aggregated correlation matrix across all primary studies (correlation-based MASEM). This step is critical as it ultimately results in the choice for a reliability coefficient. Another critical step across all procedures is the inclusion of random effects, be it for the correlations as part of the pooling of correlation matrices (correlation-based MASEM) or for the scale reliabilities derived from the primary studies as part of the pooling of the model parameters (parameter-based MASEM). We recommend the R package metaSEM (Cheung, 2018) for implementing the correlation-based MASEM procedures; for the parameter-based MASEM procedure, any structural equation modeling software can be used for the first step and almost any available meta-analytic software package for the second step (*Note:* If researchers extract multiple reliability coefficients from the primary studies, multivariate meta-analysis must be available in the meta-analytic software).

Limitations and Future Research Directions

Despite the potential MASEM has for the meta-analysis of reliability coefficients, its limitations points to areas of future research.

Data availability. The key prerequisite for conducting especially correlation-based MASEM to the meta-analysis of scale reliabilities is that the authors of the primary studies report and make available the covariance or correlation matrices needed to perform SEM (Carpenter et al., 2016). In an era striving for replicability and developing best-practice standards of reporting (AERA et al., 2014; Hedges & Schauer, 2018), we do not consider this prerequisite to be a major issue—nevertheless, authors must be aware of which statistics are to be reported.

Assumptions on the distribution and structure of the data. As noted earlier, the pooling of correlation matrices in TSMASEM and OSMASEM and the estimation of the measurement model in OSMASEM are based on the multivariate normality assumption. This assumption may however not be reasonable, especially when item-item correlations are synthesized that were derived from ordinal item responses (Li et al., 2016). Although several analytic approaches exist to accommodate possible deviations from this assumption, including robust estimation procedures and procedures correcting test statistics and standard errors, the applicability and performance of these approaches to meta-analytic correlation matrices (in TSMASEM and OSMASEM) or single-study correlation matrices (in parameter-based MASEM) is still to be examined. Currently, the WLS estimation in the second stage TSMASEM does not rely on multivariate normality. As a consequence, both stage-1 TSMASEM and OSMASEM are not without limitations, and possible deviations from multivariate normality should be considered in their future development. Parameter-based MASEM can also be based on estimators that do not require the multivariate normality assumption—even further, some procedures may even be based on polychoric or polyserial correlation matrices (Yuan, Wu, & Bentler, 2011). Despite these adjustments, however, meta-analysts should be aware that fitting structural equation models to correlation matrices instead

of covariance matrices in the first stage of parameter-based MASEM can result in biased parameter estimates.

The possible nesting of study samples in studies or other, higher-order grouping variables represents another data issue in MASEM. In parameter-based MASEM, hierarchical data structure can be modeled explicitly in the second stage through multilevel univariate meta-analysis (Cheung, 2015). In both TSMASEM and OSMASEM, such a structure has not yet been implemented; the framework of multilevel multivariate meta-analysis could offer ways to estimate the additional variance components (e.g., McShane & Böckenholt, 2018). In this way, pooled reliability coefficients could be obtained at multiple levels of analysis (Geldhof et al., 2014).

Concerning the type of reliability coefficient, meta-analyzing Cronbach's α has dominated the research on reliability generalization, with several transformations of this reliability coefficient to improve the meta-analytic results made available (see Table 1). Given the current move from using Cronbach's α to using alternative, perhaps more suitable reliability coefficients (McNeish, 2018), the applicability of these transformations and their efficiency for combining several reliability coefficients should be examined further (see also Sánchez-Meca et al., 2013).

Exclusion of primary studies. Although researchers may extract the correlation matrices from a large sample of primary studies, this sample may be reduced during several modeling steps. Studies may be excluded due to non-positive definite matrices or convergence issues associated with the estimation of the measurement models. To still retain these studies, researchers may consider adding some parameter constraints, for instance, by fixing some factor loadings to defined values. Nevertheless, these two issues should become part of the research program to advance MASEM (Yu et al., 2018).

Types of reliability coefficients. Our tutorial focused on omega-coefficients to indicate scale reliability based on factor analysis. The procedures we described in this paper are by no means limited to these types of coefficients—instead, given the flexibility of estimating different factor models and, ultimately, use the model parameters for computing reliabilities, they can accommodate alternative reliability coefficients (e.g., Bentler, 2016, 2017).

Heterogeneity of reliability coefficients. Modeling the heterogeneity of the reliability coefficients, as they are computed from the parameters of factor-analytic models (i.e., factor variances, item residual variances, item factor loadings), is still to be developed further. At the time of writing, only parameter-based MASEM allowed researchers to model explicitly the random effects in the reliability coefficients that were computed based on the stage-1 structural equation model. As noted earlier, this MASEM approach, however, may not be applicable to data sets with missing correlations (Cheung & Cheung, 2016). In correlation-based MASEM, random effects are modeled for the correlations, yet not for the parameters of the structural equation model (Jak & Cheung, 2019). OSMASEM can incorporate moderating effects of study characteristics on factor loadings, structural paths, and (co-)variance components in a model; future versions of it may also include these effects on parameters computed from these components.

Conclusion

In conclusion, we hope to stimulate the use of MASEM for reliability generalization in future meta-analyses, especially when researchers intend to synthesize evidence on the psychometric quality of tests and scales across studies. We argue that the potential that lies within MASEM—be it correlation-based or parameter-based MASEM—can be utilized for synthesizing reliability coefficients and for quantifying between-study heterogeneity. MASEM has several advantages over the commonly used univariate approaches: First, it allows researchers to synthesize correlation matrices and to flexibly establish and compare

measurement models which form the basis for selecting suitable reliability coefficients. In this sense, researchers are given the possibility to consider model-based reliability coefficients. Second, MASEM can accommodate multiple item- or subscale-level correlations and/or multiple model parameters, such as scale or subscale reliabilities, taking into account the multivariate nature of the primary study data—ultimately, MASEM circumvents the possible bias associated with univariate approaches. At the same time, the existing MASEM approaches have several limitations, and researchers should select the MASEM approach considering the features of their meta-analytic data (e.g., the presence of missing data) and their research purposes (e.g., testing a set of measurement models to derive several reliability coefficients vs. synthesizing only one type of reliability coefficients). Overall, we believe that reliability generalization through MASEM opens a new field of research that contributes to understanding the psychometric quality of psychological scales and to examining the replicability of reliability coefficients across studies.

References

** Studies marked with an asterisk were included in the systematic review.*

AERA, APA, & NCME. (2014). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA.

*Aguayo, R., Vargas, C., de la Fuente, E. I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11*(2), 343-361.

*Bachner, Y. G., & O'Rourke, N. (2007). Reliability generalization of responses by care providers to the Zarit Burden Interview. *Aging & Mental Health, 11*(6), 678-685. https://doi.org/10.1080/13607860701529965

*Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*(4), 603-618. https://doi.org/10.1177/0013164402062004005

Bates, D., Maechler, M., et al. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-17. https://cran.r-project.org/web/packages/Matrix/index.html

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507-526. http://dx.doi.org/10.1037/met0000077

Becker, B. J. (1992). Using Results From Replicated Studies to Estimate Linear Models. *Journal of Educational and Behavioral Statistics, 17*(4), 341-362. https://doi.org/10.3102/10769986017004341

*Bentler, P. M. (2016). Covariate-free and covariate-dependent reliability. *Psychometrika, 81*(4), 907-920. https://doi.org/10.1007/s11336-016-9524-y

Bentler, P. M. (2017). Specificity-enhanced reliability coefficients. *Psychological Methods, 22*(3), 527-540. https://doi.org/10.1037/met0000092

*Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63*(1), 75-95. https://doi.org/10.1177/0013164402239318

*Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*(4), 570-589. https://doi.org/10.1177/0013164402062004003

*Beretvas, S. N., Suizzo, M.-A., Durham, J. A., & Yarnell, L. M. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's locus of control

scales. Educational and Psychological Measurement, 68(1), 97-119. https://doi.org/10.1177/0013164407301529

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.

Bonett, D. G. (2002). Sample Size Requirements for Testing and Estimating Coefficient Alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335-340. https://doi.org/10.3102/10769986027004335

*Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15*(4), 368-385. https://doi.org/10.1037/a0020142

*Botella, J., & Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23*(3), 516-522.

*Botella, J., & Suero, M. (2012). Managing heterogeneity of variance in studies of reliability generalization with alpha coefficients. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8*(2), 71-80. https://doi.org/10.1027/1614-2241/a000039

*Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*(4), 386-397. https://doi.org/10.1037/a0019626

Brannick, M. T., & Zhang, N. (2013). Bayesian meta-analysis of coefficient alpha. *Research Synthesis Methods, 4*(2), 198-207. https://doi.org/10.1002/jrsm.1075

*Breidbord, J., & Croudace, T. J. (2013). Reliability generalization for Childhood Autism Rating Scale. *Journal of Autism and Developmental Disorders, 43*(12), 2855-2865. https://doi.org/10.1007/s10803-013-1832-9

*Campbell, J. S., Pulos, S., Hogan, M., & Murry, F. (2005). Reliability Generalization of the Psychopathy Checklist Applied in Youthful Samples. *Educational and Psychological Measurement, 65*(4), 639-656. https://doi.org/10.1177/0013164405275666

*Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement, 61*(3), 373-386. https://doi.org/10.1177/00131640121971266

*Capraro, R. M., & Capraro, M. M. (2002). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement*, 62(4), 590-602. https://doi.org/10.1177/0013164402062004004

Carpenter, N. C., Son, J., Harris, T. B., Alexander, A. L., & Horner, M. T. (2016). Don't Forget the Items: Item-Level Meta-Analytic and Substantive Validity Techniques for Reexamining Scale Validation. *Organizational Research Methods, 19*(4), 616-650. https://doi.org/10.1177/1094428116639132

*Caruso, J. C. (2000). Reliability generalization of the NEO Personality Scales. *Educational and Psychological Measurement, 60*(2), 236-254. https://doi.org/10.1177/00131640021970484

*Caruso, J. C., & Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences, 31*(2), 173-184. https://doi.org/10.1016/S0191-8869%2800%2900126-4

*Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. D. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*(4), 675-689. https://doi.org/10.1177/00131640121971437

Cheung, M. W.-L. (2009). Constructing Approximate Confidence Intervals for Parameters With Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(2), 267-294. https://doi.org/10.1080/10705510902751291

Cheung, M. W.-L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(1), 157-167. https://doi.org/10.1080/10705511.2013.742404

Cheung, M. W.-L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavioral Research Methods, 46*, 29-40. https://doi.org/10.3758/s13428-013-0361-y

Cheung, M. W.-L. (2015). *Meta-Analysis: A Structural Equation Modeling Approach*. Chichester, West Sussex: John Wiley & Sons, Ltd.

Cheung, M. W.-L. (2018a). Issues in solving the problem of effect size heterogeneity in meta-analytic structural equation modeling: A commentary and simulation study on Yu, Downes, Carter, and O'Boyle (2016). *Journal of Applied Psychology, 103*(7), 787-803. https://doi.org/10.1037/apl0000284

Cheung, M. W.-L. (2018b). *metaSEM: Meta-Analysis using Structural Equation Modeling (Version 1.2.0)*. Retrieved from https://cran.r-project.org/web/packages/metaSEM/index.html

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10*(1), 40-64. https://doi.org/10.1037/1082-989X.10.1.40

Cheung, M. W.-L., & Chan, W. (2009). A Two-Stage Approach to Synthesizing Covariance Matrices in Meta-Analytic Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 28-53. https://doi.org/10.1080/10705510802561295

Cheung, M. W.-L., & Cheung, S. F. (2016). Random-effects models for meta-analytic structural equation modeling: review, issues, and illustrations. *Research Synthesis Methods, 7*(2), 140-155. https://doi.org/10.1002/jrsm.1166

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*(738), https://doi.org/10.3389/fpsyg.2016.00738

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and Behavioral Statistics, 40*(2), 136-157. https://doi.org/10.3102/1076998615570945

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12*(1), 1-16. https://doi.org/10.1007/bf02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/bf02310555

*Crouch, M. K., Mack, D. E., Wilson, P. M., & Kwan, M. Y. W. (2017). Variability of coefficient alpha: An empirical investigation of the scales of psychological wellbeing. *Review of General Psychology, 21*(3), 255-268. https://doi.org/10.1037/gpr0000112

*Deditius-Island, H. K., & Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62*(4), 728-734. https://doi.org/10.1177/0013164402062004012

*Deng, J., Wang, M.-C., Zhang, X., Shou, Y., Gao, Y., & Luo, J. (2019). The Inventory of Callous Unemotional Traits: A reliability generalization meta-analysis. *Psychological Assessment, 31*(6), 765-780. https://doi.org/10.1037/pas0000698

Deng, L., & Chan, W. (2016). Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and Psychological Measurement, 77*(2), 185-203. https://doi.org/10.1177/0013164416658325

*Dunn, T. W., Smith, T. B., & Montoya, J. A. (2006). Multicultural Competency Instrumentation: A Review and Analysis of Reliability Generalization. *Journal of Counseling & Development, 84*(4), 471-482. https://doi.org/10.1002/j.1556-6678.2006.tb00431.x

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541-562. https://doi.org/10.1037/met0000083

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72-91. https://doi.org/10.1037/a0032138

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*(2), 130-139. https://doi.org/10.1027/1015-5759/a000181

Gignac, G. E., Reynolds, M. R., & Kovacs, K. (2017). Digit Span Subscale Scores May Be Insufficiently Reliable for Clinical Interpretation: Distinguishing Between Stratified Coefficient Alpha and Omega Hierarchical. *Assessment*, 1073191117748396. https://doi.org/10.1177/1073191117748396

*Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. https://doi.org/10.1016/j.jrp.2014.06.003

Gnambs, T., Scharl, A., & Schroeders, U. (2018a). The structure of the Rosenberg Self-Esteem Scale: A cross-cultural meta-analysis. *Zeitschrift für Psychologie, 226*(1), 14-29. https://doi.org/10.1027/2151-2604/a000317

Gnambs, T., Scharl, A., & Schroeders, U. (2018b). Data: The structure of the Rosenberg Self-Esteem Scale: A cross-cultural meta-analysis. https://osf.io/uwfsp/

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. *Educational and Psychological Measurement, 66*(6), 930-944. https://doi.org/10.1177/0013164406288165

*Graham, J. M., & Christiansen, K. (2009). The reliability of romantic love: A reliability generalization meta-analysis. *Personal Relationships, 16*(1), 49-66. https://doi.org/10.1111/j.1475-6811.2009.01209.x

*Graham, J. M., & Unterschute, M. S. (2015). A reliability generalization meta-analysis of self-report measures of adult attachment. *Journal of Personality Assessment, 97*(1), 31-41. https://doi.org/10.1080/00223891.2014.927768

*Graham, J. M., Diebels, K. J., & Barnow, Z. B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology, 25*(1), 39-48. https://doi.org/10.1037/a0022441

*Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A reliability generalization meta-analysis. *Journal of Marriage and Family, 68*(3), 701-717. https://doi.org/10.1111/j.1741-3737.2006.00284.x

*Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-analysis of coefficient alpha: A reliability generalization stud. *Journal of Management Studies, 55*(4), 583-618. https://doi.org/10.1111/joms.12328

Green, S. B., & Yang, Y. (2015). Evaluation of Dimensionality in the Assessment of Internal Consistency Reliability: Coefficient Alpha and Omega Coefficients. *Educational Measurement: Issues and Practice, 34*(4), 14-20. https://doi.org/10.1111/emip.12100

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of Reliability Using Coefficient Alpha and Structural Equation Modeling When Assumptions of Tau-Equivalence and Uncorrelated Errors Are Violated. *Methodology, 9*(1), 30-40. https://doi.org/10.1027/1614-2241/a000052

*Ha, J. H., Lee, S. M., & Puig, A. (2010). A reliability generalization study of the Frost Multidimensional Perfectionism Scale (F-MPS). *Psychological Reports, 107*(1), 95-112. https://doi.org/10.2466/03.09.20.PR0.107.4.95-112

Hafdahl, A. R. (2007). Combining correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics, 32*(2), 180-205. https://doi.org/10.3102/1076998606298041

Hakstian, A. R., & Whalen, T. E. J. P. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219-231. https://doi.org/10.1007/bf02291840

*Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory scale scores. *Educational and Psychological Measurement, 62*(4), 659-673. https://doi.org/10.1177/0013164402062004008

Hedges, L. V., & Schauer, J. M. (2018). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*. https://doi.org/10.1037/met0000189

*Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A Reliability Generalization Study on the Survey of Perceived Organizational Support: The Effects of Mean Age and Number of Items on Score Reliability. *Educational and Psychological Measurement, 66*(4), 631-642. https://doi.org/10.1177/0013164406288158

*Hellman, C. M., Muilenburg-Trevino, E. M., & Worley, J. A. (2008). The belief in a just world: An examination of reliability estimates across three measures. *Journal of Personality Assessment, 90*(4), 399-401. https://doi.org/10.1080/00223890802108238

*Hellman, C. M., Pittman, M. K., & Munoz, R. T. (2013). The first twenty years of the will and the ways: An examination of score reliability distribution on Snyder's Dispositional Hope Scale. *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being, 14*(3), 723-729. https://doi.org/10.1007/s10902-012-9351-5

*Henson, R. K., & Hwang, D.-Y. (2002). Variability and prediction of measurement error in a Kolb's Learning Style Inventory Scores: A reliability generalization study. *Educational and Psychological Measurement, 62*(4), 712-727. https://doi.org/10.1177/0013164402062004011

*Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*(3), 404-420. https://doi.org/10.1177/00131640121971284

*Herrington, H. M., Smith, T. B., Feinauer, E., & Griner, D. (2016). Reliability generalization of the Multigroup Ethnic Identity Measure-Revised (MEIM-R). *Journal of Counseling Psychology, 63*(5), 586-593. https://doi.org/10.1037/cou0000148

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Holland, D. F. (2015). *Reliability generalization: A systematic review and evaluation of meta-analytic methodology and reporting practice.* University of North Texas, Denton, TX. Retrieved from https://digital.library.unt.edu/ark:/67531/metadc822810/

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Hunter, J. E., & Schmidt, F. L. (2014). *Methods of meta-analysis: Correcting Error and Bias in Research Findings* (3rd ed.). Newbury Park, CA: Sage Publications.

*Huynh, Q.-L., Howell, R. T., & Benet-Martínez, V. (2009). Reliability of bidimensional acculturation scores: A meta-analysis. *Journal of Cross-Cultural Psychology, 40*(2), 256-274. https://doi.org/10.1177/0022022108328919

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine, 27*(5), 670-686. https://doi.org/10.1002/sim.2913

Jak, S. (2015). *Meta-Analytic Structural Equation Modelling*. Cham, Switzerland: Springer.

Jak, S., & Cheung, M. W.-L. (2018a). Accounting for Missing Correlation Coefficients in Fixed-Effects MASEM. *Multivariate Behavioral Research, 53*(1), 1-14. https://doi.org/10.1080/00273171.2017.1375886

Jak, S., & Cheung, M. W.-L. (2018b). Testing moderator hypotheses in meta-analytic structural equation modeling using subgroup analysis. *Behavior Research Methods, 50*(4), 1359-1373. https://doi.org/10.3758/s13428-018-1046-3

Jak, S., & Cheung, M. W.-L. (2019, in press). Meta-Analytic Structural Equation Modeling with Moderating Effects on SEM Parameters. *Psychological Methods*, https://doi.org/10.31234/osf.io/ce85j

*Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJTs). *European Journal of Psychological Assessment, 32*(3), 230-240. https://doi.org/10.1027/1015-5759/a000250

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69-92. https://doi.org/10.1037/a0040086

*Kennedy, R. S., & Turnage, J. J. (1991). Reliability generalization: A viable key for establishing validity generalization. *Perceptual and Motor Skills, 72*(1), 297-298. https://doi.org/10.2466/PMS.72.1.297-298

*Kieffer, K. M., & MacDonald, G. (2011). Exploring factors that affect score reliability and variability: In the Ways of Coping Questionnaire reliability coefficients: A meta-analytic reliability generalization study. *Journal of Individual Differences, 32*(1), 26-38. https://doi.org/10.1027/1614-0001/a000031

*Kieffer, K. M., & Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62*(6), 969-994. https://doi.org/10.1177/0013164402238085

*Kieffer, K. M., Cronin, C., & Fister, M. C. (2004). Exploring Variability and Sources of Measurement Error in Alcohol Expectancy Questionnaire Reliability Coefficients: A Meta-Analytic Reliability Generalization Study. *Journal of Studies on Alcohol, 65*(5), 663-671. https://doi.org/10.15288/jsa.2004.65.663

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

*Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement, 62*(4), 685-711. https://doi.org/10.1177/0013164402062004010

*Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A Reliability Generalization Study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66*(2), 285-304. https://doi.org/10.1177/0013164405284030

*Leue, A., & Lange, S. (2011). Reliability generalization: An examination of the positive affect and negative affect schedule. *Assessment, 18*(4), 487-501. https://doi.org/10.1177/1073191110374917

*Li, A., & Bagger, J. (2007). The balanced inventory of desirable responding (BIDR): A reliability generalization study. *Educational and Psychological Measurement, 67*(3), 525-544. https://doi.org/10.1177/0013164406292087

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. Psychological Methods, 21(3), 369–387. https://doi.org/10.1037/met0000093

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marin-Martínez, F. (2013). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics, 38*(5), 443-469. https://doi.org/10.3102/1076998612466142

*López-Pina, J. A., Sánchez-Meca, J., & Rosa-Alcázar, A. I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology, 9*(1), 143-159.

*López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marin-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., . . . Ferrer-Requena, J. (2015a). Reliability generalization study of the Yale-Brown Obsessive-Compulsive Scale for children and adolescents. *Journal of Personality Assessment, 97*(1), 42-54. https://doi.org/10.1080/00223891.2014.930470

*López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marin-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., . . . Ferrer-Requena, J. (2015b). The Yale-Brown Obsessive Compulsive Scale: A reliability generalization meta-analysis. *Assessment, 22*(5), 619-628. https://doi.org/10.1177/1073191114551954

*Luo, J., Zhao, S.-Y., Pan, Y., & Dai, X.-Y. (2013). A reliability generalization analysis of the Adolescence Time Management Disposition Inventory. *Chinese Mental Health Journal, 27*(4), 305-309.

*Maes, M., Van den Noortgate, W., & Goossens, L. (2015). A reliability generalization study for a multidimensional loneliness scale: The Loneliness and Aloneness Scale for Children and Adolescents. *European Journal of Psychological Assessment, 31*(4), 294-301. https://doi.org/10.1027/1015-5759/a000237

*Mark, W., & Toulopoulou, T. (2016). Psychometric properties of "Community assessment of psychic experiences": Review and meta-analyses. *Schizophrenia Bulletin, 42*(1), 34-44.

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524-545. https://doi.org/10.1037/met0000113

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*, 320-341. https://doi.org/10.1207/s15328007sem1103_2

McArdle, J. J. (2005). The Development of the RAM Rules for Latent Variable Structural Equation Modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (p. 225–273). Lawrence Erlbaum Associates Publishers.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

McShane, B. B., & Böckenholt, U. (2018). Multilevel Multivariate Meta-analysis with Application to Choice Overload. *Psychometrika, 83*(2), 255-271. https://doi.org/10.1007/s11336-017-9571-z

*Miller, B. K., Byrne, Z. S., Rutherford, M. A., & Hansen, A. M. (2009). Perceptions of Organizational Politics: A demonstration of the reliability generalization technique. *Journal of Managerial Issues, 21*(2), 280-300.

*Miller, C. S., Shields, A. L., Campfield, D., Wallace, K. A., & Weiss, R. D. (2007). Substance use scales of the Minnesota Multiphasic Personality Inventory: An exploration of score reliability via meta-analysis. *Educational and Psychological Measurement, 67*(6), 1052-1065. https://doi.org/10.1177/0013164406299130

*Miller, C. S., Woodson, J., Howell, R. T., & Shields, A. L. (2009). Measurements instruments scales tests: (SASSI): Assessing the reliability of scores produced by the Substance Abuse Subtle Screening Inventory. *Substance Use & Misuse, 44*(8), 1090-1100. https://doi.org/10.1080/10826080802486772

*Mji, A., & Alkhateeb, H. M. (2005). Combining Reliability Coefficients: Toward Reliability Generalization of the Conceptions of Mathematics Questionnaire. *Psychological Reports, 96*(3), 627-634. https://doi.org/10.2466/PR0.96.3.627-634

*Montano, S. A., Lewey, J. H., O'Toole, S. K., & Graves, D. (2016). Reliability generalization of the Texas Revised Inventory of Grief (TRIG). *Death Studies, 40*(4), 256-262. https://doi.org/10.1080/07481187.2015.1129370

Neale, M. C., Miller, M. B. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27*, 113-120. https://doi.org/10.1023/A:1025681223921

*Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62*(4), 647-658. https://doi.org/10.1177/0013164402062004007

Novick, M. R., & Lewis, C. J. P. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*(1), 1-13. https://doi.org/10.1007/bf02289400

*O'Rourke, N. (2004). Reliability Generalization of Responses by Care Providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement, 64*(6), 973-990. https://doi.org/10.1177/0013164404268668

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The Performance of ML, GLS, and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 557-595. https://doi.org/10.1207/S15328007SEM0704_3

Oort, F. J., & Jak, S. (2016). Maximum likelihood estimation in meta-analytic structural equation modeling. *Research Synthesis Methods, 7*(2), 156-167. https://doi.org/10.1002/jrsm.1203

*Orgiles, M., Fernandez-Martínez, I., Guillen-Riquelme, A., Espada, J. P., & Essau, C. A. (2016). A systematic review of the factor structure and reliability of the Spence Children's Anxiety Scale. *Journal of Affective Disorders, 190*, 333-340. https://doi.org/10.1016/j.jad.2015.09.055

Padilla, M. A., & Divers, J. (2015). A Comparison of Composite Reliability Estimators: Coefficient Omega Confidence Intervals in the Current Literature. *Educational and Psychological Measurement, 76*(3), 436-453. https://doi.org/10.1177/0013164415593776

*Piqueras, J. A., Martin-Vivar, M., Sandin, B., San Luis, C., & Pineda, D. (2017). The Revised Child Anxiety and Depression Scale: A systematic review and reliability generalization meta-analysis. *Journal of Affective Disorders, 218*, 153-169. https://doi.org/10.1016/j.jad.2017.04.022

Raykov, T. (1997). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. *Multivariate Behavioral Research, 32*(4), 329-353. https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*(2), 299-331. https://doi.org/10.1016/S0005-7894(04)80041-8

Raykov, T., & Marcoulides, G. A. (2013). Meta-Analysis of Scale Reliability Using Latent Variable Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 338-353. https://doi.org/10.1080/10705511.2013.769396

Raykov, T., & Marcoulides, G. A. (2014). A Direct Latent Variable Modeling Based Method for Point and Interval Estimation of Coefficient Alpha. *Educational and Psychological Measurement, 75*(1), 146-156. https://doi.org/10.1177/0013164414526039

*Reese, R. J., Kieffer, K. M., & Briggs, B. K. (2002). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement, 62*(4), 619-646. https://doi.org/10.1177/0013164402062004006

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

*Rexrode, K. R., Petersen, S., & O'Toole, S. (2008). The Ways of Coping Scale: A reliability generalization study. *Educational and Psychological Measurement, 68*(2), 262-280. https://doi.org/10.1177/0013164407310128

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373. https://doi.org/10.1037/a0029315

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. https://doi.org/10.1037/met0000045

*Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*(3), 306-322. https://doi.org/10.1037/1082-989X.11.3.306

*Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability Generalization of the Patterns of Adaptive Learning Survey Goal Orientation Scales. *Educational and Psychological Measurement, 65*(3), 451-464. https://doi.org/10.1177/0013164404272496

Rosseel, Y. (2018). *lavaan: Latent Variable Analysis (Version 0.6-3)*. Retrieved from https://cran.r-project.org/web/packages/lavaan/index.html

*Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI-2 PSY-5 Scales. *Journal of Personality Assessment, 88*(3), 264-275. https://doi.org/10.1080/00223890701293908

*Rubio-Aparicio, M., Núñez-Núñez, R. M., Sánchez-Meca, J., López-Pina, J. A., Marin-Martínez, F., & López-López, J. A. (2018). The Padua Inventory–Washington State University Revision of Obsessions and Compulsions: A Reliability Generalization Meta-Analysis. *Journal of Personality Assessment*, https://doi.org/10.1080/00223891.2018.1483378

*Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted.

British Journal of Mathematical and Statistical Psychology, 66(3), 402-425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

*Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marin-Martínez, F., Rosa-Alcázar, A. I., & Gómez-Conesa, A. A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology, 11*(3), 473-493.

*Sánchez-Meca, J., Rubio-Aparicio, M., Núñez-Núñez, R. M., López-Pina, J., Marin-Martínez, F., & López-López, J. A. (2017). A reliability generalization meta-analysis of the Padua Inventory of obsessions and compulsions. The Spanish Journal of Psychology, 20. https://doi.org/10.1017/sjp.2017.65

Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education, 128*, 13-35. https://doi.org/10.1016/j.compedu.2018.09.009

Scherer, R., Siddiq, F., & Tondeur, J. (2020). All the same or different? Revisiting measures of teachers' technology acceptance. *Computers & Education, 143*, 103656. https://doi.org/10.1016/j.compedu.2019.103656

*Schipke, D., & Freund, P. A. (2012). A meta-analytic reliability generalization of the Physical Self-Description Questionnaire (PSDQ). *Psychology of Sport and Exercise, 13*(6), 789-797. https://doi.org/10.1016/j.psychsport.2012.04.012

Schmitt, D. P., & Allik, J. (2005). Simultaneous Administration of the Rosenberg Self-Esteem Scale in 53 Nations: Exploring the Universal and Culture-Specific Features of Global Self-Esteem. *Journal of Personality and Social Psychology, 89*(4), 623-642. https://doi.org/10.1037/0022-3514.89.4.623

Sheng, Z., Kong, W., Cortina, J. M., & Hou, S. (2016). Analyzing matrices of meta-analytic correlations: current practices and recommendations. *Research Synthesis Methods, 7*(2), 187-208. https://doi.org/10.1002/jrsm.1206

*Shields, A. L., & Caruso, J. C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement, 63*(3), 404-413. https://doi.org/10.1177/0013164403063003004

*Shields, A. L., & Caruso, J. C. (2004). A Reliability Induction and Reliability Generalization Study of the CAGE Questionnaire. *Educational and Psychological Measurement, 64*(2), 254-270. https://doi.org/10.1177/0013164403261814

Sijtsma, K. (2008). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107. https://doi.org/10.1007/s11336-008-9101-0

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: SAGE Publications.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods, 7*(2), 210-227. https://doi.org/10.1037/1082-989X.7.2.210

*Sun, S., & Wang, S. (2015). The Children's Depression Inventory in worldwide child development research: A reliability generalization study. *Journal of Child and Family Studies, 24*(8), 2352-2363. https://doi.org/10.1007/s10826-014-0038-x

Tang, R. W., & Cheung, M. W.-L. (2016). Testing IB theories with meta-analytic structural equation modeling: The TSMASEM approach and the Univariate-r approach. *Review of International Business and Strategy, 26*(4), 472-492. https://doi.org/10.1108/RIBS-04-2016-0022

Teo, T., & Fan, X. (2013). Coefficient Alpha and Beyond: Issues and Alternatives for Educational Research. *The Asia-Pacific Education Researcher, 22*(2), 209-213. https://doi.org/10.1007/s40299-013-0075-z

*Therrien, Z., & Hunsley, J. (2013). Assessment of anxiety in older adults: A reliability generalization meta-analysis of commonly used measures. Clinical Gerontologist: The *Journal of Aging and Mental Health, 36*(3), 171-194. https://doi.org/10.1080/07317115.2013.767871

*Thompson, B., & Cook, C. (2002). Stability of the reliability of LibQUAL+TM scores: A reliability generalization meta-analysis study. *Educational and Psychological Measurement, 62*(4), 735-743. https://doi.org/10.1177/0013164402062004013

*Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20. https://doi.org/10.1177/0013164498058001002

*Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement, 61*(1), 45-59. https://doi.org/10.1177/00131640121971059

*Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A., & Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment, 8*(4), 391-401. https://doi.org/10.1177/107319110100800404

*Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the Life Orientation Test. *Journal of Personality Assessment, 92*(4), 362-370. https://doi.org/10.1080/00223891.2010.482016

*Vassar, M., & Bradley, G. (2012). A reliability generalization meta-analysis of coefficient alpha for the Reynolds Adolescent Depression Scale. *Clinical Child Psychology and Psychiatry, 17*(4), 519-527. https://doi.org/10.1177/1359104511424998

*Vassar, M., & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the UCLA Loneliness Scale. *Journal of Personality Assessment, 90*(6), 601-607. https://doi.org/10.1080/00223890802388624

*Vassar, M., Knaup, K. G., Hale, W., & Hale, H. (2011). A meta-analysis of coefficient alpha for the Impact of Event Scales: A reliability generalization study. *South African Journal of Psychology, 41*(1), 6-16. https://doi.org/10.1177/008124631104100102

*Victorson, D., Barocas, J., Song, J., & Cella, D. (2008). Reliability across studies from the Functional Assessment of Cancer Therapy-General(FACT-G) and its subscales: A reliability generalization. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 17*(9), 1137-1146. https://doi.org/10.1007/s11136-008-9398-2

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48*(4), 865-885. https://doi.org/10.1111/j.1744-6570.1995.tb01784.x

*Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*(2), 224-235. https://doi.org/10.1177/00131640021970475

*Wallace, K. A., & Wheeler, A. J. (2002). Reliability generalization of the Life Satisfaction Index. *Educational and Psychological Measurement, 62*(4), 674-684. https://doi.org/10.1177/0013164402062004009

*Warne, R. T. (2011). A reliability generalization of the Overexcitability Questionnaire-Two. *Journal of Advanced Academics, 22*(5), 671-692. https://doi.org/10.1177/1932202X11424881

*Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. B. (2011). A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*(1), 231-244. https://doi.org/ 10.1177/0013164410391579

Wilson, S. J., Polanin, J. R., & Lipsey, M. W. (2016) Fitting meta-analytic structural equation models with complex datasets. *Research Synthesis Methods, 7*, 121-139. https://doi.org/10.1002/jrsm.1199.

Yang, Y., & Green, S. B. (2011). Coefficient Alpha: A Reliability Coefficient for the 21st Century? *Journal of Psychoeducational Assessment, 29*(4), 377-392. https://doi.org/10.1177/0734282911406668

*Yeo, S. (2011). Reliability generalization of curriculum-based measurement reading aloud: A meta-analytic review. *Exceptionality, 19*(2), 75-93. https://doi.org/10.1080/09362835.2011.562094

*Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*(2), 201-223. https://doi.org/10.1177/00131640021970466

*Youngstrom, E. A., & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement, 63*(2), 279-295. https://doi.org/10.1177/0013164403253226

Yu, J. (J.), Downes, P. E., Carter, K. M., & O'Boyle, E. (2018). The heterogeneity problem in meta-analytic structural equation modeling (MASEM) revisited: A reply to Cheung. *Journal of Applied Psychology, 103*(7), 804-811. https://doi.org/10.1037/apl0000328

Yu, J. (J.), Downes, P. E., Carter, K. M., & O'Boyle, E. H. (2016). The problem of effect size heterogeneity in meta-analytic structural equation modeling. *Journal of Applied Psychology, 101*(10), 1457-1473. https://doi.org/10.1037/apl0000141

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology, 64*(1), 107-133. https://doi.org/10.1348/000711010X497442

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. J. P. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. https://doi.org/10.1007/s11336-003-0974-7

Tables

Table 1

Methods of Transforming Cronbach's α (see also López-López et al., 2013)

Transformation method	*Coefficient*	*Back-transformation*	*Sampling variance*
Raw coefficient	$\hat{\alpha}_i$	–	$V(\hat{\alpha}_i) = \frac{2k_i(1-\hat{\alpha}_i)^2}{(k_i-1)\{N_i-2-[(k_i-2)(J-1)]^{1/4}\}}$
Fisher's Z	$Z_i = \frac{1}{2}\ln\left(\frac{1+\hat{\alpha}_i}{1-\hat{\alpha}_i}\right)$	$\hat{\alpha}_i = \frac{e^{2Z_i}-1}{e^{2Z_i}+1}$	$V(Z_i) = \frac{1}{N_i-3}$
HW76	$T_i = \sqrt[3]{1-\hat{\alpha}_i}$	$\hat{\alpha}_i = 1-T_i^3$	$V(T_i) = \frac{18k_i(N_i-1)(1-\hat{\alpha}_i)^{2/3}}{(k_i-1)(9N_i-11)^2}$
BO02	$L_i = \ln(1-\lvert\hat{\alpha}_i\rvert)$	$\hat{\alpha}_i = 1-e^{L_i}$	$V(L_i) = \frac{2k_i}{(k_i-1)(N_i-2)}$

Note. BO02 = Bonett's (2002) approach, HW76 = Hakstian-Whalen (1976) approach. k_i = Number of items in the *i*th study, J = Number of studies, N_i = Sample size of the *i*th study.

Table 2

Comparison of Meta-Analytic Structural Equation Modeling Approaches to Synthesizing Scale Reliabilities

	Correlation-based MASEM through TSMASEM	**Correlation-based MASEM through OSMASEM**	**Parameter-Based MASEM**
Procedure			
Stage 1	Pooling the correlation matrices across primary studies under a fixed- or random-effects model	Pooling the correlation matrices and specifying the measurement model based on fixed or random effects of correlations	Specifying the measurement model based on the correlation matrices of each primary study and estimating scale reliabilities based on the resultant model parameters
Stage 2	Specifying a measurement model based on the pooled correlation matrix to obtain factor loadings, residual variances, and the factor variance		Pooling the scale reliabilities across primary studies by univariate meta-analysis (including possible effect-size transformations) or by multivariate meta-analysis of factor loadings
Synthesis of scale reliabilities	Estimated based on the stage-2 model parameters	Estimated based on the model parameters	Estimated for each primary study based on the stage-1 model parameters and then pooled
Heterogeneity across studies	Random effects of correlations	Random effects of correlations	Random effects of model parameters
Moderator analysis	Subgroup analyses at stage 2 to identify the effects of categorical moderators with few categories (invariance constraints can be imposed across subgroups)	Moderation effects on the model matrices (e.g., the A-matrix containing the factor loadings and structural coefficients, the S-matrix containing the variances and covariances in the model)	Stage-2 meta-analysis can be extended to mixed-effects models with moderators

Handling missing correlations	Possible through ML-based procedures at stage 1	Possible through ML-based procedures	Stage 1 requires complete correlation matrices

Note. FIML = Full-Information-Maximum-Likelihood procedure, MASEM = Meta-analytic structural equation modeling, ML = Maximum Likelihood, TSMASEM = Two-stage structural equation modeling, OSMASEM = One-stage MASEM.

Table 3

Suggested Steps of Synthesizing Reliability Coefficients through Correlated- and Parameter-Based MASEM

	Correlation-based MASEM through TSMASEM	**Correlation-based MASEM through OSMASEM**	**Parameter-based MASEM**
Data preparation	Extracting correlation matrices and sample sizes from the primary studies	Extracting correlation matrices and sample sizes from the primary studies	Extracting correlation matrices and sample sizes or factor loadings, residual variances, factor variances, and sample sizes from the primary studies
Analytic steps	1. Checking correlation matrices for positive definiteness and excluding non-positive definite matrices 2. Pooling the correlation matrices with fixed or random effects (stage-1 model) 3. Comparing the fixed- and random-effects models at stage 1 (e.g., heterogeneity indices, homogeneity tests) 4. Specifying, estimating, and evaluating the congeneric factor model to the pooled correlation matrix 5. Specifying, estimating, and evaluating alternative measurement models (e.g., residual covariances, bifactor structure, equal factor loadings)	1. Checking correlation matrices for positive definiteness and excluding non-positive definite matrices 2. Pooling the correlation matrices and specifying, estimating, and evaluating the congeneric factor model to the pooled correlation matrix 3. Pooling the correlation matrices and specifying, estimating, and evaluating alternative measurement models (e.g., residual covariances, bifactor structure, equal factor loadings) 4. Comparing measurement models and deciding for a final model 5. Estimating the overall scale reliability from the parameters of the final model 6. *(Optional)* Evaluating possible moderator effects on the model	*If correlation matrices are extracted:* 1. Checking correlation matrices for positive definiteness and excluding non-positive definite matrices 2. Specifying, estimating, and evaluating the congeneric factor model to the correlation matrices of each primary study 3. Specifying, estimating, and evaluating alternative measurement models (e.g., residual covariances, bifactor structure, equal factor loadings) 4. Comparing measurement models and deciding for a final model 5. Estimating the scale reliabilities from the parameters of the final model 6. Performing univariate meta-analysis to the scale reliabilities using fixed- or random-effects models

	6. Comparing measurement models and deciding for a final model (stage-2 model) 7. Estimating the overall scale reliability from the parameters of the final model 8. *(Optional)* Performing subgroup analyses to identify possible reliability differences across subgroups of studies or study samples (including possible invariance testing across subgroups)	parameter matrices, including the variance explained in the random effects of the correlations	7. Comparing the fixed- and random-effects models from step 6 (e.g., heterogeneity indices, homogeneity tests) 8. Estimating the overall scale reliability and the between-study variance 9. *(Optional)* Mixed-effects modeling to examine possible moderator effects *If model parameters are extracted:* 1. Estimating the scale reliabilities from the model parameters 2. Performing univariate or multivariate meta-analysis (depending on whether a single or multiple reliability coefficients are extracted) to the scale reliabilities using fixed- or random-effects models 3. Comparing the fixed- and random-effects models from step 2 (e.g., heterogeneity indices, homogeneity tests) 4. Estimating the overall scale reliability and the between-study variance 5. *(Optional)* Mixed-effects modeling to examine possible moderator effects

Software packages	R package `metaSEM`	R package `metaSEM`	Structural equation modeling software (e.g., M*plus*, LISREL, AMOS, R package `lavaan`) and meta-analytic software (e.g., Comprehensive Meta-Analysis; R packages `metafor`, `metaSEM`, `meta`, `robumeta`)

Table 4

Pooled Correlation Matrices Based on Fixed- and Random-Effects TSMASEM (Example 1)

Subscales	1.	2.	3.	4.	5.	6.
1. PU	1.000	0.502	0.598	0.356	0.457	0.375
2. PEOU	0.472	1.000	0.552	0.221	0.493	0.419
3. ATT	0.577	0.517	1.000	0.281	0.450	0.400
4. SN	0.382	0.254	0.298	1.000	0.243	0.246
5. TSE	0.424	0.462	0.405	0.255	1.000	0.295
6. FC	0.314	0.388	0.359	0.258	0.284	1.000

Note. Correlations in the upper diagonal are based on fixed effects. Correlations in the lower diagonal are based on random effects. PEOU = Perceived ease of use, PU = Perceived usefulness, ATT = Attitudes toward technology, SN = Subjective norms, FC = Facilitating conditions, TSE = Technology self-efficacy.

Table 5

Factor Loadings and Residual Variances of the RSES Items Obtained from the Random-Effects TSMASEM Based on the ***Nested-Factor Model*** *(Example 2)*

	Factor loadings		*Residual variances*
Items	General factor	Specific factors	
Item 1	0.756	-0.052	0.425
	[0.729, 0.782]	[-0.135, 0.028]	[0.376, 0.468]
Item 2	0.525	-0.593	0.372
	[0.501, 0.551]	[-0.637, -0.548]	[0.326, 0.415]
Item 3	0.595	0.524	0.371
	[0.561, 0.626]	[0.412, 0.636]	[0.277, 0.466]
Item 4	0.522	0.309	0.632
	[0.498, 0.546]	[0.243, 0.375]	[0.582, 0.671]
Item 5	0.518	-0.335	0.620
	[0.493, 0.545]	[-0.374, -0.293]	[0.598, 0.641]
Item 6	0.504	-0.604	0.381
	[0.484, 0.527]	[-0.646, -0.561]	[0.333, 0.425]
Item 7	0.616	0.331	0.511
	[0.588, 0.642]	[0.255, 0.410]	[0.456, 0.557]
Item 8	0.371	-0.405	0.698
	[0.338, 0.405]	[-0.460, -0.350]	[0.661, 0.730]
Item 9	0.585	-0.394	0.502
	[0.560, 0.613]	[-0.434, -0.351]	[0.476, 0.527]
Item 10	0.800	-0.035	0.358
	[0.772, 0.827]	[-0.121, 0.046]	[0.306, 0.403]
Scale reliability ω_H		0.745 [0.735, 0.755]	

Note. LBCI = Likelihood-based confidence intervals. Positively worded items: 1, 3, 4, 7, 10 (specific factor 1); negatively worded items: 2, 5, 6, 8, 9 (specific factor 2).

Table 6

Three-level univariate random-effects models synthesizing the reliability coefficients from the nested-factor model (Example 2)

Approach	Reliability ω_H	95 % Wald CI	τ_2^2	τ_3^2	I_2^2	I_3^2	$Q(33)$
Univariate-*r*	0.740	[0.714, 0.765]	0.005	-	99.1 %	-	4633.5*
Three-level univariate-*r*	0.738	[0.703, 0.773]	0.010	0.000	99.5 %	0.0 %	4633.5*
Raw reliability	0.738	[0.704, 0.773]	0.010	0.000	99.7 %	0.0 %	5889.4*
HW76	0.748	[0.715, 0.779]	0.006	0.000	99.5 %	0.0 %	7076.6*
BO02	0.754	[0.721, 0.782]	0.136	0.000	99.5 %	0.0 %	7425.0*

Note. HW76 = Hakstian-Whalen (1976) approach, BO02 = Bonett's (2002) approach.

* $p < .001$, *ns* = statistically not significant ($p > .05$). Variance estimates and I^2-statistics are reported for the study (level 2) and the study report (level 3) levels.

Figures

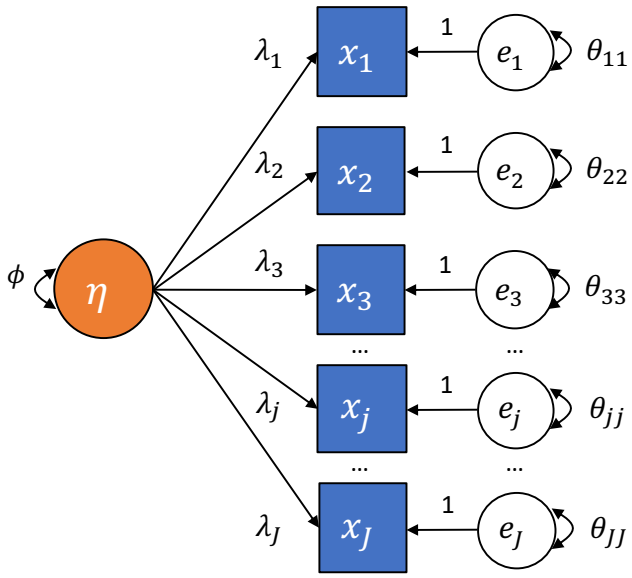

Figure 1. One-factor model for the estimation of the reliability coefficient ω_T.

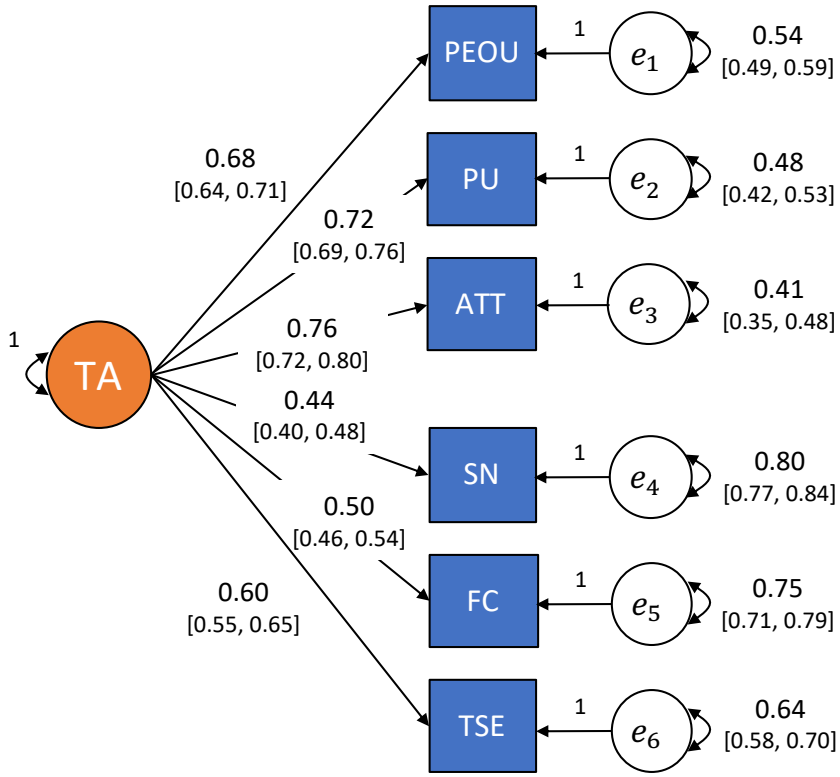

Figure 2. Factor model underlying the estimation of the scale reliability of the technology acceptance (TA) construct based on stage-1 random-effects TSMASEM (Example 1).

Note. PEOU = Perceived ease of use, PU = Perceived usefulness, ATT = Attitudes toward technology, SN = Subjective norms, FC = Facilitating conditions, TSE = Technology self-efficacy. The 95 % Likelihood-based confidence intervals are shown in brackets.

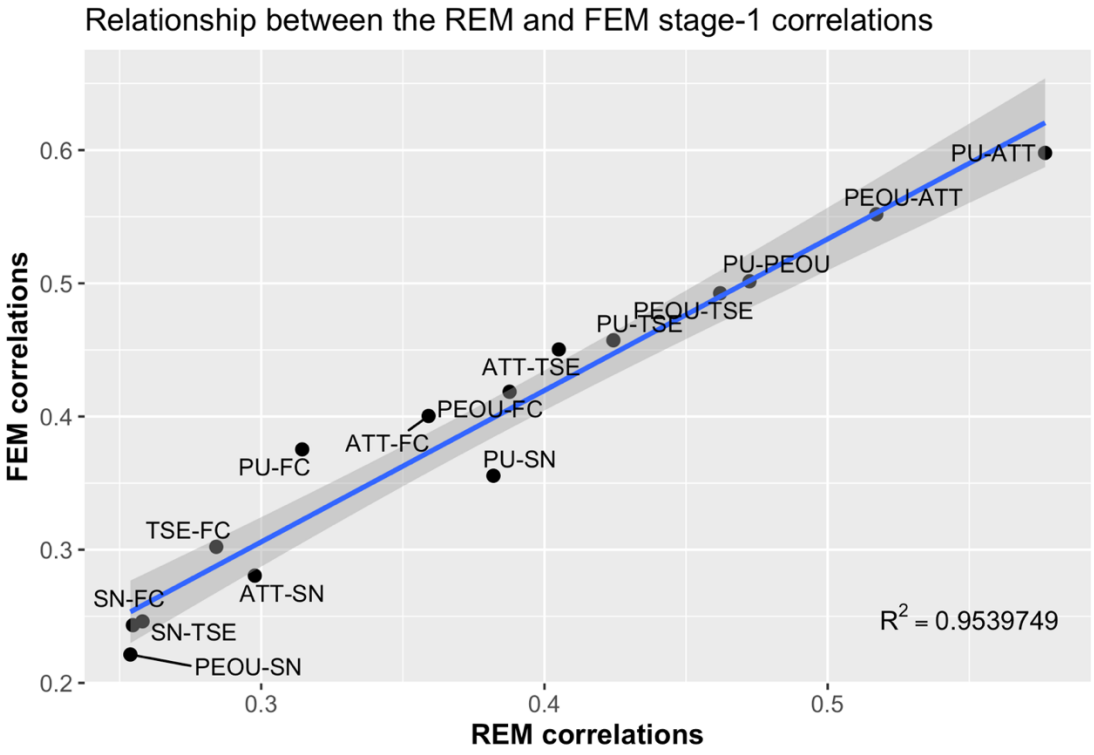

Figure 3. Relations between the TSMASEM stage-1 correlations of the model with fixed (FEM) and random effects (REM) for Example 1.

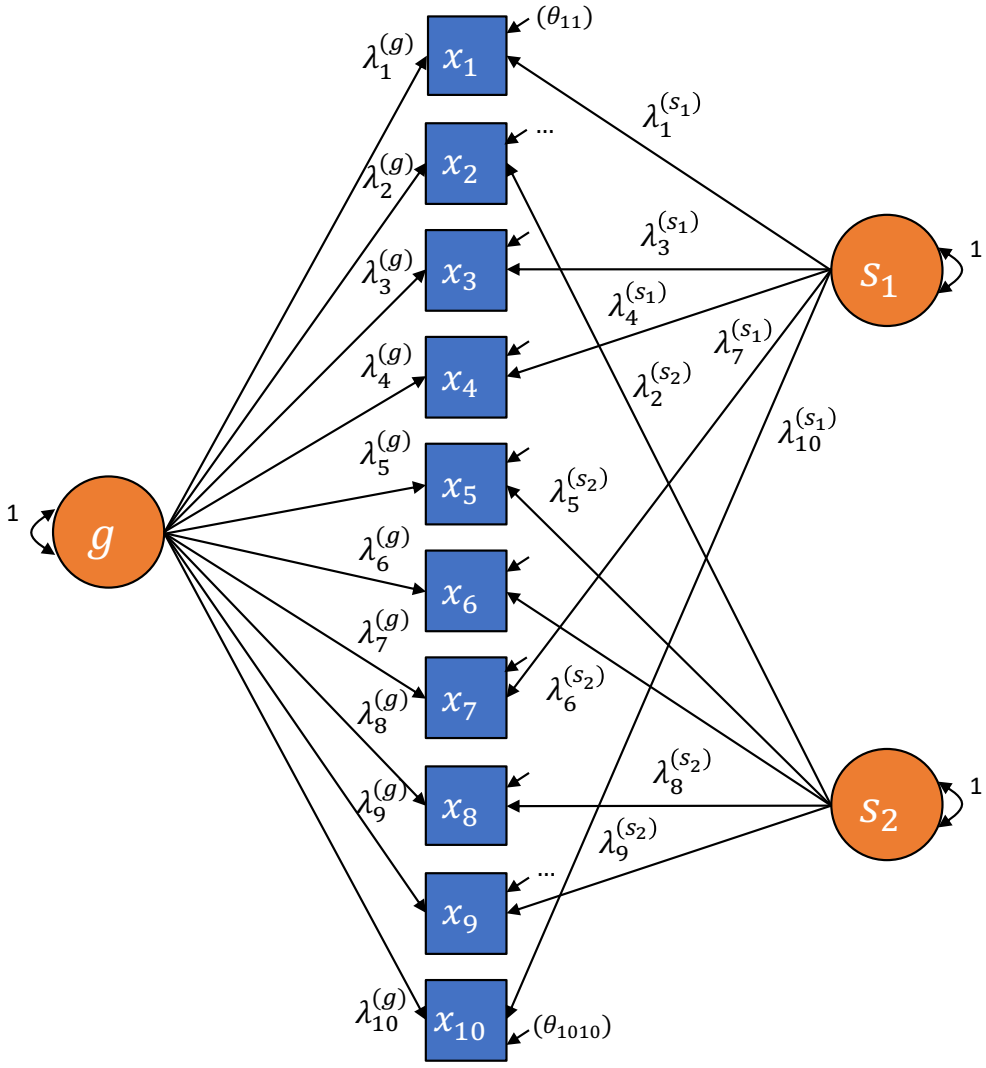

Figure 4. Nested-factor model underlying the Rosenberg Self-Esteem Scale for the estimation of the reliability coefficient ω_H (Example 2).

Note. g = General factor of self-esteem, s_1 = Specific factor representing the positively-worded RSES items, s_2 = Specific factor representing the negatively-worded RSES items. Residuals and their variances are not fully shown.