Journal of Applied Psychology
2016, Vol. 101, No. 10, 1457–1473

0021-9010/16/$12.00 http://dx.doi.org/10.1037/apl0000141

The Problem of Effect Size Heterogeneity in Meta-Analytic Structural Equation Modeling

Jia (Joya) Yu
University of Iowa

Patrick E. Downes
Rutgers University

Kameron M. Carter and Ernest H. O'Boyle
University of Iowa

Scholars increasingly recognize the potential of meta-analytic structural equation modeling (MASEM) as a way to build and test theory (Bergh et al., 2016). Yet, 1 of the greatest challenges facing MASEM researchers is how to incorporate and model meaningful effect size heterogeneity identified in the bivariate meta-analysis into MASEM. Unfortunately, common MASEM approaches in applied psychology (i.e., Viswesvaran & Ones, 1995) fail to account for effect size heterogeneity. This means that MASEM effect sizes, path estimates, and overall fit values may only generalize to a small segment of the population. In this research, we quantify this problem and introduce a set of techniques that retain both the true score relationships and the variability surrounding those relationships in estimating model parameters and fit indices. We report our findings from simulated data as well as from a reanalysis of published MASEM studies. Results demonstrate that both path estimates and overall model fit indices are less representative of the population than existing MASEM research would suggest. We suggest 2 extension MASEM techniques that can be conducted using online software or in R, to quantify the stability of model estimates across the population and allow researchers to better build and test theory.

Keywords: meta-analysis, meta-analytic structural equation modeling, validity generalization

Supplemental materials: http://dx.doi.org/10.1037/apl0000141.supp

A core analytical technique in applied psychology is meta-analysis (Cooper, Hedges, & Valentine, 2009; Schmidt & Hunter, 2015), which quantitatively synthesizes a body of knowledge. In this capacity, meta-analysis has contributed significantly to scientific advancements in applied psychology. A key limitation of meta-analysis, however, is that it traditionally focuses on a single relationship, so it cannot test more complex theoretical models. For this reason, scholars have combined the structural equation modeling (SEM) technique with meta-analysis and referred this procedure as meta-analytic structural equation modeling (MASEM). MASEM has been heralded as offering "the potential to reshape a literature's development" (Bergh et al., 2016, p. 18) because it allows researchers to examine a multivariate model for theory building and testing. Scholars are increasingly using MASEM to raise "fundamentally important questions about the viability of theoretical and conceptual frameworks" (Bergh et al., 2016, p. 4).

Notwithstanding its promise, the current approaches to MASEM do not capitalize on a key strength of meta-analysis, which is the ability to quantify effect size heterogeneity. Effect size heterogeneity refers to the variability of true score estimates in a population (Higgins, 2008). Incorporating effect size heterogeneity is important because conclusions derived solely from the analysis of mean effect sizes are limited and sometimes misleading. For example, a meta-analysis on the effectiveness of a training program, which in fact has no efficacy for women but strong efficacy for men, might conclude a medium average effect size for the general population. In this case, interpretation solely based on average effect sizes may mislead future researchers or practitioners who plan to study or implement the training program with female employees. Meta-analytic techniques address these issues through using either cred-

This article was published Online First August 8, 2016.

Jia (Joya) Yu, Department of Management & Organizations, Henry B. Tippie College of Business, University of Iowa; Patrick E. Downes, Department of Human Resource Management, School of Management and Labor Relations, Rutgers University; Kameron M. Carter and Ernest H. O'Boyle, Department of Management & Organizations, Henry B. Tippie College of Business, University of Iowa.

We are grateful to Mike Cheung for his insight into TSSEM and his feedback to help improve the FIMASEM and TS-FIMASEM procedure. We thank Christopher Berry, Talya Bauer, Mike Christian, Ellen Kossek (and her author team Shaun Pichler, Todd Bodner, and Leslie Hammer), Jennifer Nahrgang, and Jennifer Gillespie for their support in providing their data. We also would like to thank Mo Wang for his constructive comments on the earlier draft of this article. An abridged version of this article won the Best Student Paper Award in the Research Methods Division in the 2015 Academy of Management, Vancouver, British Columbia, Canada, and appeared in the Academy of Management Proceedings in 2015. Jia (Joya) Yu and Patrick E. Downes contributed equally to this article.

Correspondence concerning this article should be addressed to Jia (Joya) Yu, Department of Management & Organizations, Henry B. Tippie College of Business, University of Iowa, W317, 108 John Pappajohn Business B, Iowa City, IA 52242-1994. E-mail: joya.jia.yu@gmail.com

ibility intervals (Hunter & Schmidt, 2004; Whitener, 1990) or *Q* statistics (Hedges & Olkin, 1985). Yet similar procedures for examining effect size heterogeneity in MASEM have not been readily established. As a result, MASEM extends the same problem scholars have observed in the training program example above—only now, they can erroneously conclude the mechanism through which the misleading mean population effect occurs.

Thus, the current practice of MASEM suffers from inconsistency in its assumptions. In published MASEM studies, authors spend the first half of the results identifying and quantifying the effect size heterogeneity in bivariate relations and explaining this heterogeneity using substantive or methodological moderators. But in the second half of the article, where a multivariate SEM is introduced, the technique requires the assumption that the bivariate relationships are drawn from a population with zero effect size heterogeneity. This logical inconsistency needs to be resolved in order for researchers to draw better conclusions from MASEM results, and MASEM procedures would be greatly improved by utilizing the full information provided by meta-analytic results in estimating multivariate models. Conceptually, incorporating effect size heterogeneity into MASEM procedures would allow researchers to quantify the variability in MASEM estimates across a population, leading to more sophisticated interpretations of findings. Incorporating effect size heterogeneity into MASEM analysis would allow an examination of how model estimates vary in a population of studies. Such a technique would recognize that MASEM fit statistics—just like bivariate meta-analytic effect estimates—have true variability in the population, and that the proposed theoretical model might fit in some conditions better than others.

The purpose of this article, then, is to review the issue of effect size heterogeneity in existing MASEM research and offer a quantitative demonstration of how quantifying effect size heterogeneity may change interpretations of MASEM results. We first review the two predominant MASEM techniques: traditional MASEM as proposed by Viswesvaran and Ones (1995) and two-stage structural equation modeling (TSSEM) developed more recently by Cheung (2014, 2015). Because neither the traditional MASEM nor TSSEM has the capacity to quantify heterogeneity as the output of MASEM, we develop specific extensions for both traditional MASEM and TSSEM to quantify effect size heterogeneity. We refer to these extensions as full-information MASEM (FIMASEM), which extends traditional MASEM, and as two-stage FIMASEM (TS-FIMASEM), which extends TSSEM. We first demonstrate the performance of the traditional MASEM and TSSEM, as well as the corresponding FIMASEM procedures in a series of representative simulations in Study 1. In Study 2, we reanalyze 20 published MASEM studies to demonstrate whether substantive conclusions would have differed had the original authors been able to account for and quantify effect size heterogeneity in their analyses. Our two extensions are based on open-source R software, and we make example scripts available in the online supplemental materials. We also introduce online software so that users less familiar with R can conduct FIMASEM using a web-based application. The software allows users to upload primary study data and examine theoretical models using both versions of FIMASEM.

Current Approaches to MASEM

It is helpful to consider all forms of MASEM as consisting of two steps. The first step is to conduct meta-analyses and pool the resulting effect size estimates into a matrix of all the variables in the theoretical model. The second step is to use this pooled matrix as an input into SEM estimation. The two MASEM techniques have fundamental differences worth highlighting before advancing our respective extensions. In describing the divergence between traditional MASEM and TSSEM, we focus on the differences that are most relevant to issues of effect size heterogeneity and the FIMASEM and TS-FIMASEM applications (for a thorough review of the differences between traditional MASEM and TSSEM, we refer readers to Landis, 2013). Table 1 presents the strengths, weaknesses, and applicable contexts of the two MASEM techniques and their extensions.

Traditional MASEM and TSSEM

In traditional MASEM, the researcher first conducts a series of independent meta-analyses on each bivariate relationship. The technique relies on bivariate (or "univariate," in some traditions) meta-analysis (e.g., Hunter & Schmidt, 2004; Cooper et al., 2009), which independently estimates each effect size (ρ) and its heterogeneity (i.e., the standard deviation of the true score estimates, SD_ρ, which is used to construct credibility intervals). Effect size estimates (but not their heterogeneity) are then pooled independently into a correlation matrix. In the second step, this correlation matrix is then entered into SEM as an observed variance-covariance matrix. In TSSEM (Cheung, 2014), effect sizes are pooled using SEM-based meta-analysis, which is based on multivariate (Cheung, 2015) rather than bivariate meta-analysis (Schmidt & Hunter, 2015). In addition to producing a pooled matrix of effect sizes and information about effect size heterogeneity (quantified by τ^2), multivariate SEM-based meta-analysis produces a variance-covariance matrix of effect sizes, which represents the dependence between correlations. Step 2 of TSSEM then weights the pooled effect size matrix by the inverse of its sampling covariance matrix using WLS estimation (for a more in depth discussion of WLS and weighting, see Cheung, 2015).

A notable difference between traditional MASEM and TSSEM is the assumption of independence among effect sizes. Traditional MASEM assumes effect sizes are independent from each other and analyzes each bivariate relationship in separate meta-analysis. This approach has been criticized as ignoring the dependence between effect sizes that might be present (Cheung & Chan, 2005). In contrast, TSSEM pools effect sizes using an SEM-based multivariate approach, which acknowledges and corrects for the biasing of effects due to the potential dependence among effect sizes (Cheung, 2015). For this reason, the pooled matrices from the first step of traditional MASEM and TSSEM are likely to be different, producing differences in results in Step 2.

In addition to their differences in handling the dependence of effect sizes, traditional MASEM and TSSEM handle effect size heterogeneity differently. Although traditional MASEM produces a pooled effect size matrix and effect size heterogeneity estimates in Step 1, only the effect size matrix is used to perform SEM, and the effect size heterogeneity is discarded in Step 2. The second step of traditional MASEM is in essence a fixed effects test (Cheung, 2015), which is not appropriate for drawing random effects con-

Table 1
Overview of Traditional MASEM, TSSEM, FIMASEM, and TS-FIMASEM

MASEM procedures	Brief description of procedures	Strengths	Weaknesses	Applicable research context	Effect size heterogeneity extensions
Viswesvaran & Ones (1995)	Step 1. Conduct bivariate random effects meta-analysis and construct a pooled correlation matrix	Simple to use	Ignores effect size heterogeneity in Step 2	Independent effect sizes	FIMASEM: Step 1. Conduct bivariate random effects meta-analysis and construct pooled correlation and effect size heterogeneity matrices
	Step 2. Use the pooled matrix as the input to fit a structural equation model	Flexible with incomplete data (uses pairwise deletion)	Does not account for the dependence of effect sizes Uses correlation instead of covariance matrix when estimating SEM	No effect size heterogeneity No primary study contains a complete correlation matrix	Step 2. Bootstrap based on pooled matrices and estimate SEM on each input matrix, then calculate summary statistics for model estimates
TSSEM (Cheung & Chan, 2005; Cheung, 2014)	Step 1. Conduct a multivariate meta-analysis	Accounts for dependence of effect sizes	Does not quantify effect size heterogeneity in step two	Dependent effect sizes	TS-FIMASEM: Step 1. Conduct a multivariate meta-analysis and construct pooled correlation and effect size heterogeneity matrices
	Step 2. Calculate pooled correlation matrix (weighted by the asymptotic covariance matrix) as the input to fit a structural equation model	Controls for effect size heterogeneity	Stricter requirement on missing data (one primary studies must have a complete correlation matrix)	Primary studies are replications of an existing theoretical model and report all effect sizes for the same set of variables	Step 2. Bootstrap based on pooled matrices and estimate SEM on each input matrix, then calculate summary statistics for model estimates

Note. MASEM = meta-analytic structural equation modeling; TSSEM = two-stage structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling; TS-FIMASEM = two-stage full-information structural equation modeling; SEM = structural equation modeling.

clusions (i.e., unconditional inferences; Hedges & Vevea, 1998, p. 488). TSSEM, in contrast, is capable of a random effects approach, which derives less biased estimates by controlling for the confounding impact of effect size heterogeneity. Unlike traditional MASEM that ignores effect size heterogeneity in SEM, TSSEM considers effect size heterogeneity as statistical noise to be removed in Step 2. Simulation results (Cheung, 2014; Cheung & Chan, 2005; Furlow & Beretvas, 2005) have demonstrated that TSSEM indeed offers less biased statistical tests than does traditional MASEM.

Despite their fundamental differences, traditional MASEM and TSSEM are similar in that they emphasize the statistical significance of SEM estimates and do not quantify the impact of effect size heterogeneity, if present in Step 1, on the final output of the Step 2 model. Even if a robust statistical test shows that model estimates and confidence intervals are statistically significant, it is also important to know how those estimates are distributed in the population. For example, if a robust statistical test fails to reject the null hypothesis, it could be due to a very small population effect, or it could be because the effect is strongly positive for some subpopulations or boundary conditions, and strongly negative for others. In this case, the statistical significance test offers no information about why the effect is or is not significant. In short, the best fitting model may not always be the most generalizable model. What is missing in the current MASEM practice is an approach that determines how model statistics are distributed in the population given information about effect size heterogeneity produced in the random effects meta-analysis of Step 1.

Proposed Extensions: FIMASEM and TS-FIMASEM

As an initial attempt to incorporate effect size heterogeneity into MASEM, we propose FIMASEM and TS-FIMASEM, which seek to determine the distribution of estimates in a multivariate theoretical model given the presence of effect size heterogeneity. To do this, FIMASEM and TS-FIMASEM randomly sample values from the distributions (defined by the mean and standard deviation) of bivariate effect sizes derived from the Step 1 meta-analysis. These values are pooled into multiple effect size matrices. FIMASEM and TS-FIMASEM then iteratively estimate the SEM across each constructed matrix, allowing researchers to summarize the parameter distributions of model estimates. In either FIMASEM or TS-FIMASEM, the first step is to construct two input matrices based on meta-analyses: one representing effect sizes (i.e., ρ), and one representing effect size heterogeneity (SD_ρ or τ, depending on the meta-analytic tradition). In FIMASEM, the effects can be pooled using any bivariate meta-analysis technique that estimates effect size heterogeneity (SD_ρ and τ vary computationally but are conceptually similar). In TS-FIMASEM, effects are pooled using SEM-based multivariate meta-analysis (e.g., Cheung & Chan, 2005; Cheung, 2013), which takes into account potential dependence between effect sizes.

Once effect sizes and their heterogeneity estimates have been pooled, the second step is to use those pooled matrices as the basis to construct a random sample of matrices. This is accomplished in FIMASEM by bootstrapping matrices based on each effect size (ρ) and its heterogeneity (SD_ρ). In TS-FIMASEM, the pooled matrices and τ^2 values are extracted from the first step of TSSEM and used as the basis for the bootstrap. Because either procedure could produce non-positive-definite matrices, and because commonly used algorithms for SEM (e.g., maximum likelihood estimation) produce biased estimates for nonpositive definite matrices, nonpositive definite bootstrapped matrices are resampled until all matrices are positive definite.

After the matrices have been generated, the SEM is estimated for each bootstrapped matrix. This step results in a distribution for each estimate (e.g., path coefficient, fit index) in the model. With respect to path coefficients, we suggest summarizing the distribution by building a credibility interval of each path coefficient (credibility interval [CV_β]). Similar to a CV in bivariate meta-analysis, CV_β represents the range within which a percent (below we suggest constructing 80% CV_β) of the population parameters fall. Wider CV_β widths indicate more variability in a model estimate (Whitener, 1990). With respect to the distribution of fit indices across SEMs based on the bootstrapped matrices, we suggest quantifying the extent to which a fit index generalizes to a population of studies at conventional levels (e.g., McDonald & Ho, 2002). Although conventions for fit are controversial, they offer a succinct way to summarize the population fit indices. Our suggestion is to quantify the percent of bootstrapped input matrices that have acceptable model fit by conventional standards for a given fit index. For example, standardized root mean square residual (SRMR) generalizability could be expressed as the percent of bootstrapped matrices that have an SRMR less than .10.

Study 1: Simulation

Method

The purpose of Study 1 was to use simulated data to compare the conclusions of traditional MASEM (Viswesvaran & Ones, 1995) and TSSEM (Cheung & Chan, 2005) with our proposed extensions, FIMASEM and TS-FIMASEM, respectively. We chose two mediation structural models with single indicators, as mediation testing is a common use of MASEM in organizational research (Bergh et al., 2016). For each model, we generated a sample of primary studies corresponding to different conditions of effect size heterogeneity in the data. We used Hunter and Schmidt's (2004) bare bones psychometric random-effect meta-analysis as the Step 1 for traditional MASEM and FIMASEM. For TS-FIMASEM, we entered the primary study data into TSSEM directly and then ran TS-FIMASEM based on the Step 1 output of TSSEM.

For each simulation we generated an observed set of primary studies ($k = 50$) from a distribution where the corresponding population correlations for specified paths in Figure 1 (i.e., ρ_{ij}) had a mean of .30 and nonmodeled paths had mean population correlations of .00. We then systematically varied SD_ρ for each population correlation to create effect size heterogeneity in different magnitudes and between different constructs in the model. In Simulation 1, we tested the model without any effect size heterogeneity ($SD_\rho = .00$). In Simulation 2, we examined the presence of effect size heterogeneity ($SD_\rho = .10$) in 50% (randomly selected) of the correlations. In Simulation 3, we introduced effect size heterogeneity ($SD_\rho = .10$) into each of the correlations in the model. In Simulations 4 and 5, we increased the magnitude of effect size heterogeneity (Simulation 4: $SD_\rho = .20$; Simulation 5: $SD_\rho = .30$). Finally, in Simulation 6, we analyzed a more complex model (see Figure 2) to see how the introduction of effect size

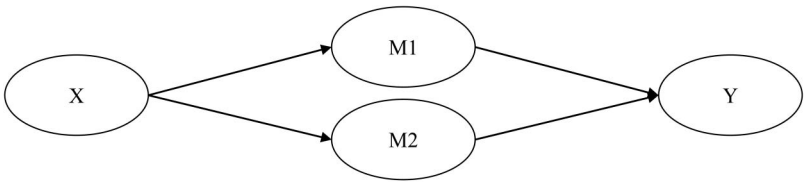


Figure 1. Study 1: Simulation 1 model. All latent variables have one indicator.

heterogeneity influences a larger model with more constructs and paths.

We bootstrapped 500 matrices in each simulation, which offered a balance of precision and computer processing power. We fixed the sample size of each primary study to 200 in each simulation. Although the sample size is arbitrary in the simulation, it does not introduce first order sampling error in this study as it would in a typical simulation. Changes in sample size would only affect the standard errors and statistical significance of the parameter estimates calculated in each simulation, which are not an emphasis of the current simulations. Put differently, the credibility intervals and generalizability estimates produced by FIMASEM and TS-FIMASEM are not dependent upon sample size.

We quantify the impact of effect size heterogeneity in two ways: by constructing CV_β for each path coefficient and computing the generalizability of the fit using χ^2 statistics, comparative fit index (CFI), and SRMR. We compute 80% CV_β intervals based on the standard deviation of the distribution of β. To offer some conventions for the width of CV_β, we interpret the difference between the upper and lower bounds of 80% CV_β (i.e., CV_β width) as a mean difference, or *d* score (Cohen, 1992). A recent empirical investigation of Cohen's (1992) conventions suggests *d* score conventions of .18 and .54 to divide the lower, middle, and upper thirds of effect sizes across the field (Bosco, Aguinis, Singh, Field, & Pierce, 2015). Thus, we suggest that CV_β width less than .18 represents small heterogeneity, a width between .18 and .54 represents medium heterogeneity, and an interval wider than .54 represents large heterogeneity.

Many fit indices could be used to calculate the generalizability of the model as a whole, and in the current research we present information on χ^2 statistics, CFI, and SRMR. Although we focus on the three fit indices, the online software outputs 16 fit indices, and the logic we propose can be applied to any fit index. We compute the percent of bootstrapped input matrices that fit at conventional levels (i.e., have an SRMR less than .10 or CFI larger than .90, McDonald & Ho, 2002). These values describe the percent of studies to which the model generalizes accounting for the effect size heterogeneity observed in random effects meta-analysis.

We technically accomplished these procedures using R. There are currently two major R packages for conducting SEM: OpenMx and lavaan. Because the two routines have slight differences in model estimation, we incorporate both lavaan and OpenMx into FIMASEM. However, TS-FIMASEM relies on Cheung's (2015) metaSEM package and must use OpenMx estimation to remain consistent with TSSEM. Regardless of which R package is implemented for SEM analysis, the estimation of all models results in a distribution of all model estimates. We provide syntax to reproduce these simulations in the supplemental materials and online software.

One important consideration about this set of simulations is that they are comparable between MASEM techniques within a simulation. However, because the simulations each generate their own set of data, and because each primary study must have a positive definite matrix, there are slight differences between the mean and standard deviation of datasets being analyzed between simulations. We take care to draw conclusions only on the comparisons between techniques within a simulation because those techniques analyze the same data.

Results

Simulation 1: Figure 1 model with no effect size heterogeneity. In Simulation 1, we assume that the effect size homogeneity assumption holds (i.e., all $SD_\rho = .00$). We generated the primary data such that zero effect size heterogeneity was present in the primary studies. We performed this simulation using a simple model presented in Figure 1 with one independent variable, two mediators and one dependent variable. As seen in Table 2, FIMASEM and TS-FIMASEM converge on identical solutions to traditional MASEM and TSSEM when there is no variability in effect sizes. Specifically, FIMASEM and TS-FIMSEM report

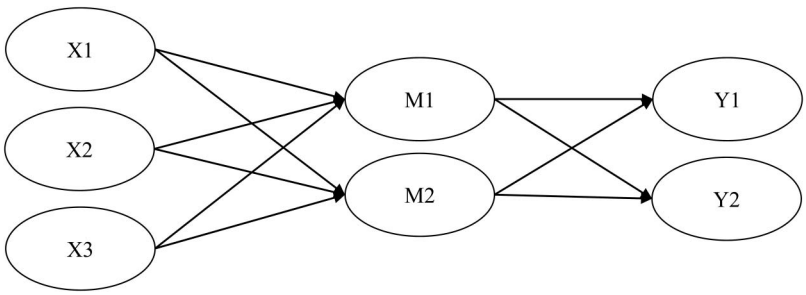


Figure 2. Study 1: Simulation 2 model. All latent variables have one indicator.

Table 2
Study 1 Results From Six Simulations

Run	Description	Traditional MASEM				FIMASEM				
		Avg. β[a]	χ^2	CFI	SRMR	Avg. $\bar{\beta}$[a]	Avg. 80% CV_β[a]	$\bar{\chi}^2$	% CFI > .90	% SRMR < .10
1	Figure 1 model: All SD_ρ = .00	.29	10.06	.92	.09	.29	[.29, .29]	10.06	100%	100%
2	Figure 1 model: 50% SD_ρ = .10	.29	15.84	.87	.09	.29	[.22, .35]	9.17	25%	54%
3	Figure 1 model: 100% SD_ρ = .10	.27	13.20	.90	.11	.28	[.16, .41]	30.84	39%	40%
4	Figure 1 model: 100% SD_ρ = .20	.25	10.06	.92	.14	.26	[.03, .47]	109.58	30%	33%
5	Figure 1 model: 100% SD_ρ = .30	.23	6.07	.93	.16	.23	[−.04, .49]	163.19	24%	25%
6	Figure 2 model: 100% SD_ρ = .10	.27	9.75	.64	.11	.21	[.15, .38]	75.29	1%	9%
		TSSEM				TS-FIMASEM				
1	Figure 1 model: All SD_ρ = .00	.29	20.00	.96	.08	.29	[.29, .29]	20.00	100%	100%
2	Figure 1 model: 50% SD_ρ = .10	.28	4,937.31	.92	.08	.30	[.24, .35]	1,018.07	0%	56%
3	Figure 1 model: 100% SD_ρ = .10	.28	1,650.44	.95	.06	.30	[.19, .40]	1,537.06	20%	40%
4	Figure 1 model: 100% SD_ρ = .20	.26	470.08	.96	.05	.26	[.07, .45]	3,763.40	16%	46%
5	Figure 1 model: 100% SD_ρ = .30	.23	275.23	.98	.04	.21	[−.07, .49]	4,994.48	8%	20%
6	Figure 2 model: 100% SD_ρ = .10	.21	655.61	.85	.12	.30	[.23, .36]	3,153.76	0%	0%

Note. MASEM = meta-analytic structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling; Avg. β = average path coefficient for traditional MASEM; χ^2 = chi-square statistic for model in traditional MASEM; CFI = comparative fit index for traditional MASEM; SRMR = standardized root mean square residual for traditional MASEM; $\bar{\beta}$ = mean path coefficient across 500 iterations; CV = credibility interval; Avg. 80% CV_β = average 80% credibility interval for the path coefficient across 500 iterations; $\bar{\chi}^2$ = mean value of χ^2 across 500 iterations; % CFI > .90 = percentage of CFI statistics that fell above the .90 cutoff across 500 iterations; % SRMR < .10 = percentage of SRMR statistics that fell below the .10 cutoff across 500 iterations; TS-FIMASEM = two-stage full-information structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling.
[a] Statistics represent average of all paths in model.

zero-width CV_β around β and that the model fits for 100% of the population, converging with the conclusions of traditional MASEM and TSSEM.

Simulation 2: Figure 1 model with effect size heterogeneity (SD_ρ = .10) in half of paths. In Simulation 2, we generated the primary study data such that effect size heterogeneity of SD_ρ = .10 was introduced to 50% (randomly selected) of the correlations between model variables. Specifically, $\rho_{X\bullet M1}$, and $\rho_{X\bullet M2}$ had effect size heterogeneity of .10. All remaining correlations had no effect size heterogeneity (SD_ρ = .00). Results from Simulation 2 are presented in Condition 2 of Table 2. The average standardized path coefficients ($\bar{\beta}$) from FIMASEM and TS-FIMASEM were nearly identical to those from traditional MASEM and TSSEM, respectively. As expected, FIMASEM and TS-FIMASEM reported a nonzero-width CV_β in the four paths of the model (FIMASEM average 80% CV_β [.22, .35]; TS-FIMASEM average 80% CV_β [.24, .35]). FIMASEM and TS-FIMASEM also demonstrated that the model fit, by conventional levels, for less than 100% of the population (i.e., for SRMR, FIMASEM: 54%, TS-FIMASEM: 56%). This means that when even half the paths are allowed to vary, approximately half of the population was not acceptably represented by the theoretical model, according to standard conventions for fit. This point is not revealed in traditional MASEM, which implies that the estimated effects and their respective statistical significance are indicative of a 100% generalizable effect.

Simulation 3: Figure 1 model with effect size heterogeneity (SD_ρ = .10) in all paths. In Simulation 3, we allowed 100% of the true correlations between X, Y, M1, and M2, to vary with a standard deviation of .10 in the Figure 1 model. Table 2 displays these results. Once again, FIMASEM and TS-FIMASEM produced CV_β (FIMASEM average 80% CV_β [.16, .41]; TS-FIMASEM average 80% CV_β [.19, .40]) that reflect the variability of correlations from meta-analytic results. In terms of generalizability of model fit, the extensions estimate that the model fits less than half of the population (i.e., for SRMR, FIMASEM: 40%, TS-FIMASEM: 40%).

Simulation 4: Figure 1 model with effect size heterogeneity (SD_ρ = .20) in all paths. The purpose of Simulation 4 was to examine how the magnitude of effect size heterogeneity impacts MASEM conclusions. We generated a sample of primary studies with SD_ρ of .20 for all effect sizes and ran the analysis on the Figure 1 model. The results in Table 2 indicate medium-width CV_β for both FIMASEM (average 80% CV_β [.03, .47]) and TS-FIMASEM (average 80% CV_β [.07, .45]). In terms of the generalizability of model fit (i.e., SRMR), the extensions report that the model fits 33% (FIMASEM) and 46% (TS-FIMASEM) of the population.

Simulation 5: Figure 1 model with effect size heterogeneity (SD_ρ = .30) in all paths. In Simulation 5, we again increased the effect size heterogeneity for all effect sizes up to SD_ρ of .30. Table 2 reports that this produced large-width CV_β on average for both FIMASEM (average 80% CV_β [–.04, .49]) and TS-FIMASEM (average 80% CV_β [–.07, .49]). Note that the conclusions of traditional MASEM and TSSEM differed slightly from the conclusions of FIMASEM and TS-FIMASEM in this simulation. Whereas traditional MASEM and TSSEM would conclude a positive, statistically significant, relationship holds along paths in the model, FIMASEM and TS-FIMASEM contend that for some

subpopulations the effect is null or even negative. This conclusion is also reflected in results for the generalizability of model fit, as only about one in four studies in the population is represented by this theoretical model (i.e., SRMR, FIMASEM: 25%, TS-FIMASEM: 20%). Traditional MASEM or TS-FIMASEM does not reveal this information in their respective calculations of fit statistics.

Simulation 6: Figure 2 model with effect size heterogeneity (SD_ρ = .10) in all paths. The purpose of Simulation 6 was to examine the impact of effect variability in a more complex model. To test a more realistic model, we added two more independent variables as well as a second dependent variable as part of Simulation 6 as displayed in Figure 2. In constructing these data we again created a sample of primary studies with effects randomly drawn from a corresponding distribution with the effect size (.30 if the path was specified, and .00 if the path is not specified in the model) and effect size heterogeneity of .10 for all paths. Results from Simulation 6 suggested a similar result in terms of CV_β as in Simulation 3. Specifically, the extensions reported small-width CV_β (FIMASEM: average 80% CV_β [.15, .38], TS-FIMASEM: average 80% CV_β [.23, .36]). More notably, the generalizability of these models was very small, with less than 10% of the population of studies demonstrating acceptable fit (i.e., SRMR, FIMASEM: 9%, TS-FIMASEM: 0%).

Discussion

These simulations, covering a broad range of conditions, illustrate the problem of assuming a single population effect size (i.e., no effect size heterogeneity) when, in fact, the effect size varies in a population. Specifically, traditional MASEM (Viswesvaran & Ones, 1995) and TSSEM (Cheung, 2015; Cheung & Chan, 2005) do not quantify the impact of effect size heterogeneity, yet offer the same conclusions for Simulations 1 through 6: the theoretical model generalizes across the entire population. However, FIMASEM and TS-FIMASEM demonstrate that this conclusion is sensitive to effect size heterogeneity. In Simulation 1, where effect size heterogeneity is absent, the extensions confirm that the model fits the entire population. However, in Simulation 2, where we introduce effect size heterogeneity in only half the paths, the model fits about half the population. When we introduce larger amounts of effect size heterogeneity into all the paths, the divergence between the existing techniques and our proposed extension is more substantial—in Simulation 5, the model fits about a quarter of the population. In this case, there probably remain subpopulations and moderators where a different model may be more appropriate. Given this conclusion, research can move to examine what those contingencies might be.

We recognize, however, that the Study 1 simulation offers an overly simple view of MASEM. Our assumptions, models, and conditions were contrived to demonstrate the value of our technique, and thus may not be applicable to real world data MASEM scholars are likely to observe. To overcome this limitation, we conducted a second study with the goal of reanalyzing published MASEM research and applying our procedures to the original authors' data. This allows a more comprehensive demonstration of a wider variety of theoretical models. Examining real world data also offers a clearer view of the potential value that our extensions could bring to the current MASEM practice.

Study 2: Reanalysis of Published MASEM Studies

Method

Literature search. To identify articles for inclusion, we completed searches in Google Scholar for articles containing the keywords *meta-analytic structural equation modeling* and *MASEM*, as well as articles having cited Viswesvaran and Ones (1995) or Cheung and Chan (2005). We limited our search to articles published within top journals as these provided the highest-quality exemplars of MASEM in leading management journals. We searched *Academy of Management Journal, Journal of Applied Psychology, Personnel Psychology, Journal of Management, Strategic Management Journal, Organization Science, Administrative Science Quarterly, Journal of Organizational Behavior, and Organizational Behavior and Human Decision Processes*. Initial search results provided 73 potential articles from the nine journals.

Inclusion criteria. To be included, meta-analyses had to meet the following criteria. First, articles had to have actually performed MASEM using a single indicator approach. Review articles, qualitative research, primary studies, and articles involving confirmatory factor analyses and metaregression were excluded. Twenty-two studies were excluded based on this criterion. Second, included MASEMs had to provide enough information to recreate their effect size (i.e., correlation) and standard deviation matrices as well as their structural equation models. Thus, articles had to provide correlations (r or ρ) between study variables as well as variance information (e.g., standard deviations, variance, and/or credibility intervals) for corresponding correlations. Additionally, articles had to include the specifications (i.e., causal paths) of their meta-analytic structural-equation models, which were generally expressed as figures. Where articles did not include all relevant information but were clearly examples of MASEM, we emailed authors of the published studies requesting the necessary information. This criterion excluded another 26 studies.

Because all selected MASEM studies were based on various SEM programs, we only included studies for which we could replicate meta-analytic results using the reconstructed matrices in the R SEM package lavaan (Rosseel, 2012). Further, replicating the original MASEM ensured that we appropriately specified the model exactly as the original authors had in their published work. We then compared our replicated results with original results in terms of path coefficients, degrees of freedom, χ^2, CFI, and SRMR. Where full or nearly full replication of path estimates was found,[1] these studies were included within our final sample. This inclusion criterion excluded five studies, leaving a final sample of 20 meta-analytic structural equation models that could be replicated using FIMASEM. All 20 articles used traditional MASEM (Viswesvaran & Ones, 1995) in their primary analysis.

Coding. The first and third authors coded the data from each study. The authors independently coded the same five articles to assure accuracy and interrater agreement. The agreement rate was high (Cohen's κ = .96; Cohen, 1960); any discrepancies were resolved through discussion. From each article we coded independent and dependent variables, k, n, ρ, and SD_ρ. When SD_ρ was not

[1] Replication results are available upon request.

reported we used the lower and upper bounds of the credibility intervals to reverse-compute it.

Because TSSEM and TS-FIMASEM require coding sheet data from the original primary studies, we obtained these values from meta-analytic results tables, which provide effect sizes from the primary studies in the meta-analysis. When these data were not available, we emailed the authors of the published MASEM studies. Through this procedure, we retrieved raw coding sheets of six included MASEM studies. Only two out of the six studies had at least one primary study with a complete effect size matrix of all study variables (a necessary condition to run TSSEM and TS-FIMASEM, Landis, 2013). As a result, we reanalyzed using Viswesvaran and Ones (1995) and FIMASEM procedures on all 20 included MASEM studies and performed TSSEM and TS-FIMASEM on two included MASEM studies that had sufficient data.

Analysis plan. We first constructed the ρ and SD_ρ matrices from 20 MASEM studies using traditional MASEM pooling for FIMASEM. Some studies provided SD_ρ only for some correlations in the meta-analysis. In these instances we filled missing data with zeros in the SD_ρ matrices (13 of the 20 studies had at least one missing SD_ρ). This decision assumes no effect size heterogeneity for unreported relationships, making the contrast between FIMASEM and traditional MASEM more conservative. The ρ and SD_ρ matrices were then uploaded as inputs for our replication and subsequent FIMASEM analyses to the website we developed for the purpose of this analysis (see supplemental materials for implementation of FIMASEM and access to the online software). For each reanalysis we specified 500 FIMASEM iterations, and we entered the sample size reported by the authors of the published studies.

Results

We organize the following results into three parts. First, we present a detailed example of a specific model from one MASEM study to illustrate the application and interpretation of FIMASEM with real data. Second, to give a broad sense of how MASEM conclusions differ when effect size heterogeneity is incorporated, we report the aggregated FIMASEM and traditional MASEM results of path estimates and fit indices from all 20 MASEM studies. Finally, we present results of TS-FIMASEM and TSSEM on the two MASEM studies that provide sufficient data to perform both TSSEM and TS-FIMASEM to facilitate comparison across both our proposed extensions.

Example FIMASEM. The model tested in published MASEM Study 14 is presented in Figure 3. It has four independent variables, one mediator, and one dependent variable. The original authors used the Viswesvaran and Ones (1995) procedure and reported using a sample size of 216 in SEM. We specified the seven structural paths and ran the model using both Viswesvaran & Ones' procedure and FIMASEM. The results are presented in Table 3.

Mean path coefficients from FIMASEM are very similar to Viswesvaran and Ones estimates. However, the divergence between traditional MASEM and the FIMASEM extension is illustrated in the 80% CV_β as reported by FIMASEM. Viswesvaran and Ones' (1995) procedure assumes that effect size heterogeneity, and thus true score variance in path coefficients, is zero in the SEM phase. This assumption is not made in FIMASEM and, as the results in Table 3 show, path coefficients vary substantially. The most narrow 80% CV_β, union instrumentality perceptions to prounion attitudes (INST→PA) is .38 (80% CV_β [.43, .81]). Although the distribution is consistently positive, the effect does vary substantially across the population. At the upper bound of the CV, a 1-*SD* change in union instrumentality perceptions is associated with almost two times larger of a change in prounion attitudes than it would be at the lower bound.

The results are more extreme for other paths. Five of the seven paths display large amounts of effect size heterogeneity (i.e., CV_β width $>$ .54) and the 80% CV_β for three paths include zero, indicating that not even the sign of the coefficient (i.e., positive or negative) between the two variables generalizes across the population. For example, job satisfaction's effect on union commitment ranges from a negative effect at the lower bound of the CV (–.43)

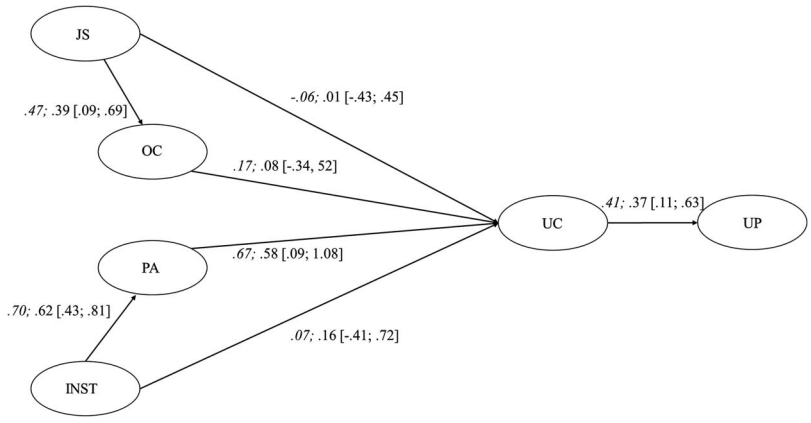
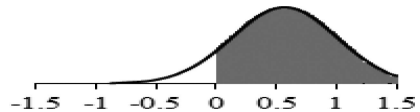
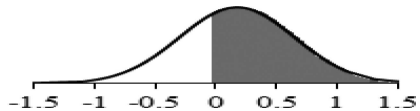
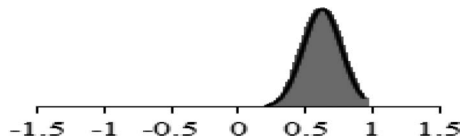


Figure 3. Traditional meta-analytic structural equation modeling (MASEM) and full-information MASEM (FIMASEM) results from Study 14. Single indicators, error terms, β, 80% CV_β limits, and residual variances are omitted for clarity. Numbers in italics represent traditional MASEM results followed by FIMASEM results. UP = union participation; UC = union commitment; OC = organizational commitment; JS = job satisfaction; PA = prounion attitudes; INST = union instrumentality perception.

Table 3
Example Reanalysis: Parameter Estimates From Study 14

	Traditional MASEM	FIMASEM				
Paths	β	$\bar{\beta}$	80% CV_β	80% CV_β width	% β above 0	β distribution
UC→UP	.41**	.37	[.11, .63]	.51	97%	
OC→UC	.17**	.08	[−.34, .52]	.85	61%	
JS→UC	–.06	.01	[−.43, .45]	.88	51%	
PA→UC	.67**	.58	[.09, 1.08]	.99	93%	
INST→UC	.07	.16	[–.41, .72]	1.13	64%	
JS→OC	.47**	.39	[.09, .69]	.59	95%	
INST→PA	.70**	.62	[.43, .81]	.38	100%	

Note. $\bar{\beta}$ = mean level of coefficients for a path across 500 iterations; MASEM = meta-analytic structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling; 80% CV_β = 80% credibility interval for the path coefficient; UP = union participation; UC = union commitment; OC = organizational commitment; JS = job satisfaction; PA = prounion attitudes; INST = union instrumentality perceptions.
** $p < .01$.

to a positive effect at the upper bound (.45). With a CV_β width of .88, FIMASEM finds wide intervals that signify that the effects vary, which could be due to boundary conditions or subpopulations.

It is critical to note that Viswesvaran and Ones' (1995) procedure does not reveal whether the path varied across subpopulations or boundary conditions. Viswesvaran and Ones' (1995) procedure reported that prounion attitudes was positively and significantly related to union commitment (i.e., $\beta = .67$, $p < .01$). The FIMASEM procedure, however, suggests that this conclusion is oversimplified. The path is positive for certain subpopulations and boundary conditions, but there are other cases where the effect is null, negative, or possibly uninterpretable (if the path represents the same construct). By accounting for the variability of the bivariate results in multivariate SEM, FIMASEM can offer a more rigorous test of the theoretical model than can Viswesvaran and Ones' (1995) procedure.

This conclusion is also reflected in comparing the distribution of fit indices from FIMASEM with the results of Viswesvaran and Ones' (1995) procedure. Viswesvaran and Ones' (1995) procedure reported that the model had acceptable fit (SRMR = .05; CFI = .98). However, FIMASEM results suggest that achieving this level of fit is rare in the population of data described by the meta-analytic results. Specifically, SRMR was less than .10 for only 21% of models generated from the distribution of effect sizes. Results with respect to CFI were even worse. Only 2% of bootstrapped models had CFI values greater than .90. These findings again highlight the value of FIMASEM; although the Viswesvaran and Ones (1995) procedure suggested the model fit the population acceptably, FIMASEM found this was rarely the case. For the remaining 79% (or 98%, interpreting CFI), there likely remain other models that better represent the nature of relationships between these constructs.

Results for FIMASEM. FIMASEM results are interpreted with two foci: the distribution of path estimates and the distribution of fit indices.

Path estimates. Table 4 presents summary statistics of original meta-analytic results, traditional MASEM path estimates and the FIMASEM results from all included MASEM studies. The left side of Table 4 presents the number of paths specified in the SEM, whether the authors concluded significant effect size heterogeneity (and thus conducted a moderator analysis) based on meta-analytic results, and summary statistics of CV_ρ widths across all correlations corresponding to specified SEM paths. For example, the SEM of Study 1 had 20 paths, and the original authors concluded significant effect size heterogeneity existed for at least some of these paths. CV_ρ widths of these paths ranged from .13 to .59 with a mean width of .32. Of the 20 MASEM studies, 16 studies performed moderator analysis for at least one bivariate relationship

Table 4
Summary Statistics From Original and Meta-Analytic Results and MASEM Results of Included Studies

	Original meta-analytic results from published articles					Traditional MASEM	FIMASEM path coefficients				
Study	# of Paths	Moder.	Min. CV_ρ width[a]	Avg. CV_ρ width[a]	Max. CV_ρ width[a]	Avg. β[b]	Avg. ($\bar{\beta}$)	$\overline{SD_\beta}$	Min. CV_β width	Avg. CV_β width	Max. CV_β width
1	20	Yes	.13	.32	.59	.15	.15	.21	.28	.54	.93
2	2	Yes	.33	.38	.44	.26	.26	.19	.41	.48	.55
3	12	No	.08	.32	.61	.08	−.16	.20	.13	.49	.77
4	11	Yes	.15	.21	.41	.21	.21	.14	.07	.35	.60
5	8	Yes	.21	.44	.72	.18	.18	.21	.21	.54	.84
6	26	Yes	.11	.42	.65	.11	.04	.38	.29	.98	1.31
7	8	Yes	.58	.63	.72	.30	.30	.18	.03	.46	.79
8	6	Yes	.28	.37	.46	−.07	−.06	.10	.00	.25	.77
9	7	Yes	.25	.33	.40	.21	.20	.21	.28	.55	1.03
10	11	Yes	.41	.54	.67	.30	.29	.18	.00	.45	.89
11	10	Yes	.36	.40	.44	.05	.05	.07	.00	.20	.88
12	8	No	.25	.30	.33	.25	.25	1.16	.34	2.97	7.56
13	8	No	.21	.37	.46	.08	.08	.41	.84	1.04	1.30
14	7	Yes	.41	.58	.76	.35	.32	.30	.38	.76	1.13
15	9	Yes	.23	.43	1.08	−.09	−.07	.29	.28	.73	1.23
16[c]	5	Yes	.20	.20	.20	.56	.56	.02	.00	.05	.19
17	6	No	.18	.25	.30	.31	.32	.11	.22	.29	.34
18	8	Yes	.27	.43	.54	.23	.23	.09	.00	.23	.53
19	4	Yes	.21	.32	.46	.12	.12	.25	.25	.63	1.00
20	23	Yes	.00	.19	.33	.04	.04	.08	.01	.19	.46
Avg.	10	—	.24	.37	.53	.18	.17	.24	.20	.61	1.16

Note. MASEM = meta-analytic structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling; Moder. = an indication of whether the authors conducted moderation analysis in the meta-analysis (Stage 1); CV = credibility interval; Min. CV_ρ width = the minimum credibility interval width across all paths in the study; Avg. CV_ρ width = the average credibility interval width across all paths in the study; Max. CV_ρ width = the maximum credibility interval width across all paths in the study; Avg. β = average path coefficient for traditional two-stage structural equation modeling; $\bar{\beta}$ = mean path coefficient across 500 iterations; $\overline{SD_\beta}$ = the average standard deviation across 500 iterations; Min. CV_ρ width = the minimum credibility interval width across 500 iterations; Avg. CV_ρ width = the average credibility interval width across 500 iterations; Max. CV_ρ width = the maximum credibility interval width across 500 iterations.

[a] Distribution of CV_ρ is based only on the number of ρ related to all structural paths of a model. [b] Average betas across all paths with traditional Viswesvaran & Ones (1995) method.

based on meta-analytic results. However, none of these studies accounted for this effect size heterogeneity in estimating the SEM.

The middle column of Table 4 presents the average path coefficient (β) across all paths of the traditional MASEM, and the right side of Table 4 presents the distributions of path coefficients from FIMASEM across all paths in the model. For example, Study 1 reported a SEM with 20 structural paths, and the middle column presents the average path coefficient across those 20 paths based on traditional MASEM. FIMASEM then presents the average distribution across these 20 paths (notated as average $\bar{\beta}$ and average SD_β). Further, we computed 80% CV_β widths for each path. Across the 20 paths in this study, 80% CV_β widths ranged from .28 to .93 with a mean width of .54. Of the 199 paths included in all 20 MASEMs, 13 paths (6%) from five published studies showed zero variability. An additional 29 paths (15%) from nine studies showed small heterogeneity (i.e., had 80% CV_β widths smaller than .18 based on our application of the conventions from Bosco et al., 2015). Collectively, these 42 paths are the estimates where we can say that the effect is relatively constant and the traditional MASEM procedure will yield approximately the same information as FIMASEM.

Sixty-eight paths (34%) from 17 published studies demonstrated moderate heterogeneity (i.e., had 80% CV_β between .18 and .54 based on our application of the conventions from Bosco et al., 2015). Finally, 89 paths (45%) in 17 models showed credibility intervals wider than .54. Overall, every study contained at least one path with an 80% CV_β wider than .18, and in 11 of the 20 models all paths had 80% CV_β's wider than .18. This means that across the 20 MASEMs we examined, every study had path coefficients with at least a moderate degree of effect size heterogeneity. In these situations, interpreting only the results derived from Viswesvaran and Ones (1995) is not optimal, as the results do not account for the effect size heterogeneity present in the meta-analytic results when estimating the SEM.

Fit indices. With respect to fit indices, Table 5 displays summary statistics for distributions of χ^2, CFI, and SRMR statistics based on FIMASEM and fit indices from Viswesvaran and Ones (1995) procedure. In 19 of the 20 models, the mean fit indices from FIMASEM indicated worse fit than the Viswesvaran and Ones (1995) procedure (Model 17 was fully saturated and thus had perfect fit). The descriptive statistics in Table 5 highlight the different interpretations FIMASEM uncovers by accounting for SD_ρ. As an example, in reanalyzing Study 9 we found that the SRMR values were above .10 for 62% of the bootstrapped input matrices. Given that the model holds, by traditional SRMR convention (i.e., SRMR $<$ .10; McDonald & Ho, 2002), for only 38% of the population, it would be difficult to argue that the theoretical model generalizes well to the population. However, traditional MASEM reported an SRMR of .07, indicating that the model has acceptable fit. FIMASEM draws a different conclusion, reporting that the model only holds for about one third of the population. Looking across the 20 published articles, we found six models where the model fit, by conventional levels (i.e., SRMR $<$ .10) for less than half of the population. For three of these studies the model fit for less than 10% of the population. These results were worse using CFI, where 13 studies fit less than half of the population, and seven studies fit less than 10% of the population.

To further illustrate the differences between traditional MASEM and FIMASEM, consider published Studies 1 and 2 in Table 5 as examples. Both studies reported SRMR of .02 in our replication

Table 5
Summary Statistics From Traditional and FIMASEM Fit Indices

	MASEM fit indices from published model			Replicated traditional MASEM fit indices			FIMASEM fit indices			
Study	χ^2	CFI	SRMR	χ^2	CFI	SRMR	$\bar{\chi}^2$	$SD(\chi^2)$	% CFI $>$ .90	% SRMR $<$ .10
1	13.11	.99	—	13.63	1.00	.02	711.02	461.79	26%	52%
2	—	.99	—	16.54	.98	.02	206.45	910.9	61%	91%
3	37.83	—	.06	42.76	.96	.04	395.91	372.45	10%	87%
4	320.97	.93	—	221.67	.93	.04	2,246.61	1,082.52	0%	48%
5	306.15	.95	.07	786.78	.81	.11	1,441.47	687.87	0%	7%
6	292.89	.97	.07	1,587.37	.83	.17	2,235.27	1054.6	7%	3%
7	27.24	.99	.03	3.22	.99	.03	430.70	912.78	49%	81%
8	79.89	.98	.02	5.43	1.00	.01	309.91	453.72	97%	100%
9	653.07	.94	—	706.62	.94	.07	5,392.39	4,501.68	7%	38%
10	120.18	.97	.04	142.74	.96	.04	830.69	976.63	44%	96%
11	596.80	.96	.05	626.11	.95	.05	902.38	659.56	90%	100%
12	232.84	.88	.07	82.93	.88	.07	1,521.54	1,115.58	3%	100%
13	5.50	.99	—	17.27	.98	.03	64.02	91.14	61%	87%
14	15.39	.98	—	14.76	.98	.05	215.20	190.02	2%	21%
15	7.07	.99	.02	36.45	.99	.02	587.49	655.31	36%	81%
16	431.48	.96	—	30.45	.95	.05	638.26	1,105.81	27%	85%
17	—	—	—	.00	1.00	.00	.00	.00	100%	100%
18	71.89	.98	.03	1,201.4	.73	.11	1,577.46	465.99	0%	0%
19	189.30	—	—	189.32	.97	.03	1,198.51	1,926.65	62%	88%
20	281.54	.97	—	290.01	.97	.03	439.70	279.52	99%	100%

Note. MASEM = meta-analytic structural equation modeling; FIMASEM = full-information meta-analytic structural equation modeling; — = not reported; χ^2 = chi-square statistic for model; CFI = comparative fit index; SRMR = standardized root mean square residual; $\bar{\chi}^2$ = mean value of χ^2 across 500 iterations; $SD(\chi^2)$ = average standard deviation of the χ^2 across 500 iterations; % CFI $>$ .90 = percentage of CFI statistics that fell above the .90 cutoff across 500 iterations; % SRMR $<$ .10 = percentage of SRMR statistics that fell below the .10 cutoff across 500 iterations.

using traditional MASEM. However, when analyzed using FIMASEM, we found that the studies differed substantially in terms of their generalizability. Published Study 2 achieved acceptable fit based on SRMR in 91% of the bootstrap iterations, while Study 1 achieved acceptable fit based on SRMR in 52% of iterations. The comparison between the two studies highlights the importance of FIMASEM: The best fitting model may not be the most generalizable. The different conclusions cast doubt on the efficacy of the existing MASEM procedure and reveal the importance of FIMASEM in testing theory.

Results for TS-FIMASEM. We retrieved sufficient data to reanalyze two published studies using both traditional MASEM and TSSEM as well as the extensions FIMASEM and TS-FIMASEM: Study 15 and Study 17. To make MASEM procedures comparable, we eliminated potential discrepancies that may be due to the use of different SEM packages and psychometric corrections by running all MASEM procedures in the OpenMx R package. For both published studies, we corrected effect sizes individually for measurement errors. We specified the model identically for each technique, and bootstrapped 500 input matrices for FIMASEM and TS-FIMASEM. The procedures allow for comparisons between each MASEM procedure and its respective extension. Although we reviewed FIMASEM and traditional MASEM results above, we include them in Table 6 and Table 7 for comparison. However, we focus our discussion here on the comparison of TSSEM with TS-FIMASEM.

As there are with traditional MASEM and FIMASEM, there are differences between TSSEM and TS-FIMASEM. In general, TSSEM estimated larger path coefficients than the mean estimates of TS-FIMASEM. More importantly, effects which are statistically significant in TSSEM have 80% CV_β which include both positive and negative effects. For example, in published Study 15 (see Table 6), TSSEM estimated the effect of burnout on accidents to be .10 ($p < .05$). By quantifying the impact of effect size heterogeneity on model estimates, TS-FIMASEM estimated an 80% CV_β from −.14 to .22. The interpretation of this finding is that even though the effect in the average study is positive and statistically significant, the effect in future studies may be positive, negative, or null. This should temper the analyst's conclusion that burnout has a positive effect on accidents, because in a substantial portion of the population of studies, the effect is negative. A similar conclusion can be drawn about the effect of F1 on S2 in published Study 17 (see Table 7). Both traditional MASEM and TSSEM suffer from this issue, reporting statistically significant effects for this study when the respective effect size heterogeneity extensions report 80% CV_β that include zero.

In terms of fit indices, published Study 17 estimated a fully saturated model, and thus exhibited perfect fit (i.e., zero degrees of freedom) in all MASEM analyses. However, we were able to assess the theoretical model of published Study 15, which demonstrated a good fit to the population according to the traditional MASEM fit indices (CFI = .99, SRMR = .02). Nevertheless, FIMASEM fit indices suggested that the theoretical model achieved acceptable fit only for 36% (based on CFI) or 81% (based on SRMR) of the population. TSSEM fit indices of the same model showed that this proposed model did not fit the population data well (CFI = .69, SRMR = .12). Extending upon these findings, TS-FIMASEM results indicated that the proposed model achieved acceptable fit for 0% of the population based on either CFI or SRMR. Thus, TSSEM and TS-FIMASEM come to consistent conclusions about model fit (i.e., the model is not a very good representation of phenomena in the population). However, the conclusion based on TSSEM and TS-FIMASEM is different from that of traditional MASEM and FIMASEM, as they are based upon fundamentally different assumptions about the dependence of effect sizes and the meta-analytic pooling techniques to be applied to the data. From our perspective of developing the two MASEM extensions, it is not clear whether TSSEM or traditional MASEM offers a more accurate picture of the model, but simply that the two

Table 6
TSSEM and TS-FIMASEM Results From Published Study 15

Structural paths	V&O β[a]	FIMASEM $\bar{\beta}$ [80% CV][b]	TSSEM β[b]	TS-FIMASEM $\bar{\beta}$ [80% CV][b]
RH→Burnout	.14*	.14 [−.03, .31]	.19*	.13 [−.02, .29]
SC→Burnout	−.07*	−.07 [−.27, .14]	−.15*	−.08 [−.23, .08]
RH→Compliance	−.43*	−.43 [−.59, −.27]	−.55*	−.43 [−.68, −.18]
SC→Compliance	.46*	.47 [.33, .62]	.49*	.41 [.22, .59]
Burnout→AE	.23*	.23 [.02, .43]	.22*	.22 [.04, .39]
Compliance→AE	−.30*	−.30 [−.47, −.13]	−.41*	−.31 [−.53, −.09]
SC→Accidents	−.09*	−.09 [−.28, .10]	−.11*	−.11 [−.26, .05]
Burnout→Accidents	.01	.00 [−.19, .19]	.10*	.04 [−.14, .22]
Compliance→Accidents	−.03	−.03 [−.25, .20]	−.14*	−.05 [−.28, .18]
Fit indices				
$\chi^2/\bar{\chi}^2$	36.45	587.49	167.30	2409.90
CFI/% CFI > .90	.99	36%	.69	0%
SRMR/% SRMR < .10	.02	81%	.12	0%

Note. TSSEM = two-stage structural equation modeling; TS-FIMASEM = two-stage full-information meta-analytic structural equation modeling; V&O = Viswesvaran & Ones; FIMASEM = full-information meta-analytic structural equation modeling; 80% CV = 80% credibility interval for the path coefficient; RH = Risks and Hazards; SC = Safety Climate; AE = Adverse Events; CFI = comparative fit index; SRMR = standardized root mean square residual. Coefficients are standardized.

[a] Based on uncorrected correlations. [b] $K = 117$ (four studies removed based on being nonpositive-definite), and all missing values were replaced with the weighted mean corrected correlation (multiple imputation would not run due to correlated missingness).

* $p < .05$.

Table 7
TSSEM and TS-FIMASEM Results From Published Study 17

Structural paths	V&O β[a]	FIMASEM β̄ [80% CV][b]	TSSEM β[c]	TS-FIMASEM β̄ [80% CV][c]
F1→S2	.05	.05 [−.09, .19]	.06*	.06 [−.03, .14]
S1→F2	.03	.02 [−.10, .15]	.02	.02 [−.03, .08]
F1→F2	.48*	.48 [.37, .60]	.53*	.46 [.35, .56]
S1→S2	.54*	.53 [.43, .63]	.63*	.53 [.45, .61]
Fit indices				
$\chi^2/\bar{\chi}^2$	.00	.00	.00	99.23
CFI/% CFI > .90	1.00	100%	1.00	4.2%
SRMR/% SRMR < .10	.00	100%	.00	1.8%

Note. TSSEM = two-stage structural equation modeling; TS-FIMASEM = two-stage full-information meta-analytic structural equation modeling; F = Family interference with work; F1 = Family interference with work time1; F2 = Family interference with work time2; S = Strain; S1 = Strain Time1; S2 = Strain Time2; V&O = Viswesvaran & Ones; FIMASEM = full-information meta-analytic structural equation modeling; % CV = 80% credibility interval for the path coefficient; CFI = comparative fit index; SRMR = standardized root mean square residual. Coefficients are standardized. Computed SD_ρ values were noticeably higher than those reported in the original manuscript.
[a] Based on uncorrected correlations. [b] Based on bare-bones meta-analysis of individually corrected correlations. [c] $K = 12$ (eight studies removed based on being nonpositive-definite).
* $p < .05$.

converge on different solutions based on their consideration of the sources of variability and dependence of effect sizes.

General Discussion

MASEM procedures, as they are currently used in applied psychology and management, are ineffective in incorporating and quantifying effect size heterogeneity, which can lead to inconsistent or even erroneous conclusions. The existing MASEM techniques, traditional MASEM (Viswesvaran & Ones, 1995) and TSSEM (Cheung, 2014), both recognize the presence of effect size heterogeneity in meta-analytic results. Yet neither fully quantifies how effect size heterogeneity leads to a distribution of SEM estimates that stem from the meta-analytic results. This presents a problem in that scholars using the technique must ignore the effect size heterogeneity they observe in meta-analysis when they estimate the SEM. Because neither MASEM technique quantifies the distributions of path estimates and model fit indices in the presence of effect size heterogeneity, we introduce FIMASEM and TS-FIMASEM in the current study as a tool to better understand the impact of effect size heterogeneity on the conclusions of MASEM studies. FIMASEM and TS-FIMASEM thus serve as supplementary tools to existing MASEM procedures and illuminate the distribution of model parameters in a population of studies. In this way, FIMASEM and TS-FIMASEM demonstrate the problem of effect size heterogeneity and suggest a continued need for MASEM techniques that offer a more complete picture of effects in the population.

We demonstrated the problem of effect size heterogeneity in two studies. First, we conducted six simulations in Study 1, which highlighted the ways in which the presence of effect size heterogeneity can lead to biased conclusions in traditional MASEM and TSSEM. In Study 2, we reanalyzed 20 published MASEM studies using all MASEM techniques available for each study. A key finding consistent across both studies was that traditional MASEM and TSSEM overestimated the extent to which MASEM results are generalizable to the population, whereas FIMASEM and TS-FIMASEM offered new information about the generalizability of the MASEM results by quantifying effect size heterogeneity and producing full information about the distributions of model estimates. Our findings suggest existing MASEM techniques, in their current form, offer a less complete assessment of theoretical models and their generalizability.

Implications for Research and Practice

FIMASEM highlighted the problem of effect size heterogeneity in MASEM research, illustrating several opportunities to advance theory and inform practice by accounting for and quantifying effect size heterogeneity. First, the width of CV_β provides additional information that can direct future research on the presence of moderators where wider credibility intervals indicate stronger moderation effects (McEvoy & Cascio, 1987; Whitener, 1990). Consistent with Edwards and Berry's (2010) call for increasing theoretical precision in management research, credibility intervals allow researchers to identify and specify boundary conditions in subsequent analysis or future empirical studies through a multivariate, rather than a bivariate, analysis.

Second, the analysis of the distribution of fit indices offers insights on when or under what conditions one particular theory receives stronger support than alternative theories. Simple models may have better overall fit, but less generalizability than complex models or vice versa. Accounting for true score effect size heterogeneity and evaluating on the impact of heterogeneity allow for the delineation of the best fitting model from the most generalizable model. FIMASEM estimates this by bootstrapping from a distribution of ρ and SD_ρ that can be further used in SEM to estimate a distribution of model parameters. Because it provides more information about the distribution of model estimates in the population of studies, FIMASEM demonstrates how the incorporation of effect size heterogeneity into MASEM can expand theory. For example, a theoretical model that only generalizes to a small portion of a population should encourage future studies to

specify boundary conditions or explore alternative theoretical explanations based on the assessment of model fit.

Third, the idea of examining a distribution of estimates through credibility intervals is consistent with Bayesian logic that moves MASEM research away from null hypothesis significance testing (e.g., Orlitzky, 2012; Zyphur & Oswald, 2015). Instead of adopting a frequentist perspective about the likelihood of obtaining a model estimate were the null hypothesis is true, researchers can make more credible inference based on the probability that a meaningful effect of a certain size or direction for a particular path. Such direct inference cannot be accomplished through null hypothesis significance testing, which comes along with several limitations (Nickerson, 2000). Future MASEM scholars can use FIMASEM and TS-FIMASEM extensions to quantify the distributions of effects and overcome some limitations of null hypothesis significance testing and pursue Bergh et al.'s (2016) recommendation to use MASEM to evaluate comparative fit of alternative models in with an experimental approach to identifying which models empirically generalize the best.

Resolving the problem of effect size heterogeneity in MASEM also would allow practitioners to assess the extent to which results from a study are likely to manifest in their organizations. We achieve this using MASEM to interpret the percent generalizability on model fit indices. For example, five out of the 20 MASEM models had SRMR value less than .10 in all of the 500 iterations (100%). This implies that the theoretical models reported in these five studies generalize to the whole population. Because there is a high probability that these models will fit in future studies of these sets of constructs, practitioners can take the results and implications of those studies with more confidence. In contrast, some studies exhibited relatively poor generalizability in the FIMASEM procedure, but they achieved acceptable model fit in the traditional MASEM procedure. More specifically, the fit indices of the fixed effect model of published Study 14 both passed the traditional thresholds (i.e., CFI = .98, SRMR = .05), whereas its random effect model suggested a low degree of generalizability. Only 21% out of 500 conditions obtained SRMR values less than .10. For about four-fifths of the population, there remain alternative models that better explain antecedents of union participation.

Practitioners may also assess the risk and effectiveness of certain workplace interventions or HR policies based on the distributions of relevant path coefficients. In our example FIMASEM, prounion attitudes was positively related to union commitment for 93% of the population, despite that the magnitude of the positive effect varied across different subpopulations. In contrast, the effects of organizational commitment and job satisfaction on union commitment were positive for only 61% and 51% of the population, respectively. For organizations designing policies related to union commitment and participation, the results of published Study 14 suggest that workers' prounion attitudes would be a more credible factor to consider than workers' level of organizational commitment and job satisfaction.

Traditional MASEM Versus TSSEM

We developed FIMASEM and TS-FIMASEM as complementary tools to the traditional MASEM and TSSEM. However, a reader might question which MASEM technique to employ in future research. As summarized in Table 1, the answer to this question lies in part in the specific research context, including: the dependence among effect sizes, the availability of primary data, and the researcher's need for statistical inference. First, if effect sizes are fully independent from one another, traditional MASEM (and thus FIMASEM) is appropriate. To the extent that the assumption of independence is violated, scholars may utilize the multivariate approach in TSSEM (Cheung, 2014) and thus TS-FIMASEM. TSSEM offers a more robust statistical significance test of model estimates by taking into account of the dependence among effect sizes than does traditional MASEM. In our reanalysis, published Study 15 and 17 had similar effects regardless of the base MASEM techniques used, which conforms to simulation results suggesting violations of effect size independence may not be as problematic as once thought (Schmidt & Hunter, 2015; Tracz, Elmore, & Polhmann, 1992). However, TSSEM should in theory offer estimates that are more robust to violations of the assumption of independence of effect sizes.

Second, one practical challenge of applying TSSEM and TS-FIMASEM is the availability of primary data for a particular MASEM. In applied psychology and management, it is not uncommon for primary studies to only examine a part of a larger theoretical model, or for MASEM scholars to synthesize theory and test effect sizes from studies in disparate literatures. For this reason, a MASEM researcher examining an integrative model may not be able to identify at least one primary study that has complete correlations among all study variables. This is a necessary condition to perform TSSEM and, by extension, TS-FIMASEM. Traditional MASEM, in combination with FIMASEM, has less strict data requirements, allowing for analysis even when no primary study has examined all of the variables in the model. For this reason, scholars may find traditional MASEM (and thus FIMASEM) to be more flexible for synthesizing diverse perspectives in a weak paradigm field (Glick, Miller, & Cardinal, 2007).

Finally, FIMASEM and TS-FIMASEM influence neither significance levels nor standard errors (nor, by extension, confidence intervals) of model estimates, but the significance tests of traditional MASEM and TSSEM differ. Both traditional MASEM and TSSEM provide standard errors that can be used to construct confidence intervals around SEM estimates. However, because TSSEM adjusts for effect size dependence and heterogeneity, it offers an advantage in conducting statistical significance tests and constructing confidence intervals (Cheung, 2015). Thus, researchers should rely on TSSEM where possible to compute more accurate confidence intervals. The FIMASEM and TS-FIMASEM extensions focus on credibility intervals, which offer distinct information in meta-analysis (Whitener, 1990). Whereas confidence intervals focus on the accuracy and statistical significant of a meta-analytic estimate, credibility intervals represent the generalizability of the meta-analytic estimate (Schmidt & Hunter, 1977). Thus, the extensions focus on a different interval than the base techniques. Researchers interested in confidence intervals should use TSSEM, if possible, and traditional MASEM, if TSSEM cannot be conducted, to retrieve standard errors required to construct confidence intervals.

Limitations and Future Directions

The FIMASEM extensions focus primarily on one criticism of the current MASEM techniques: the failure to incorporate effect

size heterogeneity from meta-analytic results to the estimation of SEMs. It is important to note that as extensions of the traditional MASEM and TSSEM, FIMASEM does not address several other valid criticisms of both current MASEM techniques. Bergh et al. (2016) present an exhaustive list of these limitations, and we advocate future research in this area to continue to improve and develop MASEM methods. However, the current study also has limitations in its handling of effect size heterogeneity. We focus this section on identifying and clarifying these limitations.

First, the present study is limited by our inability to fully replicate a small number of MASEM articles. Of the 73 articles we identified, we could only replicate, and thus reanalyze, 20 articles. Though the 53 eliminated articles were discarded for various reasons, we cannot refute the possibility that in these 53 studies, MASEM techniques might have yielded less divergent results from their respective extensions. It is possible that the 20 models we reanalyzed are in key ways different from those studies that used MASEM, but whose findings could not be replicated. For example, we excluded five studies using SEM with multiple indicators, leaving all 20 models using path analysis. This is because "lavaan" and LISREL differ in their estimation and reporting of factor loadings for constructs with multiple indicators. Differences in factor loadings are likely to have an impact on path estimates. To make a valid comparison in our study, we only included MASEM studies that could be reasonably replicated. When corrected effect sizes are pooled to construct matrices, we suggest the use of path analysis to avoid over correction of measurement errors (Podsakoff, MacKenzie, & Bommer, 1996). When uncorrected effect sizes are used, SEM with multiple indicators is generally preferred. Scholars implementing FIMASEM with multiple indicators should take care when comparing "lavaan" and "Openmx" outputs to those from other SEM software, and we advocate for conducting path analysis on individually corrected effects in TS-FIMASEM or using artifact distribution techniques (Schmidt & Hunter, 2015) for FIMASEM where possible.

Second, we applied benchmark values on the CV width that were converted from Bosco et al.'s (2015) findings based on field-wide correlations, and we used conventional levels of fit as cutoffs for CFI and SRMR (McDonald & Ho, 2002). While we emphasize and encourage reporting information about the entire distribution of model estimates, the use of cut-off points offers a strength to management researchers in terms of facilitating the communication of research. Bosco et al.'s conventions are empirically based and are the best available estimates for building such conventions. However, Bosco et al. found noticeable variation in effect sizes across content areas of applied psychology and management, suggesting that these conventions should be adjusted for different kinds of research. The MASEMs we analyzed crossed content areas and could not be easily categorized in Bosco et al.'s coding scheme, but future research should consider using the content-specific convention where possible.

Third, there are differences in estimates between FIMASEM and TS-FIMASEM that warrant future research. To some degree these differences are rooted in the divergent approaches of how traditional MASEM and TSSEM account for various sources of variance in ρ. However, the differences appear to be larger than one would expect strictly from the computational procedures between the two base techniques. Future research comparing the two base techniques would be valuable in identifying the extent to which the two base MASEM approaches produce accurate estimates of population parameters. Such work would help resolve the differences between FIMASEM and TS-FIMASEM that manifest due to the underlying base techniques.

Finally, although the results showed the path estimates from traditional MASEM and TSSEM fell within the 80% CVs derived from FIMASEM and TS-FIMASEM, there were differences, in some studies, between the mean estimate in FIMASEM and TS-FIMASEM and the point estimate in MASEM and TSSEM. It is possible that the divergence between these two values is due to the incorporation of effect size heterogeneity into MASEM. For example, in the Figure 1 model in our simulation, larger ρ_{XY} values (positive or negative) will result in lower values for β_{XM1} and β_{XM2}. Thus, larger values of $SD_{\rho XY}$ will result in generally smaller β values, meaning the mean β in FIMASEM will be different from the β estimate in MASEM. An alternative source of bias occurs through the elimination and replacement of non-positive-definite matrices. Certain combinations of ρ and $SD\rho$ in a ρ matrix may place upper or lower bounds on a given ρ_{ij}, skewing the distribution of ρ_{ij} in the bootstrapped distribution. This bias is theoretically legitimate—even if bivariate meta-analysis computes a given $SD\rho_{ij}$, it may be mathematically impossible to observe certain values of ρ_{ij} given the relationships between other variables in the model. However, there could be less theoretically legitimate reasons for the bias between the two values. For example, we take a finite number of iterations in the parametric bootstrap (i.e., 500), which may necessarily introduce sampling error. Theoretically, the distribution produced by bootstrap would be equal to the estimated value produced by the analytical counterpart if the number of iterations were infinite (Efron, 1987). Such bias could be greatly reduced by taking a larger number of bootstrap iterations (e.g., 10,000), which can easily be accomplished in R but cannot technically be accomplished using the online software application. For current FIMASEM and TS-FIMASEM researchers, we recommend assessing the skewness of ρ and β prior to constructing 80% CVs. Future research is also needed to better identify when and why mean β in FIMASEM deviates from traditional MASEM β estimates.

Conclusion

The combination of random effects meta-analysis and SEM offers a powerful way to build and test theory (Bergh et al., 2016). The current research extends a major strength of meta-analysis—identifying effect size heterogeneity—to SEM in a way that allows researchers to identify the portion of the population that is represented by a model estimate. Further, our technique builds on current MASEM techniques by quantifying the impact of effect size heterogeneity on the distribution of SEM estimates. By applying this technique in research synthesis, scholars have the opportunity to build more precise theories that more fully explain the phenomena we observe.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Bauer, T. N., Bodner, T., Erdogan, B., Truxillo, D. M., & Tucker, J. S. (2007). Newcomer adjustment during organizational socialization: A

meta-analytic review of antecedents, outcomes, and methods. *Journal of Applied Psychology, 92,* 707–721. http://dx.doi.org/10.1037/0021-9010.92.3.707

Bergh, D. D., Aguinis, H., Heavey, C., Ketchen, D. J., Boyd, B. K., Su, P., . . . Joo, H. (2016). Using meta-analytic structural equation modeling to advance strategic management research: Guidelines and an empirical illustration via the strategic leadership-performance relationship. *Strategic Management Journal, 37,* 477–497.

*Berry, C. M., Lelchook, A. M., & Clark, M. A. (2012). A meta-analysis of the interrelationships between employee lateness, absenteeism, and turnover: Implications for models of withdrawal behavior. *Journal of Organizational Behavior, 33,* 678–699. http://dx.doi.org/10.1002/job.778

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100,* 431–449. http://dx.doi.org/10.1037/a0038047

*Carr, J. Z., Schmidt, A. M., Ford, J. K., & DeShon, R. P. (2003). Climate perceptions matter: A meta-analytic path analysis relating molar climate, cognitive and affective states, and individual level work outcomes. *Journal of Applied Psychology, 88,* 605–619. http://dx.doi.org/10.1037/0021-9010.88.4.605

Cheung, M. W.-L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling, 20,* 429–454. http://dx.doi.org/10.1080/10705511.2013.797827

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

Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. West Sussex, United Kingdom: Wiley. http://dx.doi.org/10.1002/9781118957813

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

*Christian, M. S., Garza, A. S., & Slaughter, J. E. (2011). Work engagement: A quantitative review and test of its relations with task and contextual performance. *Personnel Psychology, 64,* 89–136. http://dx.doi.org/10.1111/j.1744-6570.2010.01203.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. http://dx.doi.org/10.1177/001316446002000104

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

*Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology, 92,* 909–927. http://dx.doi.org/10.1037/0021-9010.92.4.909

*Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., & Wesson, M. J. (2013). Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology, 98,* 199–236. http://dx.doi.org/10.1037/a0031757

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

*Courtright, S. H., Thurgood, G. R., Stewart, G. L., & Pierotti, A. J. (2015). Structural interdependence in teams: An integrative framework and meta-analysis. *Journal of Applied Psychology, 100,* 1825–1846. http://dx.doi.org/10.1037/apl0000027

*Eatough, E. M., Chang, C.-H., Miloslavic, S. A., & Johnson, R. E. (2011). Relationships of role stressors with organizational citizenship behavior: A meta-analysis. *Journal of Applied Psychology, 96,* 619–632. http://dx.doi.org/10.1037/a0021887

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668–689. http://dx.doi.org/10.1177/1094428110380467

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82,* 171–185. http://dx.doi.org/10.1080/01621459.1987.10478410

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods, 10,* 227–254. http://dx.doi.org/10.1037/1082-989X.10.2.227

Glick, W. H., Miller, C. C., & Cardinal, L. B. (2007). Making a life in the field of organizational science. *Journal of Organizational Behavior, 28,* 817–835. http://dx.doi.org/10.1002/job.455

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Englewood Cliffs, NJ: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. http://dx.doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology, 37,* 1158–1160. http://dx.doi.org/10.1093/ije/dyn204

*Hom, P. W., Caranikas-Walker, F., Prussia, G. E., & Griffeth, R. W. (1992). A meta-analytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology, 77,* 890–909. http://dx.doi.org/10.1037/0021-9010.77.6.890

*Hong, Y., Liao, H., Hu, J., & Jiang, K. (2013). Missing link in the service profit chain: A meta-analytic review of the antecedents, consequences, and moderators of service climate. *Journal of Applied Psychology, 98,* 237–267. http://dx.doi.org/10.1037/a0031666

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

*Jiang, K., Liu, D., McKay, P. F., Lee, T. W., & Mitchell, T. R. (2012). When and how is job embeddedness predictive of turnover? A meta-analytic investigation. *Journal of Applied Psychology, 97,* 1077–1096. http://dx.doi.org/10.1037/a0028610

*Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100,* 298–342. http://dx.doi.org/10.1037/a0037681

*Kossek, E. E., Pichler, S., Bodner, T., & Hammer, L. B. (2011). Workplace social support and work–family conflict: A meta-analysis clarifying the influence of general and work-family-specific supervisor and organizational support. *Personnel Psychology, 64,* 289–313. http://dx.doi.org/10.1111/j.1744-6570.2011.01211.x

Landis, R. S. (2013). Successfully combining meta-analysis and structural equation modeling: Recommendations and strategies. *Journal of Business and Psychology, 28,* 251–261. http://dx.doi.org/10.1007/s10869-013-9285-x

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82. http://dx.doi.org/10.1037/1082-989X.7.1.64

McEvoy, G. M., & Cascio, W. F. (1987). Do good or poor performers leave? A meta-analysis of the relationship between performance and turnover. *Academy of Management Journal, 30,* 744–762. http://dx.doi.org/10.2307/256158

*Monnot, M. J., Wagner, S., & Beehr, T. A. (2011). A contingency model of union commitment and participation: Meta-analysis of the antecedents of militant and nonmilitant activities. *Journal of Organizational Behavior, 32,* 1127–1146. http://dx.doi.org/10.1002/job.736

*Nahrgang, J. D., Morgeson, F. P., & Hofmann, D. A. (2011). Safety at work: A meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. *Journal of Applied Psychology, 96,* 71–94. http://dx.doi.org/10.1037/a0021484

*Ng, T. W. H., & Feldman, D. C. (2015). Ethical leadership: Meta-analytic evidence of criterion-related and incremental validity. *Journal of Applied Psychology, 100,* 948–965. http://dx.doi.org/10.1037/a0038246

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. http://dx.doi.org/10.1037/1082-989X.5.2.241

*Nohe, C., Meier, L. L., Sonntag, K., & Michel, A. (2015). The chicken or the egg? A meta-analysis of panel studies of the relationship between work-family conflict and strain. *Journal of Applied Psychology, 100,* 522–536. http://dx.doi.org/10.1037/a0038012

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods, 15,* 199–228. http://dx.doi.org/10.1177/1094428111428356

Podsakoff, P. M., MacKenzie, S. B., & Bommer, W. H. (1996). Transformational leader behaviors and substitutes for leadership as determinants of employee satisfaction, commitment, trust, and organizational citizenship behaviors. *Journal of Management, 22,* 259–298.

*Robbins, S. B., OH, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and retention as mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *Journal of Applied Psychology, 94,* 1163–1184. http://dx.doi.org/10.1037/a0015738

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1–36. http://dx.doi.org/10.18637/jss.v048.i02

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540. http://dx.doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

*Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. *Personnel Psychology, 46,* 259–293. http://dx.doi.org/10.1111/j.1744-6570.1993.tb00874.x

Tracz, S. M., Elmore, P. B., & Polhmann, J. T. (1992). Correlational meta-analysis: Independent and nonindependent cases. *Educational and Psychological Measurement, 52,* 879–888. http://dx.doi.org/10.1177/0013164492052004007

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

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315–321. http://dx.doi.org/10.1037/0021-9010.75.3.315

*Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61,* 309–348. http://dx.doi.org/10.1111/j.1744-6570.2008.00115.x

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41,* 390–420.

Received April 20, 2015
Revision received May 27, 2016
Accepted May 31, 2016 ■