American Psychological Association

Journal of Applied Psychology

0021-9010/18/$12.00

2018, Vol. 103, No. 7, 804–811
http://dx.doi.org/10.1037/apl0000328

REPLY

The Heterogeneity Problem in Meta-Analytic Structural Equation Modeling (MASEM) Revisited: A Reply to Cheung

Jia (Joya) Yu
University of Nebraska-Lincoln

Patrick E. Downes
Rutgers University

Kameron M. Carter
University of Iowa

Ernest O'Boyle
Indiana University

Yu, Downes, Carter, and O'Boyle (2016) introduce a new technique to incorporate effect size heterogeneity into meta-analytic structural equation modeling (MASEM) labeled full information meta-analytical structural equation modeling (FIMASEM). Cheung's (2018) commentary raises concerns about the viability of FIMASEM and provides its initial validation. In this reply, we briefly respond to those concerns noting how they relate to Yu et al.'s original conclusions, general MASEM practices, and operational decisions within the FIMASEM procedure. We synthesize Cheung's criticisms and build on his findings to lay out a research agenda for the future of MASEM and the role that our technique might play in it. In doing so, we clarify the conceptual nature of FIMASEM, identity inferential mistakes that current MASEM studies are likely to make, and offer specific and actionable recommendations in terms of the types of research questions FIMASEM is best suited to address and how FIMASEM results can best be interpreted and reported.

Keywords: meta-analysis, structural equation modeling, meta-analytical structural equation modeling

Supplemental materials: http://dx.doi.org/10.1037/apl0000328.supp

Yu, Downes, Carter, and O'Boyle (2016) identified inconsistency in the current practice of meta-analytic structural equation modeling (MASEM). Specifically, MASEM assumes effect size heterogeneity at the stage of meta-analysis but then treats the parameters as homogenous or fixed at the stage of structural equation modeling (SEM). Put differently, in the meta-analysis portion of MASEM, researchers assume that no one true correlation generalizes across the full population. Rather, the true correlation varies depending on the specific subpopulation under examination. However, at the SEM stage of MASEM, those same correlations are assumed to not vary and not be moderated by substantive and/or methodological variables. Regardless of whether researchers used Viswesvaran and Ones's (1995) technique or Cheung and Chan's (2009) two-stage structural equation modeling (TSSEM) technique, most published MASEM studies have made these inconsistent assumptions.

Yu et al. (2016) developed a procedure called full information MASEM (FIMASEM) that attempted to reconcile the inconsistent assumptions of whether an effect size varies across the population. Through a simulation study and a reanalysis of existing MASEM data sets, Yu et al. demonstrated that when effect size heterogeneity is incorporated into both the meta-analysis step as well as the SEM step, results and interpretations of robustness can be substantially different than when the heterogeneity is ignored.

Cheung's (2018) commentary on Yu et al. (2016) is a critical evaluation of the FIMASEM approach. He introduces issues regarding (a) the stability of $SD\rho$, (b) the handling of non-positive-definite (NPD) matrices, (c) the performance of a generalizability index based on fit indices, (d) the interpretation of misspecified models, and (e) technical coding errors. Of these five issues, Cheung concludes through his independent simulation that three of them (i.e., Issues 1, 2, and 4) are "quite mild" (p. 28) and have little effect on parameter estimates, variance estimates, and overall conclusions. Nevertheless, his commentary brings up the need to further conceptually clarify how heterogeneity can be incorporated into MASEM research and how FIMASEM procedures might be used to improve rigor and answer new research questions.

Jia (Joya) Yu, Department of Management, College of Business, University of Nebraska-Lincoln; Patrick E. Downes, Department of Human Resource Management, Rutgers University; Kameron M. Carter, Department of Management & Organizations, Henry B. Tippie College of Business, University of Iowa; Ernest O'Boyle, Department of Management and Entrepreneurship, Indiana University.

Correspondence concerning this article should be addressed to Jia (Joya) Yu, Department of Management, College of Business, University of Nebraska-Lincoln, CoB 325H, 730 North 14th Street, P.O. Box 880491, Lincoln, NE 68588-0491. E-mail: joya.jia.yu@gmail.com

The purpose of this reply is not to rebut each of Cheung's (2018) criticisms. Some of his criticisms are valid and helpful to correct the published record (e.g., there *were* errors in the code). Other criticisms are entirely accurate but general to common practices in MASEM (e.g., reliance on cutoff values to interpret model fit). We do diverge on a few points that we briefly clarify/contest below, but there is far more alignment in the perspectives of both articles than misalignment—the most crucial point of alignment being that effect size heterogeneity has the potential to be very problematic in MASEM when it is present but ignored. Thus, the purpose of this reply is to outline how to move the field forward when it comes to MASEM. After the aforementioned clarifications, we lay out a research agenda concerning the future and viability of FIMASEM and, more broadly, MASEM. In particular, we specify different scenarios for current MASEM practice with an updated review and highlight when ignoring heterogeneity can lead to inaccurate inferences. In addition, we elaborate on the types of research questions that FIMASEM is best suited to address.

Clarifying the Conceptual Nature of the FIMASEM Approach

Cheung (2018) summarizes the current statistical assumptions underlying MASEM into two categories: correlation-based MASEM and parameter-based MASEM. He argues that FIMASEM falls into the parameter-based MASEM approach (p. 788). However, this description is not accurate, and before moving forward, we need to clarify the exact nature of FIMASEM. FIMASEM and Two-Stage FIMASEM (TS-FIMASEM) are extensions of correlation-based approaches such as Viswesvaran and Ones (1995) and TSSEM (Cheung & Chan, 2009). This means that FIMASEM and TS-FIMASEM first meta-analyze the correlations in primary studies, then build a correlation matrix of meta-analyzed effects, and finally specify a structural model. This is fundamentally different from parameter-based MASEM, which fits a structural equation model for each primary study then meta-analyzes the parameters (i.e., path coefficients and factor loadings).

Whereas parameter-based approaches conceptualize heterogeneity solely on the parameters, correlation-based approaches conceptualize heterogeneity solely on the correlations. FIMASEM treats effect size heterogeneity consistently throughout all phases of analysis. The central argument of Yu et al. (2016) was that assumptions about heterogeneity should be consistent across both stages of MASEM. FIMASEM maintains this consistency by conceptualizing heterogeneity on both the correlations and the path coefficients. Essentially, FIMASEM (or TS-FIMASEM, for that matter) intends only to address the inconsistency of heterogeneity assumptions between the two steps by adding an intermediate bootstrapping procedure.

In Step 1, FIMASEM is identical to a standard Schmidt and Hunter (2015) psychometric meta-analysis. It generates estimates of population correlations from primary studies and partitions observed variance into that which can be explained by sampling error and statistical artifacts and that which can be attributed to true-score variance (e.g., moderators). A correlation matrix and a true-score variance matrix are then built based on these estimates.

Next is an intermediate step and it is here where FIMASEM diverges from Viswesvaran and Ones's (1995) procedures. Rather than solely using information about the correlations at the mean (ρ), FIMASEM also uses information about how much that correlation varies across the population (specifically, $SD\rho$). Using both ρ and $SD\rho$, FIMASEM resamples or bootstraps multiple correlation matrices. The amount that the correlations in these bootstrapped matrices vary is directly proportional to their true-score variation identified in Step 1. For example, if a correlation (e.g., $\rho = .30$) had a small amount of true score variance (e.g., $SD\rho = .01$), then across the hundreds or thousands of bootstrapped matrices, that particular correlation would be relatively constant (e.g., $.28 < \rho s < .32$). The larger the true score variance, the larger the range of values ρ could take in any given matrix.

In Step 2, the SEM step, FIMASEM fits the theoretical model to those bootstrapped matrices following Viswesvaran and Ones's (1995) procedure. The difference is that rather than fitting one model based exclusively on the mean estimate of each correlation, FIMASEM fits the model to all the correlation matrices generated in the intermediate step. The result is that instead of getting one set of results based on one model of mean estimates (e.g., RMSEA = .07), FIMASEM provides a distribution of results (e.g., across 500 iterations, the RMSEA ranged from .00 to .18, with a mean of .07). The central logic of FIMASEM is that one model specification may not fit well across all subpopulations, but it is important to know the extent or percentage of subpopulations that do and do not fit a particular model specification.

Brief Reply to Issues Raised in Cheung (2018)

We now turn to the five specific criticisms Cheung (2018) raised. We present our responses not in the original order but in three categories: (a) whether they can be improved and affect the conclusions of Yu et al. (2016), (b) whether they apply to the general practice of meta-analysis/MASEM but do not jeopardize the proposed FIMASEM procedure, and (c) whether they are related to operational decisions of FIMASEM that may require future simulation work.

Issues Related to Yu et al.'s (2016) Conclusions

Syntax errors. Cheung (2018) points out two technical errors in Yu et al. (2016) regarding (a) the Reticular Action Modeling (RAM) specification of paths, and (b) the error variance specifications. He is entirely correct. Our syntax in part of the Study 1 simulation incorrectly specified error variances and the direction of paths.[1] For the model in Figure 1 (Yu et al., 2016, p. 1461), this error has no impact on results because the model is symmetric: The results are as if X and Y were exchanged in the Figure. However, for Figure 2 (Yu et al., 2016, p. 1461), the error resulted in us effectively testing a different model with two predictors and three outcomes (instead of three predictors and two outcomes). New results after correction for both errors are presented in Table SM1 of Supplementary Material 1 in the online supplemental materials.

[1] The elements of Study 1 and Study 2 using lavaan specification are correct. The technical errors are only present in the Study 1 simulation syntax using OpenMx specifications. Results from Study 2 are from the FIMASEM website, where the syntax correctly specified sigma, the direction of paths, error variances, and so on. Users can continue to use the FIMASEM website to generate results.

Different results, same conclusion—effect size heterogeneity is present and not currently incorporated in MASEM. Nevertheless, it is important to correct the record, and we thank Cheung for identifying the errors.

Model-implied matrices. Cheung (2018) points out that the models presented in Figure 1 and 2 in Yu et al. (2016) are not the models actually tested. Cheung's criticism reflects the earlier assumption about FIMASEM being a parameter-based technique as opposed to its correct classification as a correlation-based technique. Focusing on the path coefficient, Cheung assumes that the parameter (i.e., the β coefficient) from X→Y in Figure 1 is zero. If accurate, this would mean that the model-implied correlation between X and Y is not zero due to the indirect paths between X and Y. From a parameter-based standpoint, this interpretation is reasonable, but this is not how the model was conceptualized and it is inconsistent with FIMASEM's correlation-based nature. Yu et al. fixed the bivariate correlation between X and Y to zero, not the multivariate path between X and Y (p. 1460). As such, there are no errors in the model-implied matrices for Figure 1 or Figure 2 (in Yu et al., 2016), but we understand where the confusion is derived from and we hope this reply solidifies FIMASEM as an extension of the correlation-based technique.

Issues Related to General Meta-Analytic/MASEM Practice

The stability of the population estimates of SDρ. Cheung (2018) rightly points out that when the number of included studies for any correlation in the MASEM is low, then the stability of *SD*ρ is reduced. This is a critical consideration and one that is often overlooked, as much of the focus in applied psychology has been on the stability of the mean estimate of the correlation, not the stability of the variance estimate. Steel and Kammeyer-Mueller (2008) made a similar argument to Cheung's that *SD*ρ is affected by second-order sampling error. This criticism that variance estimates lose stability as the number of studies and/or overall sample size goes down (ceteris paribus) is not new to meta-analysis and not specific to the FIMASEM approach. Ultimately, Cheung's criticism is valid, but it is a criticism relevant to all meta-analysis practices that attempt to empirically partition true score variance from observed variance with a small *k* (including TSSEM). Our recommendation, and one likely shared by Cheung, is that when some cells, or even one cell, of the MASEM matrix is based on a small number of studies, great care should be taken in interpreting the variance estimates or perhaps even avoiding any form of MASEM including FIMASEM or TSSEM.

Impact of NPD matrices. Cheung's (2018) simulation tests two ways of handling the NPD problem (i.e., nearPD and replacement), and he concludes that they both provide similar results and that the conclusions are largely unaffected, but the nearPD technique performs slightly better. FIMASEM is agnostic when it comes to which technique to use (any NPD handling method of the researchers' choosing can be built into the code), but we support Cheung's conclusion.

However, we are hesitant to recommend that researchers simply ignore the frequency of NPD matrices during the bootstrapping procedure. A high frequency of NPD matrices could indicate problems with the point and variance estimates in the overall matrix. Just as in standard SEM, NPD could indicate multicollinearity, out of bounds correlations, excessive missing data, or other problems (see Worthke, 1993, for a detailed discussion). Recognizing these NPD matrices is a strength of FIMASEM. In bivariate meta-analysis, multivariate relationships are ignored when effect size heterogeneity is estimated. This is problematic because independent meta-analyses of related constructs across a literature could produce credibility intervals that are logically incompatible. For example, assume three constructs (e.g., A, B, and C), where A-B and A-C have true score correlations of .60. In order for the 3 × 3 matrix to be positive definite, the B-C correlation logically must be between −.28 and .99. This is ignored in bivariate meta-analysis and traditional MASEM, such that the credibility interval from an independent meta-analysis of the B-C correlation could be below −.28. In FIMASEM, a bootstrapped B-C correlation below −.28 is logically impossible and would contribute to an NPD matrix.

Our perspective is that NPD issues have theoretical value and should be explored, not statistically masked. After researchers investigate the causes of NPD issues, they can decide whether it would be prudent to (a) switch to other estimators such as asymptotically distribution-free or weighted least squares (ADF/WLS), (b) exclude a variable causing linear dependency, or (c) examine the correlation matrix for strong correlations combined with large *SD*ρ and constrain distributions that exceed a certain value (e.g., .80) and justify a course of action. At the very least, we encourage researchers to report the percentage of matrices found to be NPD.

Issues Related to Operational Decisions of FIMASEM

Choice of bootstrap iterations. We chose 500 bootstrap iterations in the Yu et al. (2016) article because 500 iterations reduced bootstrap error to an acceptable level (see Supplementary Material 2 of the online supplemental materials). FIMASEM demands significant computer processing power as the model becomes more complex. Each bootstrap involves an iterative SEM estimation, and fitting hundreds of models can be quite time-consuming. Five hundred iterations appears to reasonably balance bootstrap error with technical feasibility. However, we are in complete agreement with Cheung (2018) that more iterations are better.

Choice of sample size. Cheung (2018) highlights that the issue of sample size is critical and will have an impact on the performance of fit indices. In Yu et al. (2016), we used the mean sample size in Study 1 simulations. In Study 2, we used the same sample sizes that were used in the original published MASEM studies—usually the harmonic mean. Because in both studies we used the same sample size for comparison, the fit indices from the traditional MASEM and FIMASEM will be equally impacted. In the current literature, there is scant definitive evidence for choosing the sample size in MASEM. Whether the total sample size, harmonic mean, or some alternative measures of central tendency should be used in MASEM require far more simulation and investigation before any definitive recommendation can be made (see Bergh et al., 2016, for a review).

Use of fit index cutoffs. Artificial dichotomization of any continuous statistic into "good" and "bad" leads to a host of problems (Lance & Vandenberg, 2009; O'Boyle, 2017; Williams & O'Boyle, 2011; Williams, O'Boyle, & Yu, 2017). Despite the fact that fit index cutoffs are logically and computationally prob-

lematic, their use has become standard practice in SEM and MASEM and are typically how models are evaluated. If a model fits well based on some threshold values, it is interpreted as showing theoretical support. Even those of us that recognize the problems of cutoffs often still use them for illustrative purposes (e.g., Cheung, Leung, & Au, 2006; O'Boyle & Williams, 2011). Ultimately, this is not a point of contention, as few in the research methods community, including Cheung and ourselves, advocate for cutoffs to be the sole means to evaluate model quality.

Moving Forward With FIMASEM Fit Indices

Cheung's (2018) simulation reveals that commonly used fit indices in SEM (e.g., chi square, comparative fit index [CFI], root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR]) and their commonly used cutoff values have poor performance in FIMASEM represented by the low coverage of what is considered as "good fit." It is worth noting that these investigated fit indices are primarily proposed for multiple indicator models in which a large component of error exists within the measurement model. It is less reasonable for them to perform in models that are farther away from the confirmatory factor models upon which they were initially validated (West, Taylor, & Wu, 2012). A typical MASEM, in fact, is a path model, or a single indicator model at best. With the absence of the measurement component, commonly used fit indices may not be appropriate. Indices also vary in how they treat various aspects of the model (e.g., model complexity) and data (e.g., sample size) in estimating error. The current cutoff values have been derived from either simulation work based on simpler, multiple indicator SEMs or the general experiences of the developers of the indices (Hu & Bentler, 1998; Lai & Green, 2016; Marsh, Balla, & McDonald, 1988). Taken together, the current fit indices and their cutoff values might be inappropriate for the MASEM context.

Perhaps now is the time to take a new approach to evaluating model fit in MASEM. The key conceptual shift we sought to make was to advocate for a more logically consistent treatment of effect size heterogeneity. In this vein, if an effect is particularly strong for some subpopulations (and particularly weak for other subpopulations), then it follows that a theoretical model might fit some subpopulations particularly well (and others particularly poorly). This represents a meaningful conceptual divergence from how fit indices are used in practice. Rather than asking, "Is the misspecification small enough that I can conclude my model fits?" we encourage researchers to ask, "Based on the patterns of correlations I estimate to exist in the population [i.e., the bootstrapped correlation matrices based on meta-analytic results], does this model fit a small or large portion of the population?"

Yu et al. (2016) initially proposed answering this question by codifying the fit index cutoff and assessing the percent of bootstrapped correlation matrices that cleared that cutoff. We agree with Cheung's (2018) arguments that this approach suffers from the same limitations as the use of fit index cutoffs in practice. In Supplementary Material 3 of the online supplemental materials, we expand on our original recommendations and propose that researchers holistically consider the results from three procedures. First, researchers should produce the summary statistics of the distributions of relevant fit indices across the bootstrapped correlations. Second, researchers should produce a density plot of these fit indices. These first two steps allow researchers to gauge whether the model fit is widespread in the population or isolated to only certain subpopulations. Third, we recommend researchers assess the correlations across fit indices to see if multiple fit indices converge on the same conclusion regarding model fit evaluation.

It should be noted that these recommendations follow the best of our present understanding of MASEM and effect size heterogeneity. Research on this issue is sorely needed. To begin with, we need to understand differences in fit index computations when assessing bootstrapped correlation matrices in order to understand which fit indices are best suited for comparison when using FIMASEM. This would involve searching for alternative ways of evaluating overall fit (i.e., different types of fit indices) customized to FIMASEM. First, FIMASEM fit indices need to account for using correlation matrices instead of covariance matrices. For example, Cheung (2015) introduced a modified version of SRMR to remove the diagonals of the matrices from the formula as they were fixed to 1 in MASEM. Second, FIMASEM fit indices need to consider the role of sample size and model complexity if they are to be used to compare a model across different contexts or to compare different models within a context. In addition, fit indices adjusting for parsimony need to be evaluated to ensure that the alternative model is not better simply because it is more or less complex. Finally, additional simulation work is necessary to set the realistic bounds of fit indices' values by introducing different complexities frequently encountered in MASEM practice.

Future methodological research is also needed on how to handle poorly fitting models in MASEM. We disagree with Cheung's (2018) statement that poorly fitting models indicate a problem with the FIMASEM technique. It seems plausible that most theoretical models have little predictive value for certain subpopulations under very specific conditions. In this sense, we should expect to find samples in which the model fits very poorly with extreme—perhaps even nonsensical—fit index values. It is exactly what one should expect to see in FIMASEM results: that some models fit well in some subpopulations but fit very poorly in other subpopulations. If we were to exclude poor-fitting models, then FIMASEM results would no longer represent heterogeneity but only a summary of good fitting models, which would not be the goal that FIMASEM is serving. However, this is an area for future research to develop more robust statistical theory and computational approaches to address poorly fitting models.

When Is Ignoring Heterogeneity Problematic in MASEM Research?

Because MASEM is a hybrid procedure, researchers make independent assumptions and draw conclusions based on both the Step 1 meta-analysis and the Step 2 SEM. For example, in Step 1, a typical MASEM study first assumes whether the relationship between X-Y in the population is homogenous (e.g., one true score) or heterogeneous (e.g., a distribution), deciding to either conduct a fixed- or a random-effect meta-analysis (Schmidt, Oh, & Hayes, 2009). As a part of the meta-analytical procedure, the actual homogeneity/heterogeneity in a bivariate relationship is tested through Q-statistics (Hedges & Olkin, 1985), credibility interval (CV; Whitener, 1990) width, or the percentage of variance attributable to artifact (Schmidt & Hunter, 2015). In Step 2, the

stage of SEM, researchers also need to make (and test) the assumption of homogeneity before conclusions about SEM results can be drawn. Theoretically, assumptions and conclusions made in both steps are expected to remain logically consistent in order to avoid making inferential mistakes. That is, findings in Step 1 should inform assumptions in Step 2.

When the assumptions about variance across the two steps are inconsistent, two inferential mistakes can result: Type II errors (false negatives) and Simpson's paradox. In the context of MASEM, researchers typically find population heterogeneity at the meta-analysis stage. If heterogeneity is found in the meta-analysis but not assumed or detected in the SEM, a Type II error is made—the path estimates and model fit values are concluded to be invariant when in fact they do differ across the population. The other inferential mistake, Simpson's paradox, describes a situation when erroneous conclusions occur because data drawn from heterogeneous populations are pooled and analyzed as if they were from a single homogenous population (Simpson, 1951). That is, depending on how data are divided up, the true effect can be nullified or even reversed. This paradox arises when a moderator is a confounding factor but is not accounted for in the model (see Lindley & Novick, 1981, for a detailed treatment).

In Table 1, we specify eight scenarios of different assumptions and conclusions about population heterogeneity when conducting MASEM. In two of the eight scenarios (Scenarios 1 and 7), assumptions are consistent at both stages (e.g., fixed effects meta-analysis followed by Viswesvaran & Ones's [1995] MASEM) *and* the correct inference is made (e.g., nonsignificant Q-statistic for correlation and the SEM parameter generalizes). Of 35 MASEM studies published in management journals from 1992 to February 2018,[2] none fit Scenario 1 and 13 fit Scenario 7. Among those in Scenario 7, 10 studies did not quantify the heterogeneity at the output of SEM but either conducted multigroup SEM analysis with a known moderator or controlled for heterogeneity using TSSEM. The remaining three studies used the FIMASEM procedure proposed in Yu et al. (2016).

We now turn to those scenarios in which the assumptions are consistent at both stages *but* the inferences are incorrect. We observed two MASEMs that initially assumed a fixed effect meta-analytic model but then found several bivariate relationships to be heterogeneous. Nevertheless, they still assumed homogeneity at the stage of SEM (i.e., Scenario 2). Although the assumptions were consistent across the meta-analysis and SEM, these two MASEMs might have failed to detect heterogeneity when it is actually present in the population (Type II error and possible Simpson's paradox). A corresponding case for a random-effect model is Scenario 8, in which researchers incorrectly assume heterogeneity in both the meta-analysis and SEM portions of MASEM. Here, researchers assume heterogeneity and correspondingly run a random-effects model in the meta-analysis step but instead find homogeneity. Despite evidence for a fixed-effects model in the meta-analysis, they continue to assume heterogeneity in the SEM. Worth mentioning is that despite being incorrect inferences in both steps of the MASEM, this error is unlikely to affect overall conclusions. This is because running a random effect meta-analysis does not prevent the detection of a fixed effect in the population (Schmidt et al., 2009), nor does running a FIMASEM prevent the detection of parameters that do not vary in the population. No study in our review fit Scenario 8.

The remaining scenarios are those in which the homogeneity/heterogeneity assumptions are inconsistent. In Scenarios 3 and 4, homogeneity is assumed in the meta-analysis but heterogeneity is assumed in SEM. We did not identify any studies that fit either Scenarios 3 or 4. In Scenarios 5 and 6, heterogeneity is assumed in the meta-analysis but homogeneity is assumed in the SEM. Twenty published MASEM studies fit Scenario 5, in which researchers first conduct a random effect meta-analysis, detect heterogeneity in the correlations, but then assume those same correlations are homogenous at the stage of SEM. Studies of this kind are likely to make Type II errors and may be vulnerable to Simpson's paradox. Scenario 6 is a variant of Scenario 5 in that studies of this kind begin with a random effect meta-analysis but discover evidence for assuming homogeneity in SEM. In Scenario 6, researchers are inconsistent in assumptions between Stage 1 and Stage 2 but make correct inference when homogeneity is detected in the result of Stage 1 and assumed at Stage 2.

Overall, out of eight possible scenarios of conducting MASEM research, only in two scenarios (Scenarios 2 and 5) can researchers potentially make inferential mistakes. However, it is unfortunate that the majority of published MASEMs ($n = 22$; 63%) have fallen into these two categories. The danger of making such inferential errors is either the true scores or distributions in the population are undetected or erroneous conclusions about model robustness are made.

When Is Quantifying Effect Size Heterogeneity in MASEM Useful?

FIMASEM provides a framework to make correct inferences in MASEM research, as it is one of only a few approaches to statistically incorporate effect size heterogeneity in MASEM. Alternative techniques seeking to account for the heterogeneity in the stage of SEM involve conducting multigroup SEM analysis for a certain moderator and using WLS estimation in TSSEM. However, only FIMASEM *quantifies* effect size heterogeneity. This feature of FIMASEM highlights a different perspective of conducting MASEM research and one that presents a significant departure from traditional MASEM approaches—especially in terms of how hypotheses could be formulated and how results of MASEM should be interpreted. When the relationships are indeed homogeneous, FIMASEM can be used as the supplementary analysis to the traditional MASEM model as a validation that accurate inferences are made. For example, Greer, de Jong, Schouten, and Dannals (2018) used FIMASEM to supplement their fixed effect model to demonstrate that their assumption of homogeneity in running a fixed effect model was warranted. Nevertheless, the primary intention of using FIMASEM is not to evaluate *if* the relationship is significant from zero but to examine the distribution of true effect size estimates in the population. This moves applying FIMASEM beyond testing the null hypothesis of a relationship or a mediating effect. If a MASEM researcher's goal is only to reject a null hypothesis (i.e., if the X-Y relationship differs from zero holding all other relationships constant), FIMASEM is of a limited utility and this could be just as easily accomplished with TSSEM.

[2] See Supplementary Material 4 of the online supplemental materials for our search and coding methodology.

Table 1
Eight Scenarios, Assumptions, and Inferences of MASEM

Scenarios	Assumption of Step 1 meta-analysis	Conclusion of Step 1 meta-analysis	Assumption of Stage 2 SEM	Conclusion of Step 2 SEM	Assumption consistency	Inferential mistakes	Number of MASEM studies
Scenario 1	Assumes a single, true score of the X-Y relation in population. Thus, fixed effects model performed.	Confirm homogeneity (e.g., nonsignificant Q-statistic).	Assume the population is homogeneous	The path coefficient of X-Y is the single, true value in the population.	Consistent assumption at both steps	Correct inference	0
Scenario 2		Instead, find heterogeneity (e.g., large, significant Q-statistic).			Consistent assumption at both steps	**Type II error; Simpson's paradox**	**2**
Scenario 3		Confirm homogeneity (e.g., nonsignificant Q-statistic).	Assume the population is heterogeneous	The path coefficient of X-Y varies across populations.	Inconsistent assumption from Step 1 to Step 2	Correct inference	0
Scenario 4		Instead, find heterogeneity (e.g., large, significant Q-statistic).			Inconsistent assumption from Step 1 to Step 2	Correct inference	0
Scenario 5	Assumes a distribution of true scores in population. Thus, random effects model performed.	Confirm heterogeneity (e.g., wide credibility interval).	Assume the population is homogeneous	The path coefficient of X-Y is the single, true value in the population.	Inconsistent assumption from Step 1 to Step 2	**Type II error; Simpson's paradox**	**20**
Scenario 6		Instead, find homogeneity (e.g., SD-rho close to zero).			Inconsistent assumption from Step 1 to Step 2	Correct inference	0
Scenario 7		Confirm heterogeneity (e.g., wide credibility interval).	Assume the population is heterogeneous	The path coefficient of X-Y varies across populations.	Consistent assumption at both steps	Correct inference	13
Scenario 8		Instead, find homogeneity (e.g., *SD*-rho close to zero).			Consistent assumption at both steps	Correct inference	0

Note. MASEM = Meta-Analytic Structural Equation Modeling; SEM = Structural equation modeling. Bold data indicates the inferential mistakes people make.

The true utility of FIMASEM can be extended to examine and interpret heterogeneity in three ways. First, FIMASEM provides quantitative measures for assessing the overall generalizability of a theoretical model to the population and sheds light on where in the model (i.e., which particular structural path) a moderator(s) may potentially exist. In other words, FIMASEM can address such research questions as (a) To what degree (what percentage of the population) does this theoretical mechanism/model apply? and (b) Which of the structural paths from the model may meaningfully fluctuate in the population? For example, Lapierre et al. (2017) used FIMASEM to evaluate the generalizability of the proposed model as well as individual structural paths through the distributions of SEM fit indices and the credibility intervals of path coefficients. They concluded that "the hypothesized model . . . is largely generalizable across populations" (p. 11) with 83.4% or 85% TLI above .90 and 80.7% or 80% RMSEA smaller than or equal to .08. Despite the overall generalizability of their proposed model, they identified substantial variation that exists in some structural paths of the model. For example, the 80% CV from social support at work to family work enrichment is [−.30, .24], with a mean of −.03. If the CV were not present, one could have concluded that there was no direct path from social support at work to family work enrichment. Instead, this CV suggests that the relationship can range from moderately negative to moderately positive across subpopulations. Another study by Goering, Shimazu, Zhou, Wada, and Sakai (2017) comprehensively interpreted the FIMASEM distribution of effect sizes. In doing so, they detected variability in the paths of the conventional job demand-resource model, with some holding relatively constant in the population, whereas others demonstrated significant variation.

Second, researchers can use FIMASEM to quantify the robustness of a theoretical model across different contexts or given a particular theoretical/methodological moderator. In other words, FIMASEM can answer the research question "To what extent is the proposed theoretical model more robust in Context A than in Context B?" This can be accomplished if the ρ and *SD*ρ matrices of a MASEM study are separately constructed by a moderator of interest (e.g., male vs. female groups). Our review of published MASEM studies revealed that it is not uncommon (nine MASEMs; 25.7%) to find sufficient data broken down by at least one moderator. Researchers can compare the different FIMASEM models in terms of whether there is a substantial change in the percentage of model fit or substantial differences in path estimates. Interested readers can refer to Supplementary Material 5 of the online supplemental materials for an example of a reanalysis of Hom, Caranikas-Walker, Prussia, and Griffeth (1992). Overall, FIMASEM not only concludes if a theoretical model works differently *across* contexts but also informs the degree of generalizability of the theoretical model *within* these contexts.

Third, FIMASEM can be used to evaluate different theoretical explanations of the same phenomenon or context (e.g., "To what extent is Model A more robust/generalizable than Model B in the same research context?"). To answer this type of research question, researchers can specify two different models based on competing theories and perform independent FIMASEM analysis on each model based on sample matrices generated from the same population (i.e., same set of ρ and *SD*ρ matrices). FIMASEM researchers can quantitatively present the degree of generalizability of each model and conclude which model has a higher or lower generalizability/robustness in the same research context by comparing the percentages of model fit across different model specifications.

References

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. http://dx.doi.org/10.1002/smj.2338

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

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

Cheung, M. W.-L., & Chan, W. (2009). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling, 16,* 28–53. http://dx.doi.org/10.1080/10705510802561295

Cheung, M. W.-L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology, 37,* 522–541. http://dx.doi.org/10.1177/0022022106290476

Goering, D. D., Shimazu, A., Zhou, F., Wada, T., & Sakai, R. (2017). Not if, but how they differ: A meta-analytic test of the nomological networks of burnout and engagement. *Burnout Research, 5,* 21–34. http://dx.doi.org/10.1016/j.burn.2017.05.003

Greer, L. L., de Jong, B. A., Schouten, M. E., & Dannals, J. E. (2018). Why and when hierarchy impacts team effectiveness: A meta-analytic integration. *Journal of Applied Psychology*. Advance online publication. http://dx.doi.org/10.1037/apl0000291

Hedges, L. V., & Olkin, O. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hom, P. W., Caranikas-Walker, F., Prussia, G. E., & Griffeth, R. W. (1992). A meta-analytic 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

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. http://dx.doi.org/10.1037/1082-989X.3.4.424

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51,* 220–239. http://dx.doi.org/10.1080/00273171.2015.1134306

Lance, C. E., & Vandenberg, R. J. (2009). *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. New York, NY: Taylor & Francis.

Lapierre, L. M., Li, Y., Kwan, H. K., Greenhaus, J. H., DiRenzo, M. S., & Shao, P. (2017). A meta-analysis of the antecedents of work–family enrichment. *Journal of Organizational Behavior*. Advance online publication. http://dx.doi.org/10.1002/job.2234

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *Annals of Statistics, 9,* 45–58. http://dx.doi.org/10.1214/aos/1176345331

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410. http://dx.doi.org/10.1037/0033-2909.103.3.391

O'Boyle, E. H. (2017). Validity generalization as a continuum. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10,* 485–488. http://dx.doi.org/10.1017/iop.2017.46

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology, 96,* 1–12. http://dx.doi.org/10.1037/a0020539

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781483398105

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97–128. http://dx.doi.org/10.1348/000711007X255327

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B. Methodological, 13,* 238–241. Retrieved from http://www.jstor.org/stable/2984065

Steel, P. D., & Kammeyer-Mueller, J. (2008). Bayesian variance estimation for meta-analysis: Quantifying our uncertainty. *Organizational Research Methods, 11,* 54–78. http://dx.doi.org/10.1177/1094428107300339

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

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. J. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

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

Williams, L. J., & O'Boyle, E. H., Jr. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods, 14,* 350–369. http://dx.doi.org/10.1177/1094428110391472

Williams, L. J., O'Boyle, E. H., & Yu, J. (2017). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives. *Organizational Research Methods*. Advance online publication. http://dx.doi.org/10.1177/1094428117736137

Worthke, W. (1993). Nonpositive definite matrices in structural equation modelling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.

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

Received December 10, 2017
Revision received April 30, 2018
Accepted May 2, 2018 ■