

Addressing Endogeneity in Meta-Analysis: Instrumental Variable Based Meta-Analytic Structural Equation Modeling

Zijun Ke

Sun Yat-sen University

Yucheng Zhang 

University of Southampton

Zhongwei Hou

Northwest Normal University

Michael J. Zyphur

University of Queensland

In management research, meta-analysis is often used to aggregate findings from observational studies that lack random assignment to predictors (e.g., surveys), which may pose challenges in making accurate inferences due to the correlational nature of effect sizes. To improve inferential accuracy, we show how instrumental variable (IV) methods can be integrated into meta-analysis to help researchers obtain unbiased estimates. Our IV-based meta-analytic structural equation modeling (IV-MASEM) method relies on the fact that IVs can be incorporated into SEM, and meta-analytic effect sizes from correlational research can be used for MASEM. Conveniently, IV-MASEM does not require that each primary study measures all relevant variables, and it can address typical types of endogeneity, such as omitted variable bias. We clarify how the principles of IV-SEM can be applied to MASEM and then conduct three simulations to study the validity of IV-MASEM versus Univariate Meta-Analyses (UMA) and MASEMs that exclude IVs when the instruments were appropriate, inappropriate, and missing from a subset of primary

Acknowledgment: The Zijun Ke and Yucheng Zhang are co-first authors due to their equal contribution to the work presented in this paper, who are listed alphabetically. This research was supported by National Natural Science Foundation of China (Grant No. 31700986, 72343035, and 72272048) and Guangdong Basic and Applied Basic Research Foundation (Grant No.2022A1515011986).

Supplemental material for this article is available with the manuscript on the JOM website.

Corresponding author: Yucheng Zhang, Southampton Business School, University of Southampton, University Road, Southampton, SO17 1BJ, UK.

E-mail: yucheng.eason.zhang@gmail.com

studies. We also offer an illustrative study to demonstrate how to apply IV-MASEM to address endogeneity concerns in meta-analysis, which includes a new R function to test the qualifying conditions for IVs. We conclude with limitations and future directions for IV-MASEM.

Keywords: *endogeneity in meta-analysis; instrumental variable; correlational effect sizes*

Meta-analysis (MA) represents a robust methodological approach that enables researchers to synthesize and integrate primary research findings across multiple studies on shared topics. By doing so, it serves to validate and advance management theories while guiding organizational practices (Cooper, Hedges, & Valentine, 2009; Hedges & Olkin, 2014; Hunter & Schmidt, 2004). Nonetheless, traditional meta-analyses have primarily concentrated on effect size synthesis. By using Univariate Meta-Analysis (UMA), developed by Hedges and Olkin (2014) and Hunter and Schmidt (2004), most previous meta-analyses have focused on examining the relationship between two variables. This conventional meta-analytic paradigm, however, is constrained in its capacity to delve into the mechanisms governing variable relationships, such as mediation effects when the model encompasses more than three variables (Bergh et al., 2016). Accordingly, the Meta-Analytic Structural Equation Model (MASEM) was developed as an integrative method that incorporates SEM into meta-analysis (Becker, 2009; Cheung & Cheung, 2016). MASEM allows researchers to conduct complex multivariate analysis (e.g., mediation tests) using meta-analysis (Bergh et al., 2016; Bergh, Boyd, Byron, Gove, & Ketchen, 2022), and it has witnessed a growing prevalence in management research (e.g., Hancock, Allen, Bosco, McDaniel, & Pierce, 2013; Zhang, Liu, Xu, Yang, & Bednall, 2019).

However, in both of UMA and MASEM research, endogeneity issues (e.g., those arising from omitted variables and reverse causality) are often overlooked by researchers. This oversight can be attributed to the use of UMA and MASEM in fields such as psychology, where experimental primary studies are more likely to be found, and thus the effect sizes included in UMA and MASEM are more likely to be unbiased. In contrast, effect sizes derived from survey data or archival data are more significantly affected by endogeneity, and therefore UMA and MASEM are more likely to produce estimates that are biased due to endogeneity (Bergh et al., 2016). This issue is especially relevant for management research, where non-experimental data are commonly used (Antonakis, Bendahan, Jacquart, & Lalive, 2010). Hill, Johnson, Greco, O'Boyle, and Walter (2021) likened endogeneity to a "disease" that affects empirical research in management. We argue that this "disease" is highly contagious, allowing endogeneity issues to transfer from primary studies to any meta-analysis reliant on them. If unaddressed, endogeneity can significantly influence the estimation based on UMA and MASEM. As Semadeni, Withers, and Certo (2014) noted, even minimal levels of endogeneity can increase the likelihood of researchers committing Type-I errors.

Endogeneity manifests in multiple ways in UMA and MASEM research. First, effect sizes from initial studies are often biased due to many primary studies not addressing endogeneity by design. Second, modeling based on UMA and MASEM often omit potentially important control variables, resulting in biased estimation. Specifically, given its univariate nature, UMA cannot address omitted-variable issues. Although MASEM can include control variables for addressing omitted variables, researchers rarely can ensure all omitted variables are

included. Hünermann and Louw (2023) and Mändli and Rönkkö (2023) comprehensively clarify the use of control variables for enhancing causal inference. In the context of MASEM, Jak and Cheung (2018) and Furlow and Beretvas (2005) identified reasons why MASEM studies cannot eliminate all relevant omitted variables, including variables beyond the scope of the study, novel variables lacking operationalization, and the decision to not report variables that do not reflect expected benefits. Despite the fact that exhaustive inclusion of control variables in MASEM is possible in certain contexts, it is difficult to ensure that all control variables are also exogenous, which might lead to greater biases. Indeed, over-controlling for variables to address omitted variables can harm the accuracy of inferences (Darlington & Hayes, 2017: 538). Third, MASEM research may be subject to additional endogeneity concerns, such as those inherent within the data generation model itself. For instance, relationships between variables in the model could be affected by reverse causation, which typical MASEM cannot handle.

In the realm of management studies, the impact of endogeneity on research results has emerged as a critical concern. Researchers have actively engaged in developing and using methods that address endogeneity in non-experimental studies (see Antonakis et al., 2010). The IV method features prominently as a notable strategy in this regard, having gained traction for its efficacy in mitigating endogeneity (see Angrist & Krueger, 2001; Didelez, Meng, & Sheehan, 2010). Initially prevalent in economics, the IV method is now gaining traction in management, psychology, and related fields. A pivotal advancement in this area is the innovative integration of the IV approach with SEM, which is suggested by Maydeu-Olivares, Shi, and Rosseel (2019) and Maydeu-Olivares, Shi, and Fairchild (2020), offering a robust framework for estimating unbiased effects using SEM. However, a systematic method to specifically address endogeneity in meta-analysis remains elusive in management research, highlighting a gap that warrants urgent scholarly attention.

Given the similarity of MASEM and SEM, we suggest that IV-SEM has potential to address endogeneity in meta-analyses. However, the integration of IV-SEM with MASEM and the efficacy of IV-SEM to resolve endogeneity in a MASEM framework have been overlooked. Considering the substantial potential of IV methods to address endogeneity, we introduce IV-MASEM as an innovative application of the IV-SEM method in MASEM, showing it is valid for analyzing effect-sizes data (e.g., correlations) and that it can be applied with various MASEM techniques—allowing IV-MASEM to both synthesize research and estimate unbiased effects.

Our paper makes key contributions by first reviewing why IV-SEM can be integrated into MASEM as a solution for solving endogeneity in meta-analysis. Second, it details the IV-MASEM method, explaining how IV-SEM can be incorporated into MASEM to reduce bias in both UMA and MASEM estimates. Third, considering statistical validation, we conduct three comparative Monte Carlo simulations to assess the estimation effectiveness of IV-MASEM, UMA, and MASEM in various scenarios. Finally, an illustrated study provides evidence for the efficacy of IV-MASEM compared to UMA and MASEM in actual applications. To supplement our method, we introduce a novel R function for testing weak IVs based on IV-MASEM results. This is critical because IV-MASEM estimates should only be used when IVs are effective. We also cover three meta-analytic procedures to demonstrate how to apply IV-MASEM using different MASEM techniques, which is available in our online supplemental materials. To achieve these objectives, we first outline the rationale for integrating IV-SEM into MASEM research.

Integrating Instrumental Variable Methods With MASEM

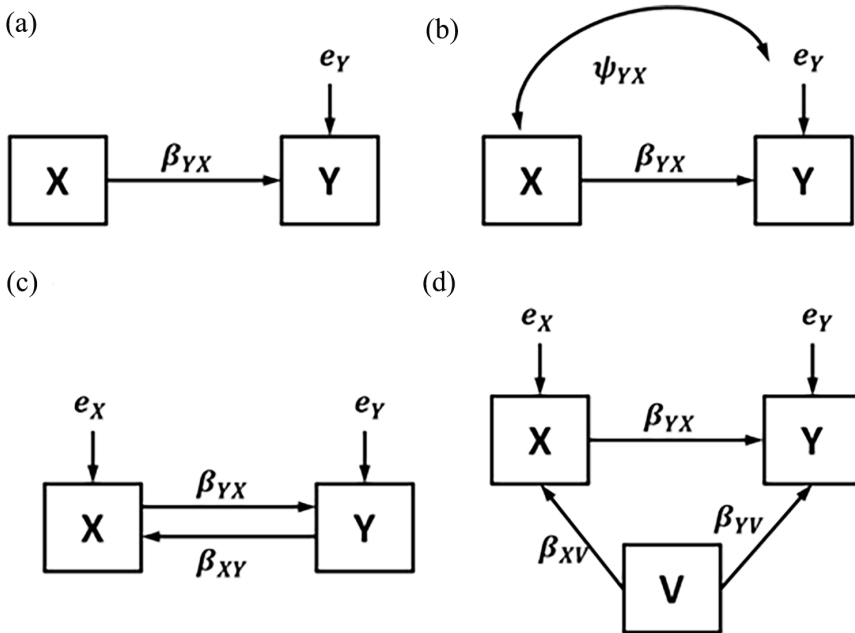
The idea of integrating the instrumental variable (IV) method into MASEM research to solve the problem of endogeneity caused by omitted variables and reverse causality is mainly due to recent research in management and psychology that models IVs in a SEM framework. Specifically, Maydeu-Olivares et al. (2020) and Maydeu-Olivares et al. (2019) have proposed IV-SEM as an alternative to Instrumental Variables Regression (IVR), which is a prevalent method within the econometrics field. This method uses a SEM perspective to show how IVs can be added to the theoretical or “structural” model of a SEM to obtain unbiased estimates.

The IV-SEM method aims to incorporate IVs in a theoretical model to achieve unbiased estimation of the relationship between variables X and Y . An unbiased estimate of an effect of X on Y (see Figure 1a) cannot be estimated using bivariate correlations in the presence of endogeneity (Aldrich, 1995; Antonakis et al., 2010). In a formal sense, endogeneity occurs when a predictor X is endogenous, causing it to be correlated with what should be estimated as the residual of Y (see Figure 1b). When this occurs, the estimate X 's effect includes the correlation of X 's endogenous component with the error of Y , biasing the effect estimate. To give a quick practical example, if X is job satisfaction and Y is job performance, the estimated effect of X may appear to be positive. However, this could be due to reverse causation if higher job performance Y leads to more job satisfaction X . In this case, Y 's residual should include that part of it which causes job performance X , but because X is endogenous the true error of Y is correlated with X , which in turn biases the empirically estimated effect of X on Y (because it also includes the reverse-causal effect of Y on X).

Antonakis et al. (2010) and Bollen (2012) suggest the most common forms of endogeneity are omitted variables and reciprocal causation (see Figures 1c and d). The IV method addresses endogeneity by adding IVs to fitted models as in Figure 1b to ensure consistent estimation (Bollen, 2012; Maydeu-Olivares et al., 2020). Antonakis et al. (2010) and Maydeu-Olivares et al. (2019) validated and demonstrated how to integrate IVs into SEM (IV-SEM). For using IV-SEM, it is necessary to employ appropriate IVs that meet two critical conditions, including exclusion and relevance (Antonakis et al., 2010; Bollen, 2012). As shown in Figure 2, exclusion requires IVs (e.g., Z_1 and Z_2 in Figure 2) to be uncorrelated with the error of Y and have no direct effect on Y (e.g., Bollen, 2012). The relevance condition is met when IVs are correlated with X (Bollen, 2012). Specifically, IV models like that in Figure 2 can be estimated as a SEM in a single step, rather than two-stage IVR which historically has been common (Maydeu-Olivares et al., 2019; Maydeu-Olivares et al., 2020), and has been used in management research since the 1990s (e.g., Frone, Russell, & Cooper, 1994).

The minimum number of IVs is equal to the number of endogenous predictors X (Maydeu-Olivares et al., 2020: 41). Such a model is just-identified and the parameters can be estimated, but without positive degrees of freedom the model's fit cannot be assessed. Thus, Maydeu-Olivares et al. (2019) recommended that at least one more IV should be included than X . Additionally, all paths between IVs and Y should be fixed to zero for the exclusion condition. Assumption checking is necessary to validate model specification, which is critical to examine whether the assumptions of IV models are violated—exclusion and relevance (Table 1 summarizes our notation; see also Bollen, 1989; for a general treatment see Maydeu-Olivares et al., 2020). Exclusion can be checked based on model fit using the chi-square test or modification indices. Relevance can be tested through a t test or F test of individual or

Figure 1
Path Diagrams of Possible Relationships Between Two Variables

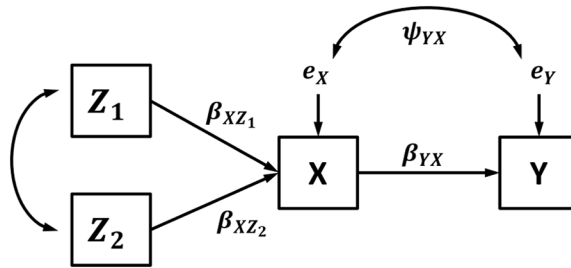


- (a) Standard model.
 (b) Correlated errors.
 (c) Reciprocal associations.
 (d) Omitted variables.

joint significance of the paths from IVs to predictors (or the z test and Wald test in large samples), respectively, and R^2 can be used to measure the strength of the associations. If the exclusion and relevance conditions are satisfied, then we recommend that it is appropriate to include these IVs in the model (Maydeu-Olivares et al., 2020: 246). In this case, IV-SEM ensures consistency in estimating the effect of X on Y, and it is valid to draw accurate inferences based on null hypothesis significance testing with p values.

From the perspective of SEM, incorporating IVs into a SEM's structural model is akin to adding several variables with the restrictive condition that they only predict X. However, the benefits of incorporating IVs into SEM are manifold. It solves the endogeneity problem in management research in a relatively concise and efficient way, without any notable increases to computational difficulty. In addition to some stable individual factors that can be used as IV, such as personality traits which are exogenous (Antonakis et al., 2010), individuals may encounter unpredictable exogenous events at any time in their daily work, such as encountering a traffic jam or an unexpected work assignment. These unexpected events may be used as IVs when they impact proximal variables such as emotions (X), which, in turn, can impact more distal outcome variables such as job performance (Y). In this way, random events

Figure 2
Model Identification in an IV-SEM Framework



encountered at work can be modeled as IVs by having effects on Y that are fully mediated by predictors of interest X (as in Figure 2).

By integrating IVs into a SEM, unbiased parameter estimates can be obtained. Given the common characteristics of MASEM and SEM research, integrating the IV method into MASEM is feasible and can enhance the robustness of meta-analytic estimates. As in SEM, MASEM can use correlation matrices (or covariance matrices) as data inputs for analyzing relationships among multiple variables, and fit indices can be used to decide whether to reject the corresponding theoretical model. Conveniently, MASEM typically benefits from larger sample sizes in meta-analysis, leading to more stable and reliable parameter estimates and fit statistics (Landis, 2013). However, there are some notable differences between SEM and MASEM.

First, the data types are different for these two methods. The covariance matrix applicable to SEM research is generated based on an original dataset, where all individual cases are typically observed and hierarchical structure is conventionally not considered. Error in a SEM study typically comes from sampling or measurement error, such that bias caused by endogeneity exists in the original data. The correlation matrices used in MASEM are often aggregated based on the effect sizes in different primary studies. Hierarchical structures that distinguish within- and between-study effects are typically considered. The within-study errors in such studies can come from sampling error or from a range of endogeneity problems in the primary studies, but bias may also come from differences at the between-study level, such as differences in the country or industry from which primary data were drawn.

Second, there is the problem of the source of heterogeneity. In SEM, it is typically valid to assume that modeled effects are homogeneous in the sample being studied. However, in MASEM, between-study heterogeneity is typically assumed as a possibility that should be investigated. This is directly reflected in the models of MASEM, which use random effects for pooling correlation coefficients across studies. This approach is particularly effective for accommodating between-study heterogeneity when primary studies are collected from different populations (e.g., data from employees and data from senior managers).

Finally, there is the question of assumption checking for using IVs, because existing methods of assumption checking on IVs by using SEM are all based on primary data. However, MASEM relies on effect sizes rather than original data. Additionally, heterogeneity could exist in the degree to which IV assumptions are violated. Both factors raise

Table 1
Notation Used in IV Models

Notation	Definition
Y	Dependent variable
X	Endogenous predictor of Y that is correlated with the error of Y
Z	Instrumental variable (IV) having a direct effect on X but no direct effect on Y
$\beta_{YX}; \beta_{YZ}$	Standardized effect of X on Y; effect of Z on X

questions about whether existing methods for testing the IV conditions of exclusion and relevance are valid for MASEM research. Below, we begin by elaborating how IV methods can be integrated into MASEM research and how to test two IV assumptions in the context of MASEM. Additionally, we rely on three simulations and an illustrative study to validate IV-MASEM.

Applying IV Methods to MASEM

MASEM is a valuable approach for conducting multivariate analysis using effect sizes. It goes beyond UMA by examining relationships among multiple variables. There are three main MASEM approaches in management research: univariate MASEM (U-MASEM, Viswesvaran & Ones, 1995); full-information MASEM (FI-MASEM, Yu, Downes, Carter, & O’Boyle, 2016); and one-stage MASEM (OS-MASEM, Jak & Cheung, 2020). These methods pool correlation matrices and estimate parameters differently. Specifically, U-MASEM is widely used in management research, but it does not quantify between-study heterogeneity. FI-MASEM overcomes this limitation by allowing researchers to quantify path-estimate heterogeneity for evaluating generalizability and detecting potential boundary conditions (Yu, Downes, Carter, & O’Boyle, 2018: 810). In contrast to U-MASEM and FI-MASEM, which rely on meta-analytic effect sizes as data, OS-MASEM uses correlation matrices from primary studies as data inputs (Jak & Cheung, 2020).¹ This approach addresses the shortcomings of MASEM that are based on meta-analytic effect sizes, such as using a single value for the study sample sizes.

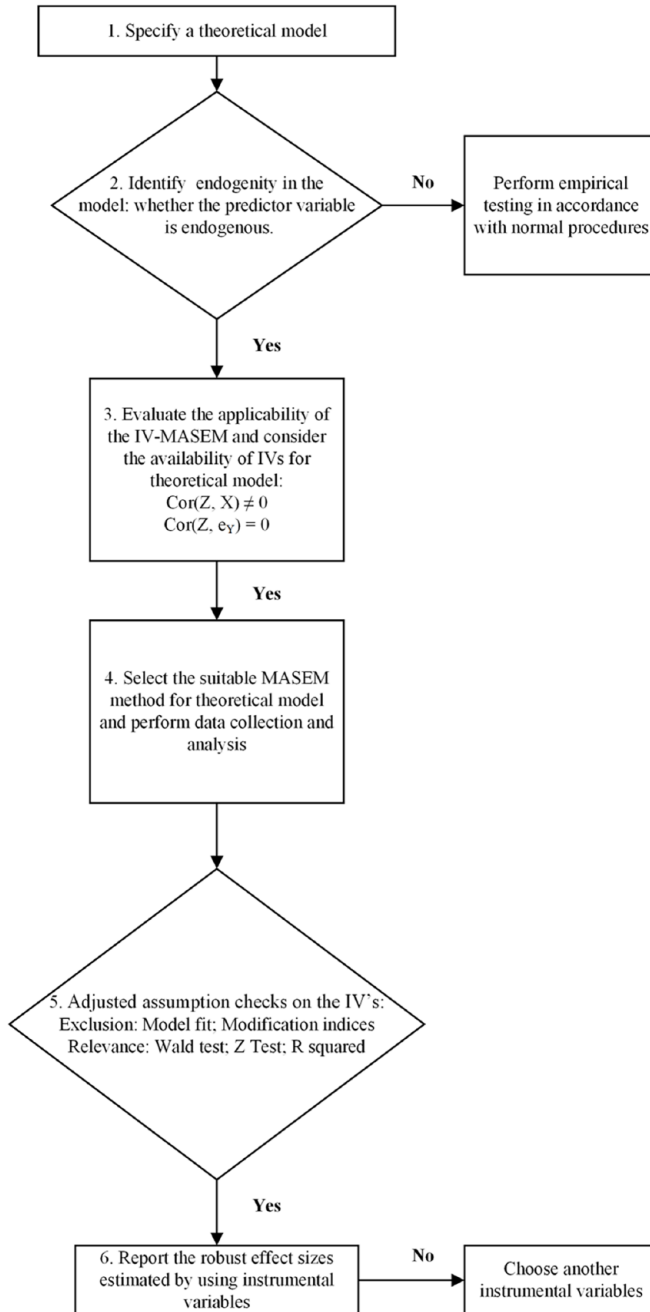
All MASEM methods are equipped for multivariate analysis, irrespective of whether the primary studies are experimental or observational. In scenarios where the primary studies are based on experimental designs, MASEM remains unaffected by endogeneity. Conversely, in cases where primary studies are observational, they are inherently susceptible to issues of endogeneity, which must not be overlooked. In essence, using experimental data in MASEM research inherently avoids endogeneity issues through its research design. However, for non-experimental data, researchers cannot address endogeneity at the research design stage and must manage it during the modeling process. IV-MASEM offers a potential solution to address endogeneity at the modeling stage by incorporating the features of IV-SEM into the MASEM method. This makes the IV approach suitable for different MASEM methods, which can then solve the endogeneity problems caused by primary studies, such as omitted variables and reverse causality. Although each MASEM method has different ways of pooling correlation matrices for data preparation, the procedure for integrating MASEM and IV methods is consistent.

As shown in Figure 3, the IV method can be integrated with each MASEM method based on six steps. Conveniently, the first three steps are the same for all three methods. First, a researcher needs to specify a theoretical model that includes all relationships between independent and dependent variables, as is typical in SEM. Second, it is necessary to determine whether the model has endogeneity problems given the data that may bias parameter estimates (e.g., omitted variables or reverse causality). If the researcher can conclude that there is no endogeneity problem, then it is reasonable to use a typical MASEM to examine the relationships between the variables without using IV-MASEM. If there is an endogeneity issue, however, then in the third step researchers can select appropriate IVs for the endogenous predictors (Step 3, Figure 3), or otherwise try to estimate relationships using effect sizes from experimental studies or effect sizes from primary studies that have been corrected for endogeneity issues. For the former, as we noted, any IVs used need to satisfy two conditions: relevance, that is, $\text{Cor}(Z, X) \neq 0$, and exclusion, that is, $\text{Cor}(Z, e_Y) = 0$.

In the fourth step, researchers must choose a MASEM method according to their data type (i.e., using primary or aggregated effect sizes) and research question (i.e., whether they need to model path-coefficient heterogeneity). Researchers must choose the appropriate MASEM method based on their research questions, design, and data collection difficulty. Each method has advantages and disadvantages. The primary advantage of the U-MASEM method lies in its streamlined analytical procedure utilizing aggregated matrices as data input. However, a limitation of this method is its approach to treating effect sizes as independent, and disregarding heterogeneity. If researchers do not need to consider the dependency among different effect sizes and heterogeneity across studies, they may opt for this method. Interested readers can refer to Viswesvaran and Ones (1995) and Bergh et al. (2016) for more detailed information on this approach. In contrast, the FI-MASEM method's strength lies in quantifying the heterogeneity of effect sizes across different studies. FI-MASEM requires not only the collection of aggregated effect sizes but also the collection of indices of heterogeneity. When heterogeneity of effect sizes in meta-analysis is significant, FI-MASEM is effective to estimate the distribution of effects among variables of interest. Interested readers can refer to Yu et al. (2016) and Yu et al. (2018) for detailed guidance of analytical procedure. OS-MASEM offers the advantage of modeling primary effect sizes, which is important for accommodating missing data and effectively assessing moderating effects. It assumes a hierarchical structure of the data (see Jak & Cheung, 2020). Overall, U-MASEM and FI-MASEM are suitable for aggregated effect sizes, while OS-MASEM is suited for primary effect sizes. Researchers should select a method based on the specific research question and the availability of data.

In the fifth step, three procedures for IV-SEM should be applied in MASEM, including (a) model specification, (b) assumption checking, and (c) statistical inference. For model specification, a target model is specified, including IVs (as in IV-SEM), but used with a between-study covariance matrix. Following this, the exclusion condition for IVs is checked by evaluating the overall fit of the IV model. This checks whether IVs are exogenous. Generally, the number of IVs is one more than the number of endogenous predictor variables in the model, otherwise there are no degrees of freedom to detect possible violations of the exclusion condition. This is different from econometrics, where it is often sufficient to have an equal number of instrumental and predictive variables. Next, researchers need to check whether IVs meet the relevance condition. Researchers could test the paths from IVs to the predictor, individually and jointly, using the z test and Wald test (only for one endogenous

Figure 3
Steps of Implementing IV-MASEM



predictor), respectively; and computing R^2 measures for the relationships between Z and X , which indicates the effect size of the relationship among the IVs and the endogenous predictor. Appendix A (in the online supplemental material) offers details on calculating these R^2 measures using results from the three MASEM approaches, and we have written an R function to automate this for researchers, which has been uploaded to OSF as Supplemental Material.

If the assumption check shows IVs do not meet the two conditions, researchers need to go back to the third step to select other IVs, which may require collecting additional data and returning to the subsequent steps. If the assumption check shows IVs meet the two conditions, after finishing the assumption check, the robust effect can be estimated as the mean value β_{YX} , with assessments of “practical significance” using standardized values. In this case, the effects estimated by models can be reported as the meta-analytic estimates. Also, confidence interval estimates provide a way to assess the “reliability” of the effect estimate—a wider interval indicates a less reliable estimate of the effect.

In sum, with this approach, IV-MASEM can help solve the endogeneity problem faced in MASEM research in a relatively efficient way and obtain robust estimates of parameters, which further expands existing MASEM research and researchers’ ability to test theories (see Appendix B in the online supplemental materials for a summary of the differences between MASEM and IV-MASEM). To support this point, in what follows we offer three comparative simulations to show the effectiveness of IV-MASEM in comparison with UMA and MASEM across different research contexts. In addition, for meta-analysts who are interested in using IV-MASEM following the steps suggested above, we provide an illustrative study that includes procedures and results of the IV-MASEM methods for analyzing a real meta-analytic dataset. Data, annotated R code, and relevant results of both our simulation study and illustrative study are available as supplemental materials, which can be downloaded from OSF (downloaded link: https://osf.io/k4bv6/?view_only=0d5d5e6273b74665bcd6d4d046998be7).

Comparative Simulation Study Using IV-MASEM

We conducted three simulations across different scenarios to compare IV-MASEM with MA techniques excluding IVs (i.e., UMA and MASEM) in estimating and testing the strength of a relationship (i.e., β_{YX}) in the presence of endogeneity. Specifically, Simulation 1 evaluated the ideal situations for IV-MASEM when the relevance and the exclusion assumptions were satisfied. In this situation, IV-MASEM is based on a properly specified model whereas MA techniques excluding IVs (i.e., UMA and MASEM) are not. The objective is to demonstrate that in addition to a better model fit, the target effect β_{YX} would be less biased, and the Type-I error rates to test β_{YX} would be better controlled with IV-MASEM than with UMA or MASEM. Simulation 2 assessed the adverse instances for IV-MASEM where IVs were weakly relevant or endogenous. Note that when IVs are weakly relevant or endogenous, IV-MASEM may not be based on a properly specified model and it is unclear whether IV-MASEM could still outperform UMA and MASEM. Simulation 3 assessed the impact of missing IVs on the estimation and testing of the target effect.² In the rest of this section, we first describe the research design in each simulation and then present the results and summarize findings for each simulation.

Simulation 1: Comparison of IV-MASEM With MA Techniques Excluding IVs When IVs Are Appropriate

Simulation Design

Objectives. The objective of this simulation is to compare IV-MASEM based on relevant and exogenous IVs with MA techniques excluding IVs (i.e., UMA and MASEM) in estimating and testing the target effect β_{YX} in the presence of endogeneity.

Data-generating models. We generated data based on the two models in Figures 4a and 4b. In both models, IVs, Z_1 and Z_2 , were exogenous and relevant in that they were not correlated with the error of the outcome Y , had no direct effects on the outcome, and were strongly correlated with the endogenous predictor X . To create endogeneity, we included a confounding variable V that simultaneously affected both the predictor X and the outcome Y in the model in Figure 4a and we considered a reverse-causal effect from Y to X in the model in Figure 4b.

Estimation methods. The fitted model for IV-MASEM, as shown in Figure 2, was a standard IV-MASEM model with two IVs and the X - Y residual correlation. It did not contain the confounder V or the reverse path from Y to X , which were present in the two data-generating models. For comparison, we used the UMA and MASEM approaches to estimate and test the X - Y relationship without IVs.

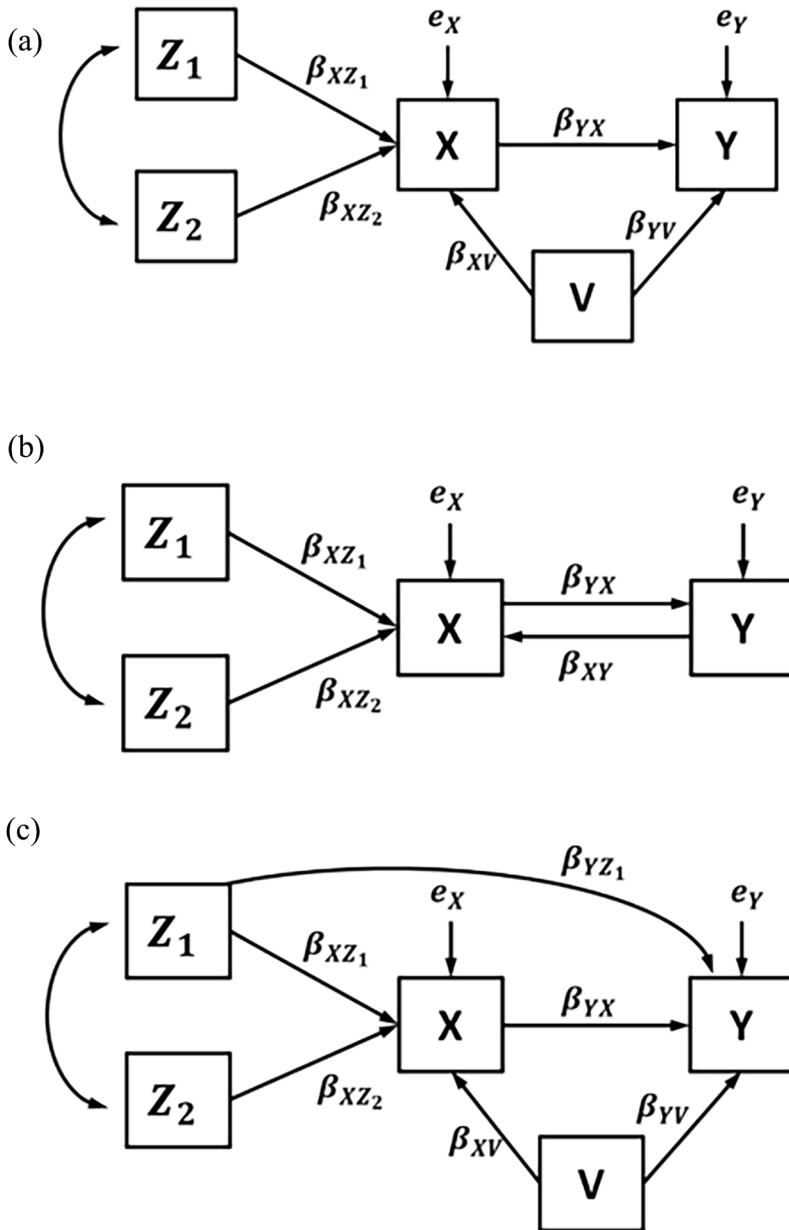
Manipulated factors and parameter values. We considered the following experimental conditions and assigned the following parameter values and sample size settings to the data-generating models in Figures 4a and 4b:

- (a) Two types of endogeneity: an omitted variable model in Figure 4a with meta-analytic slopes from V to X and to Y fixed at $\beta_{XV} = \beta_{YV} = \sqrt{.20}$ and a reverse-causality model in Figure 4b with a meta-analytic slope from Y to X (β_{XY}) calculated by inverting the model-implied correlation matrix equation so that the resulting meta-analytic correlation matrix \mathbf{P} equaled that of the omitted-variable model (see Appendix C in the online supplemental material for details);
- (b) Meta-analytic slope from X to Y , three values: $\beta_{YX} = .00, .10, .40$; and
- (c) Number of studies: $k = 10, 50$.

In addition, we considered a strong inter-IV correlation $\rho_{Z_1, Z_2} = .50$, strong IV-predictor relationships $\beta_{XZ_1} = \beta_{XZ_2} = .40$, and moderate size of heterogeneity in bivariate correlations $\tau = .10$. We fixed the variances of all variables in the data-generating models to 1 by inverting the model-implied correlation matrix equation to solve for appropriate residual variances of X and Y using the R function `nlm`. We present the algorithm used to calculate residual variances of X and Y in Appendix C.

All parameters were chosen following previous simulation studies on IV methods and on MASEM methods (Cheung, 2018; Maydeu-Olivares et al., 2019). Notably, we only considered the high-level confounding conditions (i.e., with $\beta_{XV} = \beta_{YV} = \sqrt{.20}$) according to the previous simulation study on IV methods from Maydeu-Olivares et al. (2019). This amounted to introducing a spurious correlation of .20 to the true target effect β_{YX} . Furthermore, when data were generated using the omitted variable model in Figure 4a, IVs Z_1 and Z_2 were set

Figure 4
Data-Generating Models in Simulation Studies



- (a) Model A had an omitted common cause variable V .
- (b) Model B had a reverse effect from Y to X .
- (c) Model C had an omitted common cause variable V and a direct effect of instrument Z_1 on the outcome Y .

to be uncorrelated with the confounding variable V . As a result, the exclusion assumption that IVs were not correlated with the error of the outcome and had no direct effects on the outcome were satisfied and the IVs were appropriate. A list of all chosen parameter values is available in the supplemental materials.

In sum, this simulation contained 2 (type of endogeneity) $\times 2$ (number of studies) $\times 3$ (slope from X to Y) = 12 conditions. In each condition, we simulated 1,000 datasets, which is a common sample size used in simulation studies and follows common simulation practices (Burton, Altman, Royston, & Holder, 2006; Morris, White, & Crowther, 2019).

Data-generating procedures. To simulate a meta-analytic dataset, we first computed the meta-analytic correlation matrix \mathbf{P} based on meta-analytic parameter values using the equation $\mathbf{P} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{V}_e [(\mathbf{I} - \mathbf{B})^{-1}]'$ (Appendix C explains how to construct \mathbf{P} using the prespecified parameters). For data simulated based on the model in Figure 4a, we removed the row and the column relating the confounding variable V from \mathbf{P} . We then simulated population bivariate correlations from normal distributions with means set to the corresponding values in \mathbf{P} and an SD for heterogeneity τ to construct each study-specific population correlation matrix \mathbf{P}_i . Next, we generated k sample sizes (N_i ; $i = 1, 2, \dots, k$) for individual studies by resampling with replacement the within-study sample sizes from our illustrative example. For each primary study, we simulated a sample of N_i from a multivariate normal distribution with a mean vector of zero and a covariance matrix \mathbf{P}_i . We then calculated the observed correlation matrix \mathbf{R}_i . The k sets of simulated \mathbf{R}_i and N_i constituted a meta-analytic dataset for IV-MASEM.

Performance measures. The bias of β_{YX} was used to check whether the meta-analytic estimates of β_{YX} would converge on its true value, calculated as $\bar{\hat{\beta}}_{YX} - \beta_{YX}$ where β_{YX} was the true value and $\bar{\hat{\beta}}_{YX}$ was the average estimate of β_{YX} across simulated datasets in each condition. Confidence interval (CI) coverage of β_{YX} or the proportion of simulated datasets with estimated 95% confidence intervals that contained the true β_{YX} was used to evaluate the trustworthiness of interval estimates. The ideal coverage rate should be close to the preset confidence level, that is, 95% here. Generally, a coverage lower than 90% is unacceptable when the chosen confidence level is 95%. Null hypothesis (H_0) rejection rates or the proportion of simulated datasets with significant β_{YX} estimates was used to evaluate the performance of the three studied methods in testing the target effect β_{YX} . Note that based on the definition of Type-I error rates and statistical power, rejection rates were the observed Type-I error rates when the true $\beta_{YX} = 0$ (i.e., the null hypothesis was true and it was incorrectly rejected or a significant result was observed), and the statistical power when the true $\beta_{YX} \neq 0$ (i.e., the null was not true and it was correctly rejected or a significant result was observed). Ideally, Type-I error rates should be close to the predetermined 5% significance level and power should better be as high as possible. Often, 80% power is desired.

Notably, while assessing the validity of statistical inferences, we only used datasets determined to have converged solutions during estimation as well as relevant and exogenous IVs in finite samples, in contrast to the prior study by Maydeu-Olivares et al. (2019). This was done to replicate a common practice used in real data analysis, where researchers only make conclusions after determining that the chosen IVs pass the IV diagnostics. In particular, convergence rate, the proportion of simulated replications with converged solutions, was used to exclude replications with untrustworthy estimation results.

Results

Inclusion criteria. MA techniques excluding IVs (i.e., UMA and MASEM) estimation converged normally for all simulated datasets and did not rely on IVs. Therefore, their results were based on all simulated datasets. For IV-MASEM, the convergence rates of its estimation algorithm were all above 93%. Also, less than 2% of the simulated datasets were determined to have weak IVs (i.e., β_{XZ1} and β_{XZ2} were jointly and individually not significant) and less than 8% of the datasets were determined to violate the exclusion condition across conditions (i.e., the chi-square test of model fit for the model in Figure 2 found significant results). Therefore, analysis of IV-MASEM's results excluded the above-mentioned samples with non-converged estimation solutions and IVs that were diagnosed as inappropriate.

Mean estimates of β_{YX} . As shown in Figure 5a, when IVs were relevant and exogenous, IV-MASEM provided much less biased estimates of β_{YX} than the two methods without IVs (IV-MASEM: .00 ~ .01; UMA and MASEM: .15 ~ .20). This pattern was consistent regardless of the number of studies, the effect size (i.e., β_{YX}), and the data-generating model (i.e., the omitted variable model in Figure 4a or the reverse-causality model in Figure 4b).

CI coverage of β_{YX} . As shown in Figure 5b, IV-MASEM produced confidence intervals for β_{YX} with excellent coverage rates above 90% across conditions. The coverage rates of UMA and MASEM were far below the acceptable range, that is, around zero. These near-zero coverage rates were likely due to the large estimation bias discussed above.

H0 rejection rates of β_{YX} . As shown in Figure 5c, IV-MASEM showed the best rejection rate results across the three studied methods. Specifically, IV-MASEM had good control of Type-I error rates, as indicated by the fact that rejection rates under conditions where $\beta_{YX} = 0$ were close to the nominal level, that is, .05. The other two methods without IVs showed highly inflated Type-I error rates, with rejection rates around 1 in conditions with a null effect. Again, the highly inflated Type-I error rates were largely due to the estimation bias mentioned previously. Because it is less meaningful to discuss statistical power if Type-I error rates are highly inflated, we did not compare the three methods in terms of statistical power, given that only IV-MASEM had acceptable Type-I error rates.

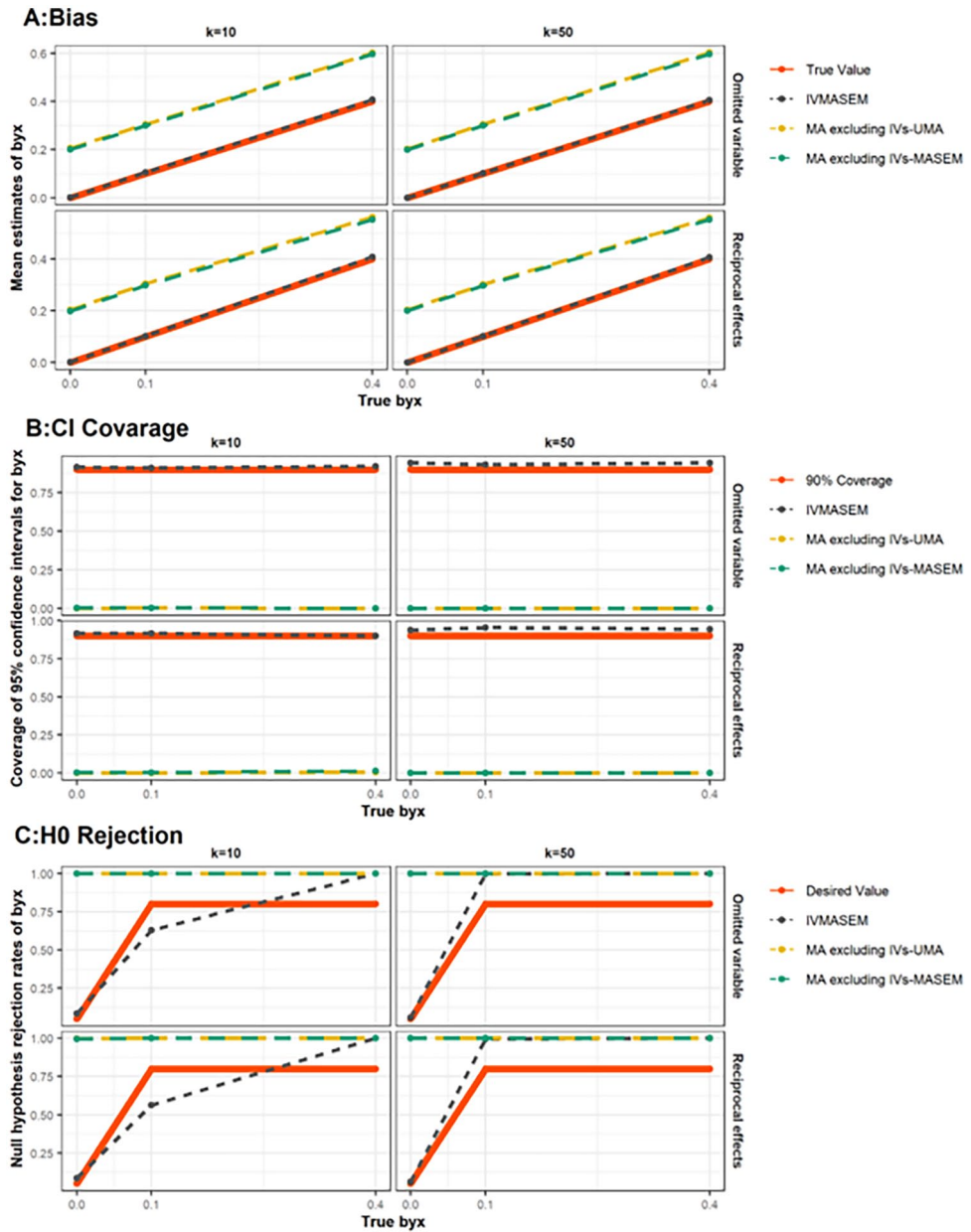
Summary. Results of this study suggest that in the favorable conditions where IVs are relevant and exogenous, IV-MASEM produces less biased estimates of β_{YX} , yields 95% interval estimates of β_{YX} with much better coverage, and provides a significance test of β_{YX} with better Type-I error rates, as compared to the two MA methods without IVs, i.e., UMA and MASEM estimates of the X-Y correlation.

Simulation 2: Comparison of IV-MASEM With MA Techniques Excluding IVs When IVs Are Inappropriate

Simulation Design

Objectives. The objectives of this simulation are twofold: (a) to compare IV-MASEM based on either weakly relevant or endogenous IVs with MA techniques excluding IVs (i.e.,

Figure 5
Simulation 1: Performance of IV-MASEM and MA Techniques Excluding IVs (UMA and MASEM) When IVs Are Appropriate in Conditions With Various Numbers of Studies (k), Sizes of Effect (byx), and Types of Endogeneity (Omitted Variable vs. Reciprocal Effects)



UMA and MASEM) in estimating and testing the target effect β_{YX} , and (b) to investigate the effectiveness of the diagnostic tests of the two IV assumptions.

Data-generating models and estimation methods. We generated data based on either the model in Figure 4c (in which the exclusion assumption was violated and IVs were endogenous) or the model in Figure 4a (in which the relevance assumption was violated and IVs were weakly relevant). To focus on the evaluation of the consequences of using weakly relevant or endogenous IVs, we only considered the omitted variable confounding mechanism in this study. The same three estimation methods as in Simulation 1 were considered: IV-MASEM, UMA, and MASEM.

Manipulated factors and parameter values. We adopted a similar but slightly different simulation design. Specifically, we had 4 (type of IV inappropriateness) \times 2 (sample size) \times 3 (slope from X to Y) = 24 conditions in this simulation. The assigned parameter values and sample size settings were as follows:

- (a) Types of IV inappropriateness: weak IVs with the following three sets of meta-analytic slopes from the IVs to the endogenous predictor: $\beta_{XZ1} = .00$ and $\beta_{XZ2} = .10$; $\beta_{XZ1} = .10$ and $\beta_{XZ2} = .10$; and $\beta_{XZ1} = .10$ and $\beta_{XZ2} = .4$. An endogenous IV, Z_1 , with a nonzero direct effect on the outcome was also included: $\beta_{YZ1} = .10$;
- (b) Meta-analytic slope from X to Y with three values: $\beta_{YX} = .00$; .10; and .40; and
- (c) Number of studies: $k = 10$ and 50.

In addition, we considered a strong inter-IV correlation $\rho_{Z1,Z2} = .50$ and a moderate size of heterogeneity in bivariate correlations $\tau = .10$. For all other parameters, we applied simulation settings identical to those in Study 1. We did not consider conditions with a strong IV-outcome residual correlation (e.g., $\beta_{YZ1} = .20$) as in a relevant previous study (Maydeu-Olivares et al., 2019). This was because the previous study showed that when the IV-outcome residual correlation was as strong as .20 and the IVs were weak ($\beta_{XZ1} = \beta_{XZ2} = .10$), the nonconvergence issue was so severe that even for the population the fitted model would not converge (Maydeu-Olivares et al., 2019). It is therefore natural to expect that the power rates of detecting this level of misspecification would be excessively high and thus results are less informative about the effectiveness of diagnostics for IV assumptions.

Performance measures. To investigate the estimation accuracy and the quality of hypothesis testing regarding the target effect made using IV-MASEM over UMA and MASEM, we again used the three performance measures in Simulation 1: bias; CI coverage; and H0 rejection rates.

In addition to the benefits of IV-MASEM in estimation and testing of the target effect, we were also interested in evaluating how effective IV-MASEM was in detecting violations of the two IV assumptions. The rate of detected weak IVs or the proportion of simulated datasets determined to have weak IVs was used to evaluate the effectiveness of the diagnostics of the relevance assumption. IVs were diagnosed as weak instruments if the meta-analytic slopes from IVs to the predictor (i.e., $\beta_{XZ1}\beta_{XZ2}$) were both individually and collectively nonsignificant. The rate of detected endogenous IVs or the proportion of simulated datasets determined to have significant direct effects of IVs on the outcome Y or nonzero IV-outcome

residual correlations was used to assess the effectiveness of the diagnostics of the exclusion assumption. IVs were considered endogenous if the chi-square test of model fit for the IV-MASEM model in Figure 2 returned a significant result.

Results

Inclusion criteria. MA techniques excluding IVs (i.e., UMA and MASEM) estimation converged normally in nearly all simulated datasets.³ For IV-MASEM, the convergence rates of the estimation algorithm reported in Table 2 were 100% when the number of studies was at least 50. Additionally, the convergence rates of IV-MASEM decreased as the strength of IVs and/or the number of studies decreased. The following analyses regarding the diagnostics of the two IV assumptions were based on datasets with converged results.

Detection of weak IVs. Results in Table 2 showed that the power rates to detect weak IVs were low. Specifically, when there were only 10 studies and the IVs were weak in the population ($\beta_{XZ1} = 0$ and $\beta_{XZ2} = .1$), 26% to 32% of simulated datasets were determined to have weak IVs in finite samples. For all the other conditions, the proportions of datasets having weak IVs was less than 6%. Despite low power rates of detecting weak IVs, however, as will be shown later when examining IV-MASEM's estimation and testing of β_{YX} in the presence of weak IVs, IV-MASEM still provided less biased estimates and better control of Type-I error rates to test β_{YX} compared to UMA and MASEM. This might suggest that instrument relevance was a small-sample problem or if the (aggregated) sample size was sufficiently large, IVs weakly correlated with the predictor in the population were still sufficiently strong in finite samples.

Detection of endogenous IVs. We mainly focused on the six conditions where the first IV had a direct effect on the outcome ($\beta_{YZ1} = .1$) and the exclusion condition was violated in the population. As shown in Table 2, when there were only 10 studies, about 20% of the simulated datasets were diagnosed as having endogenous IVs. When there were as many as 50 studies, the power to detect endogenous IVs rose to between 53% and 61%. For the remaining 18 conditions with exogenous IVs or IVs only correlated with the outcome through the endogenous predictor, the rates of falsely detecting endogenous IVs were less than 7%. Overall, the chi-square test of model fit was effective in controlling the Type-I error rates—rates of falsely detecting endogenous IVs—and showed moderate statistical power to detect endogenous IVs.

Mean estimates of β_{YX} . When inappropriate IVs were weakly correlated with the predictor and had no direct effects on the outcome (see the three plots except for the bottom right one in Figure 6a), IV-MASEM yielded unbiased estimates of β_{YX} , compared to UMA and MASEM (i.e., IV-MASEM: $-.04 \sim .02$ versus UMA and MASEM: $.20 \sim .21$). The bias of UMA and MASEM did not decrease as the number of studies increased. When inappropriate IVs were strongly correlated with the predictor but had direct effects on the outcome (see the bottom right plot in Figure 6a), β_{YX} estimates from IV-MASEM were biased (around .13 regardless of the true value of β_{YX}). On the positive side, IV-MASEM was still less biased than the two methods without IVs (which produced bias around .26).

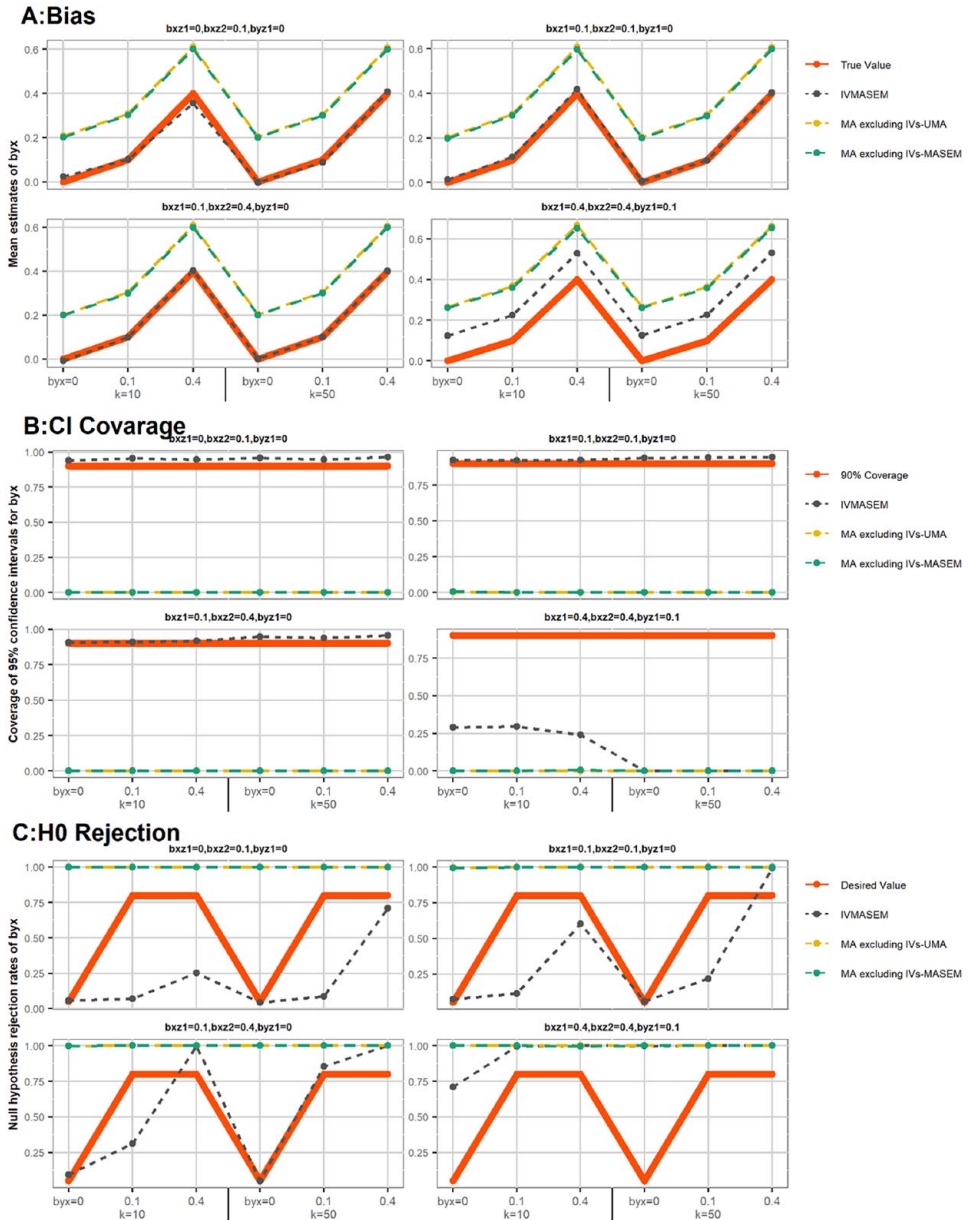
Table 2
Results of Simulation 2 Regarding the Diagnostics of the Two Instrument Assumptions When Instruments Were Inappropriate

<i>k</i>	β_{YX}	Rate of Detected Weak Instruments	Rate of Detected Endogenous Instruments	Convergence Rate of IV-MASEM
Inappropriate IVs Condition 1: $\beta_{XZ1} = 0$, $\beta_{XZ2} = 0.1$, and $\beta_{YZ1} = 0$				
10	.00	.32	.05	.80
10	.10	.32	.03	.85
10	.40	.26	.04	.91
50	.00	.00	.05	1.00
50	.10	.00	.06	1.00
50	.40	.00	.04	1.00
Inappropriate IVs Condition 2: $\beta_{XZ1} = 0.1$, $\beta_{XZ2} = 0.1$, and $\beta_{YZ1} = 0$				
10	.00	.06	.05	.86
10	.10	.03	.07	.92
10	.40	.04	.07	.90
50	.00	.00	.05	1.00
50	.10	.00	.05	1.00
50	.40	.00	.06	1.00
Inappropriate IVs Condition 3: $\beta_{XZ1} = 0.1$, $\beta_{XZ2} = 0.4$, and $\beta_{YZ1} = 0$				
10	.00	.00	.06	.97
10	.10	.02	.07	.93
10	.40	.01	.07	.95
50	.00	.00	.05	1.00
50	.10	.00	.06	1.00
50	.40	.00	.05	1.00
Inappropriate IVs Condition 4: $\beta_{XZ1} = 0.4$, $\beta_{XZ2} = 0.4$, and $\beta_{YZ1} = 0.1$				
10	.00	.01	.19	.95
10	.10	.02	.21	.94
10	.40	.02	.18	.95
50	.00	.00	.59	1.00
50	.10	.00	.61	1.00
50	.40	.00	.53	1.00

CI coverage of β_{YX} . As shown in the three plots, except for the one on the bottom right in Figure 6b, UMA and MASEM showed coverage that was almost zero due to the large bias previously discussed. IV-MASEM produced interval estimates of β_{YX} with excellent coverage > 91% when inappropriate IVs were weakly correlated with the predictor and had no direct effects on the outcome. However, when inappropriate IVs had direct effects on the outcome (see the bottom right plot in Figure 6b), IV-MASEM showed unacceptably low coverage < 30%. The estimation bias under such circumstances, noted previously, was a substantial factor here.

H0 rejection rates of β_{YX} . When the inappropriate IVs were weakly correlated with the predictor and had no direct effects on the outcome (see the three plots except for the one on the bottom right in Figure 6c), IV-MASEM was the only method that demonstrated proper control over Type-I error rates at < .09. UMA and MASEM showed Type-I error rates close

Figure 6
Simulation 2: Performance of IV-MASEM and MA Techniques Excluding IVs (UMA and MASEM) When IVs Are Not Appropriate in Conditions With Various Numbers of Studies (k), Sizes of Effect (byx), Instrumental Relevance (bxz1 and bxz2), and Violations of Exclusion Assumption (byz1)



to one and estimation bias was likely a major factor. We only discuss the statistical power of IV-MASEM because discussing the statistical power of the two approaches without IVs would be meaningless due to the significantly exaggerated Type-I error rates. The power rates of IV-MASEM to detect significant β_{YX} increased noticeably as the slopes from IVs to the predictor (β_{XZ1} and β_{XZ2}) and/or the number of studies increased. Particularly, the slopes from IVs to the predictor had a considerable impact. When the slopes from the IVs to the predictor were strong and the true value of β_{YX} and the number of studies were held constant, the power rates of the significance test of β_{YX} could be multiple times higher than in cases where the slopes from the IVs to the predictor were weak. For example, with weak IVs ($\beta_{XZ1} = .00$ and $\beta_{XZ2} = .10$), the power rate of IV-MASEM was 25% in the condition with $\beta_{YX} = .40$ and the number of studies was 10. However, this increased to 100% with strong IVs ($\beta_{XZ1} = .10$ and $\beta_{XZ2} = .40$). We observed a different pattern of results when IVs had direct effects on the outcome (see the bottom right plot in Figure 6c). IV-MASEM produced inflated Type-I error rates $> .71$ for the test of β_{YX} , suggesting untrustworthy hypothesis testing results from IV-MASEM in such conditions.

Summary

Results of Simulation 2 showed that the power rates to detect weak IVs were low. Fortunately, weak IVs appeared to be a small-sample size issue. As long as the (aggregated) sample size was sufficiently large, weak IVs had a small impact on IV-MASEM point estimates, interval estimates, and Type-I error control of β_{YX} , despite showing a negative effect on the power to detect β_{YX} . Simulation results also showed that the power rates to detect endogenous IVs were relatively high compared to those of detecting weak IVs. However, the power rates still fell short of the desired 80% level. In addition, applying IV-MASEM to datasets that met the diagnostic criteria for the exclusion assumption still led to biased estimates, interval estimates with unacceptably low coverage rates, and liberal hypothesis testing results. The encouraging finding was that even in unfavorable conditions with endogenous IVs, IV-MASEM still outperformed MA techniques excluding IVs (i.e., UMA and MASEM) in terms of bias and coverage rates. Therefore, although MASEM might not be perfect, it still outperformed traditional meta-analytic methods.

Simulation 3: Comparison of IV-MASEM With MA Techniques Excluding IVs When IVs Are Available for Only a Subsample of the Primary Studies

Simulation Design

Objectives. In the context of meta-analysis, it is not uncommon to find that IVs are only available in a subset of primary studies or, equivalently, bivariate correlations involving IVs are missing in a subsample of primary studies. This simulation aims to evaluate the impact of missing IVs on the performance of IV-MASEM and MA techniques excluding IVs (i.e., UMA and MASEM).

Data-generating models and estimation methods. We generated data based on the model in Figure 4a and omitted IV-related correlations randomly with two missing rates. The fitting

model was again the one in Figure 2 and we only considered the omitted variable confounding mechanism in this study. The same three estimation methods as in Simulation 1 were considered: IV-MASEM; UMA; and MASEM.

Manipulated factors and parameter values. We adopted a similar but somewhat modified simulation design. Specifically, our conditions were 2 (missing rate) \times 2 (sample size) \times 3 (slope from X to Y) = 12. The specific assigned parameter values and sample size settings were listed below:

- (a) Missing rate for $X - Z_1$, $X - Z_2$, $Y - Z_1$, and $Y - Z_2$ correlations of two levels $\pi = 60\%$ and 80% ;
- (b) Meta-analytic slope from X to Y of three values $\beta_{YX} = .00$, $.10$, and $.40$; and
- (c) Number of studies $k = 10$ and 50 .

In addition, we considered a high missing rate for the $Z_1 - Z_2$ correlation (90%), a strong inter-IV correlation ($\rho_{Z_1, Z_2} = .50$), a strong IV-predictor correlation ($\beta_{XZ_1} = \beta_{XZ_2} = .40$), and a moderate size of heterogeneity in bivariate correlations ($\tau = .10$). Apart from this, we applied simulation settings identical to those in Simulations 1 and 2. The missing rates were chosen to mimic the high missingness found in some MASEM studies.

Missing data generation. Based on the data-generating procedure described in Simulation 1, we generated missing bivariate correlations in the following way. We first specified a missing rate π for $X - Z_1$, $X - Z_2$, $Y - Z_1$, and $Y - Z_2$ bivariate correlations. We then randomly selected: the first IV Z_1 to be missing at a rate = $.90 - \pi$; the second IV Z_2 to be missing at a rate = $.90 - \pi$; and both IVs Z_1 and Z_2 to be missing at a rate = $2\pi - .90$. If a variable was missing in a primary study, then all bivariate correlations involving this variable were missing. For example, if Z_1 was absent from the first primary study, the $X - Z_1$, $Y - Z_1$, and $Z_1 - Z_2$ correlations were all missing. According to this rule, the missing rates for the $X - Z_1$, $X - Z_2$, $Y - Z_1$, and $Y - Z_2$ correlations were $.90 - \pi + 2\pi - .90 = \pi$, as noted previously. For the $Z_1 - Z_2$ correlation, its missing rate was $.90 - \pi + .90 - \pi + 2\pi - .90 = .90$. In conditions with 10 primary studies, by chance alone the IV-related correlations were absent from nearly all primary studies and MASEM was not estimable. In these cases we “refilled” or eliminated the missingness of the IV-related correlations in three randomly selected primary studies to satisfy the minimum number for meta-analysis (Chambless & Hollon, 1998; Seibert, Wang, & Courtright, 2011). This was done when the IV-related correlations were observed in fewer than three studies in a meta-analysis dataset.

Performance measures. As in Simulation 2, we used the power rates to detect weak and endogenous IVs to evaluate the impact of missing IVs on the effectiveness of the IV assumption diagnostics. Also, we used the following three performance measures to evaluate the effect of missing IVs on IV-MASEM’s estimation and testing of β_{YX} : bias; CI coverage; and H0 rejection rates. Results regarding IV-MASEM in estimation and testing were calculated based on samples with converged estimation and IVs diagnosed as relevant and exogenous.

Table 3
Results of Simulation 3 Regarding the Diagnostics of the Two Instrument Assumptions in the Presence of Missing Instruments When Instruments Were Strong and Exogenous in the Population

k	β_{YX}	Rate of Detected Weak Instruments	Rate of Detected Endogenous Instruments	Convergence Rate of IV-MASEM
Missing Rate Condition 1: 60%				
10	.00	.33	.09	.48
10	.10	.26	.09	.55
10	.40	.39	.08	.46
50	.00	.03	.05	.92
50	.10	.02	.07	.94
50	.40	.02	.06	.95
Missing Rate Condition 2: 80%				
10	.00	.23	.09	.60
10	.10	.34	.09	.47
10	.40	.32	.10	.52
50	.00	.08	.08	.85
50	.10	.06	.07	.87
50	.40	.06	.07	.86

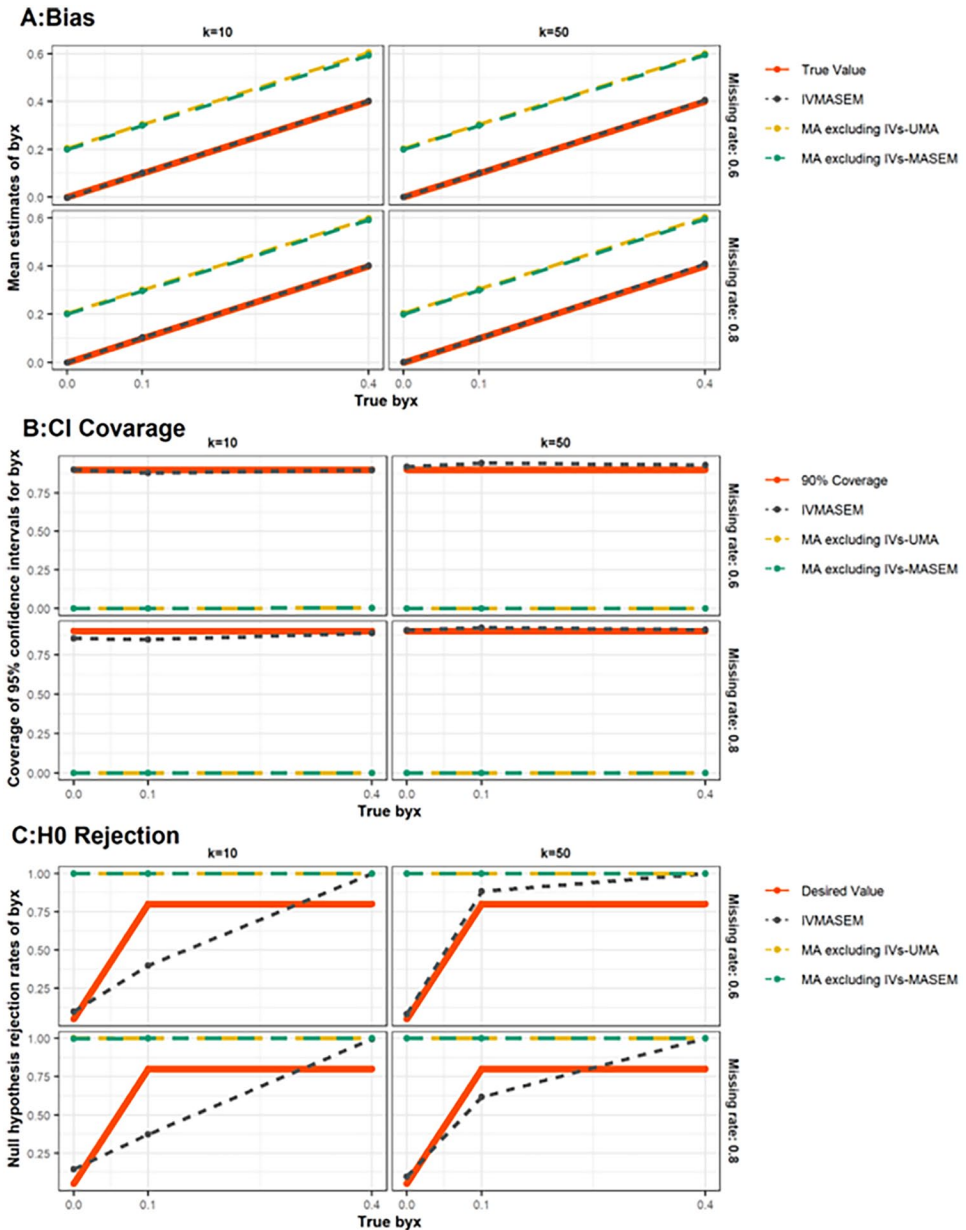
Results

Inclusion criteria. The analyses on the detection of weak and endogenous IVs were based on samples with converged estimation. Specifically, as shown in Table 3, UMA and MASEM on the X-Y relationship converged normally in all simulated datasets, partly because there were no missing X-Y correlations, and thus their results were based on all simulated datasets. For IV-MASEM, the convergence rates were above 85% when there were 50 studies, regardless of the missing rate π . When there were 10 studies, the convergence rate could be unacceptably low: 46% ~ 60%. Therefore, IV-MASEM's results were based on subsets of simulated datasets with converged estimation.

Detection of weak IVs. When IVs were available only in a subset of studies due to missingness, the probability of falsely identifying weak IVs was nonnegligible despite IVs being strong in the population ($\beta_{XZ1} = \beta_{XZ2} = .4$). Specifically, 23% to 39% of the simulated datasets were determined to have weak IVs when the number of studies was 10. For all other conditions, the proportion of falsely detecting weak IVs was less than 8%.

Detection of endogenous IVs. Regardless of the missing rate for IV-related correlations and the number of studies, about 10% of the simulated datasets were falsely determined to have endogenous IVs. Given that IVs were actually generated to be exogenous in the population, these detection rates were Type-I error rates to detect endogenous IVs. They were slightly inflated but remained in an acceptable range in the presence of missing IV-related correlations.

Figure 7
Simulation 3: Performance of IV-MASEM and MA Excluding IVs (UMA and MASEM) When IVs Are Available in a Subset of Primary Studies in Conditions With Various Numbers of Studies (k), Sizes of Effect (byx), and Missing Rates



Mean estimates and CI coverage of β_{YX} In the presence of missing IVs, IV-MASEM still produced less biased estimates of β_{YX} compared to UMA and MASEM, as shown in Figure 7a. With respect to interval coverage, IV-MASEM was the only method that provided 95% interval estimates with acceptable coverage rates (i.e., > 85%), as shown in Figure 7b. Interval estimates obtained using UMA and MASEM had coverage rates close to zero, possibly because of the large upward bias in the estimation of β_{YX} , as shown in Figure 7b. Notably, the coverage rates of IV-MASEM intervals were lower when the number of studies was small (i.e., 10 rather than 50). The decrease in coverage rates was likely due to the downward bias in SE with only 10 studies.

H0 rejection rates of β_{YX} Rejection rate results in Figure 7c showed patterns similar to those in Simulation 1. IV-MASEM showed better control of Type-I error rates to test β_{YX} in the presence of missing IVs, compared to UMA and MASEM. However, its Type-I error rates (.08~.15) were higher than .05, and higher than rates when IVs were available in all studies as in Simulation 1.

Summary

Results of this study suggest a nontrivial impact of missing IVs in a subset of studies on the detection of weak IVs, coverage rates, and Type-I error rates for the significance test of β_{YX} . Although IV-MASEM still outperformed MA techniques excluding IVs (i.e., UMA and MASEM), it yielded inflated Type-I error rates to falsely detect weak IVs, produced 95% interval estimates for β_{YX} with lower coverage rates, and showed inflated Type-I error rates to test β_{YX} , when IVs were available in only a subset of studies.

Illustrative Study for Using IV-MASEM

To illustrate IV-MASEM and compare it with MA techniques excluding IVs (i.e., UMA and MASEM), we used IV-MASEM methods, UMA, and MASEM excluding IVs to explore the effect of abusive supervision (X) on supervisor-directed deviance (Y). For IV-MASEM, we used two personality traits as exogenous IVs, according to the suggestion of Antonakis et al. (2010), including social desirability (Z_1) and negative affectivity (Z_2). The relationship between abusive supervision and supervisor-directed deviance has been an important topic in management research, and it has been synthesized in previous correlational meta-analyses (Mackey, Frieder, Brees, & Martinko, 2015; Zhang & Liao, 2015). Abusive supervision is defined as the subordinate's perception of the supervisor's persistently displaying hostile verbal or nonverbal behavior excluding physical contact (Mawritz, Greenbaum, Butts, & Graham, 2017; Tepper, Duffy, & Shaw, 2001). For example, supervisors yell at, ridicule, intimidate, and publicly humiliate their employees (Kluemper, Mossholder, Ispas, Bing, Iliescu, & Ilie, 2019; Tepper, 2000, 2007).

When repeatedly exposed to abusive behaviors by supervisors, employees experience anger, frustration, helplessness, and anxiety (Aryee, Sun, Chen, & Debrah, 2008; Harvey, Stoner, Hochwarter, & Kacmar, 2007; Restubog, Scott, & Zagenczyk, 2011; Tepper, Moss, Lockhart, & Carr, 2007). According to negative social exchange theory, subordinates who experienced abusive supervision often took some actions in the form of supervisor-directed deviance that were harmful to the supervisor as a response to the abusive supervision (Burton

& Barber, 2019; Mitchell & Ambrose, 2007, 2012). Although research on abusive supervision and supervisor-directed deviance has made great progress, few previous studies have examined the causal relationship between abusive supervision and supervisor-directed deviance. In particular, previous research suggests the relationship between subordinate misbehavior and abusive supervision could be reciprocal (Lian, Ferris, Morrison, & Brown, 2014), implying bi-directionality or “reverse” causality. Additionally, abusive supervision and supervisor-directed deviance are often correlated with the same third variables, such as job satisfaction (Tepper, Carr, Breaux, Geider, Hu, & Hua, 2009), negative reciprocity (Mitchell & Ambrose, 2007), leader–member exchange (Chen & Liu, 2019), and work-related negative affect (Michel, Newness, & Duniewicz, 2016). To gain a more comprehensive and accurate understanding of the relationship between abusive supervision and supervisor-directed deviance in the meta-analysis, we explored this potential causal effect using the three IV-MASEM methods mentioned in the manuscript. As Antonakis et al. (2010) suggested, personality traits could be IVs in management research because they are stable and exogenous. Accordingly, we use two personality traits—social desirability and negative affectivity—as IVs, which are correlated with abusive supervision while not directly correlating with employee-based supervisor-directed deviance (Lian, Lance Ferris, & Brown, 2012; Mitchell & Ambrose, 2012).

Sample and Coding

In order to expand the meta-analysis database of primary studies as much as possible, based on the four variables involved in the theoretical model in this study, we first searched all related empirical studies in the following databases: Web of Science (SSCI); Scopus; PsycINFO; EBSCO; ERIC; ABI /INFORM; Google Scholar; and CNKI. Second, we further enlarged our meta-analysis database with highly cited qualitative and quantitative reviews (e.g., Tepper, 2007). Additionally, we searched Google Scholar, Web of Science (SSCI), PsycINFO/Dissertation, ProQuest Digital Dissertations, SCOPUS, and PsycINFO for unpublished studies, including conference papers, working papers, dissertations, book chapters, and reports. To identify valid samples, we developed the following inclusion criteria: (1) the sample articles contain at least one bivariate relationship in our theoretical model; (2) the sample articles must be empirical; (3) the sample articles must contain correlations. Based on these three criteria, a total of 85 published studies were identified. We developed a coding scheme based on Krippendorff (2018). Specifically, the articles were first independently coded to avoid data inaccuracies due to coding errors. Second, three main aspects of information were extracted from each article, including (a) the sample size, (b) the reliability of each variable, and (c) the correlations. During the coding process, some subtle differences were also resolved through discussion. All effect sizes we coded and references included in our meta-analysis are presented in online supplemental material in the OSF.

Analysis

First, we calculated the number of independent studies (k), sample size (N), the weighted mean correlation (r) of each bivariate effect with its standard deviation (SDr), the mean true-score correlation (ρ) of each bivariate effect with its standard deviation ($SD\rho$), the 95%

confidence interval for the mean effect, and the 80% credibility interval according to Hunter and Schmidt (2004).

Second, we calculated the causal effect of abusive supervision on supervisor-directed deviance by using UMA, MASEM, and three new IV-MASEM methods: IV-U-MASEM; IV-FI-MASEM; and IV-OS-MASEM. Consistent with the analytical procedure for IV-MASEM we proposed, we specified a structural model. Following the suggestion of Antonakis et al. (2010), we used two personality traits as exogenous IVs, including social desirability (Z_1) and negative affectivity (Z_2), to predict abusive supervision (i.e., an endogenous predictor X), which affects supervisor-directed deviance Y (see Figure 1 for the model diagram). Crucially, we specified correlated errors among X and Y to allow for endogeneity in X , and the two IVs Z_1 and Z_2 were specified as having no direct relationship with Y . R code and outputs (see supplemental material in the OSF) and functions used in the analysis (see supplemental material in the OSF) have been uploaded to OSF to help interested readers replicate our analyses. Specifically, in the interest of article length, we present the results of IV-U-MASEM, which is the most commonly used MASEM method in management research. A full report of results based on all three IV-MASEM methods (i.e., IV-U-MASEM, IV-FI-MASEM, and IV-OS-MASEM) is available in our online supplemental materials, where interested readers can access detailed results and compare the analytical processes of these three methods. For our analysis, a total of 85 studies were included. All the effect sizes we coded and references included in our meta-analysis are presented in the online supplemental materials.

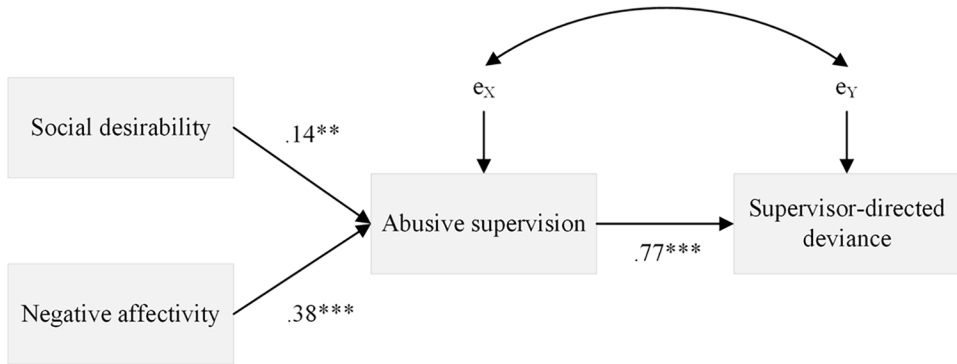
Results of MA Techniques Excluding IVs

We considered two MA techniques excluding IVs: UMA and MASEM. For UMA, we computed the sampling variance for the X-Y correlation from each primary study using Equation 3.4 in Schmidt and Hunter (2015: 99). Results based on UMA showed that the meta-analytic correlation between abusive supervision and the supervisor-directed deviance was moderate: $\hat{\beta}_{YX} = \hat{\rho}_{YX} = 0.43$, $SE = 0.03$, $p < .001$. For MASEM, we reformatted the meta-analysis dataset from a vector of X-Y correlations to a list of 2-by-2 correlation matrices with X-Y correlations set to the off-diagonal elements as data input for MASEM analysis. Results of MASEM showed that the relation between abusive supervision and the supervisor-directed deviance was moderate: $\hat{\beta}_{YX} = \hat{\rho}_{YX} = 0.43$; $SE = 0.03$; $p < .001$. These results indicate there is a remarkable alignment between the findings of our illustrative and simulation studies. Specifically, the results of UMA and MASEM are essentially equivalent in the analysis of univariate relationships.

Results of IV-MASEM

To proceed with our analysis, we checked the IV assumptions by first assessing the fit of the model. This provides a global assessment of whether the two IVs are uncorrelated with the error of Y . Model fit assessed using the χ^2 showed adequate fit ($\chi^2 = 3.26$, $df = 1$, $p = .07$), suggesting both IVs satisfied the condition of being unrelated to Y 's error term. Second, we checked the other IV condition that the two IVs were strongly correlated with X . A Wald test of the two IV's coefficients predicting X was highly significant ($\chi = 58.97$, $df = 1$, $p < .001$).

Figure 8
The Model Results of the Illustration Study



Note. Asterisks denote levels of statistical significance: * indicates $p < .05$, ** indicates $p < .01$, and *** indicates $p < .001$.

This indicates that at least one of the IVs was correlated with X, but, given the large sample size and omnibus nature of this test, we evaluated each instrument separately based on its coefficients. Social desirability was statistically and practically related to abusive supervision ($b_{X-IV1} = .14, SE = .05, p = .004$), and thus social desirability did appear to be an IV for the endogenous variable abusive supervision. Negative affectivity was also statistically and practically related to abusive supervision ($b_{X-IV2} = .38, SE = .05, p < .001$), indicating negative affectivity appears to be a good IV for abusive supervision. As indicated earlier, we also recommend using R^2 to assess the overall strength of these associations, which in this case for the two IVs was .14, bigger than Cohen's criterion of medium R^2 effect sizes (.13). This indicated no weak instruments. In the third step, we proceeded to explore causality. As shown in Figure 8, the meta-analytic path coefficient between abusive supervision and supervisor-directed deviance was $b_{YX} = .77, SE = .08, p < .001$. This result showed that abusive supervision was indeed an important predictor of supervisor-directed deviance, and the effect between the two variables was supported by the IV-U-MASEM estimated effect.

Summary

In conclusion, according to the results from UMA, MASEM, and IV-U-MASEM methods, we can see that by incorporating IVs into MASEM, researchers may obtain very different meta-analytic effect size estimates, as opposed to the traditional UMA and MASEM methods without IVs. Specifically, UMA and MASEM showed that the effect of abusive supervision on supervisor-directed deviance was moderate, whereas IV-MASEM showed that this effect was large. Given that these results followed a pattern that was consistent with the findings of the simulation studies, and as these results might complement other findings in the literature showing similar effects, we suggest that IV-MASEM can be a useful complement to existing meta-analytic testing.

Discussion

In management research, meta-analyses using UMA and MASEM excluding IVs are increasingly common, and can be potent methods for synthesizing findings from multiple studies and obtaining robust estimates. UMA and MASEM excluding IVs demonstrate a robust capacity for integrating effect sizes from both experimental and non-experimental data, allowing for efficient amalgamation of diverse data types. However, issues of endogeneity in the aggregation process are often overlooked by researchers. While endogeneity is not a concern with experimental effect sizes for UMA and MASEM excluding IVs, it poses challenges with non-experimental effect sizes. UMA cannot address these issues, and although MASEM has the capability by integrating IV, it is rarely explicitly addressed by scholars utilizing this approach. In response, we propose IV-MASEM as a potential solution. The methodological innovation of IV-MASEM represents a significant advancement by applying the IV-SEM framework to meta-analysis, thereby improving the accuracy and validity of synthesized findings. Our paper offers three comparative simulations that explore the efficacy of IV-MASEM, UMA, and MASEM excluding IVs under various conditions, shedding light on their advantages and limitations. The results show that the introduction of IVs in MASEM, even in the face of challenges caused by weak IVs, endogenous IVs, or missing data, can significantly enhance the MASEM model's ability to manage endogeneity. Specifically, Simulation 1 compared IV-MASEM with UMA and MASEM excluding IVs when IVs were strong and exogenous. We found that IV-MASEM gave much more accurate estimates of the target effect than UMA and MASEM excluding IVs, showing better control of Type-I error rates to test the target effect. In Simulation 2, we compared the performance of these three methods when the IVs were invalid, that is, weakly relevant or endogenous. According to the results of the Simulation 2, under the condition of large sample sizes, weak IVs lead to low statistical power and endogenous IVs bias point estimates, interval estimates, and Type-I error control of β_{YX} within IV-MASEM. However, compared to UMA and MASEM excluding IVs, the biases produced by IV-MASEM are smaller. This illustrates the comparative robustness of IV-MASEM against potential distortions in parameter estimation when faced with weak or endogenous IVs. Finally, Simulation 3 showed that IV-MASEM recovers the effect with negligible bias even when the proportion of missing IVs is high, even though the accuracy of interval estimation, and significance tests of β_{YX} decreased. This partly addresses the issue of missing IVs, which is relevant because primary studies frequently measure different variables—as we elaborate on below. In sum, our results show potential advantages of IV-MASEM over UMA and MASEM excluding IVs. We now explore the implications of our findings and offer suggestions for future research.

Implications

Two overall objectives of management research and scientific research more generally are to find unbiased effects and systematically summarize the empirical research on a given topic (e.g., Bilgili, Calderon, Allen, & Kedia, 2017; Zhang et al., 2019). When effect sizes from randomized experiments are not available, these two objectives often cannot be realized simultaneously, because unbiased inferences are hard to make based on effect sizes from observational primary studies. The major challenge is that the effect of an endogenous predictor X on an outcome Y cannot be estimated consistently because of endogeneity. The IV-MASEM overcomes this by including IVs in MASEM, allowing an X - Y error correlation

to be estimated (Antonakis et al., 2010; Bollen, 2012; Maydeu-Olivares et al., 2020). The net result is that the IV-MASEM can help researchers achieve two key objectives simultaneously: enhancing the accuracy of inference and synthesizing the findings of primary studies meta-analytically, even when primary studies are non-experimental and some may not contain all required variables and correlations.

From the perspective of method development, our paper advances previous work on IV methods, which were limited to single empirical studies. By incorporating the logic of IV-SEM, IV models can be estimated using three types of MASEM: U-MASEM; FI-MASEM; and OS-MASEM. All of these can accommodate an IV-SEM structure, with different advantages with each. An advantage of all these methods is that even if each primary study does not investigate all relevant variables, and thus does not allow IV modeling within a given study, the IV methods can still be performed at the aggregate level between studies. Specifically, any single study might not measure enough of these variables, making IV methods impossible. In contrast, by synthesizing information from multiple studies, MASEM allows management researchers to apply IV methods at the aggregate level even in the presence of missing data on an IV—as we demonstrate in Simulation 3. In such cases, the primary correlational study can still improve estimates of meta-analytically derived correlations, which can be incorporated with information about IVs from other primary studies that have measured them.

Although IV-MASEM is an application of IV-SEM for analyzing effect size data within meta-analysis, significant distinctions exist between IV-SEM and IV-MASEM across three aspects. First, IV-SEM generally fits a single covariance matrix, whereas IV-MASEM (e.g., IV-OS-MASEM) fits multiple matrices to account for between-study heterogeneity, which is a unique strength. Second, the methods of collecting IV data differ, with IV-SEM using primary data and IV-MASEM typically gathering effect size data from published studies that extends the scope for identifying IVs. Lastly, the nature of result reporting differs between the two: IV-SEM reports on the fit of a single model, whereas IV-MASEM not only calculates relationships but also considers between-study heterogeneity through effect size distribution estimation.

From the perspective of method application, we have detailed and illustrated analytical procedures of IV-MASEM. Specifically, before the application of IV-MASEM, researchers should examine the two conditions for applying IV methods in the context of meta-analysis, including research design and modeling conditions. Regarding research design, UMA can be applied when endogenous predictors in most primary studies are manipulated through randomized experiments, whereas IV-MASEM should be applied when most primary studies are observational. Regarding modeling, both a relevance assumption (i.e., IVs are correlated with the endogenous predictors) and an exclusion assumption (i.e., IVs are independent of endogenous outcomes after accounting for endogenous predictors) are important for IV-MASEM. Overall, IV-MASEM will offer more accurate estimates than traditional meta-analysis when these two conditions are satisfied: (1) if the researcher finds that in primary studies there is insufficient effect size information from randomized experiments (i.e., a research design condition); and (2) the researcher finds IVs that satisfy the relevance and exclusion assumptions (i.e., IV modeling condition). In these cases, we recommend IV-MASEM as a preferred method that researchers should consider.

In the process of applying IV-MASEM, the selection of a specific MASEM method and IVs are critical. It is necessary to take the type of correlational matrix as data input into account when selecting between three IV-MASEM methods, as per our fourth step for IV-MASEM. Both IV-U-MASEM and IV-FI-MASEM can analyze meta-analytic correlation matrices, whereas IV-OS-MASEM relies on primary-study correlations. For using meta-analytic correlation matrices, Landis (2013: 256) notes “meta-analytic correlations reported elsewhere should be the most viable values” to build matrices as data inputs for MASEM when the average correlation matrix is incomplete. This can be an effective strategy as many published papers show (Bergh et al., 2016; Friend, Jaramillo, & Johnson, 2020), but it is based on aggregated correlation effect sizes and thus requires U-MASEM or FI-MASEM. For IV-MASEMs with primary study data, IV-OS-MASEM can be adopted. Accordingly, if researchers plan to use IV-MASEM based on meta-analytical effect sizes from previously meta-analysis, then IV-U-MASEM and IV-FI-MASEM are more feasible. When choosing between IV-U-MASEM and IV-FI-MASEM, it is important to consider whether the model path coefficients are heterogeneous. If researchers are interested in modeling the heterogeneity of path coefficients, IV-FI-MASEM should be applied.

Regarding the selection of IVs, each primary study may use a different IV. Thus, more primary studies may increase the number of IVs. Modeling many IVs presents challenges, so, rather than address the complexity of numerous IVs during the modeling phase, we suggest a strategic approach during the inclusion/exclusion phase. This involves predetermining an effective number of IVs suitable for IV methods. Subsequently, the most common IVs across the primary studies are selected. For instance, if six studies focus on variables X and Y, necessitating two IVs, researchers would identify the most common IVs, such as Z_1 and Z_2 , and incorporate them into the IV-MASEM. This will reduce missing data rates compared to including all IVs, and the issues caused by missing data shown in Simulation 3. It is important to note that choosing to include only a subset of IVs does not indicate that primary studies utilizing other IVs will not be included in a meta-analysis. If those studies report the X-Y correlation, then they should also be included. What can be excluded are the correlations involving alternative IVs from primary studies. This strategic selection not only addresses the challenge of managing numerous IVs, but it can also enhance the usefulness of IV-MASEM by contributing to accurate meta-analytic inferences.

Directions for Further Research

Given the potential for IV-MASEM to estimate unbiased effects meta-analytically and to reanalyze previously published meta-analytic data, several issues are worth future research. First, the IV method can only be used to strengthen the accuracy of inferences with meta-analytic correlation matrices. When between-study heterogeneity exists, especially with heterogeneous effects, the average effect across studies will lose some of its utility for generalization. Accordingly, including moderators (i.e., between-study covariates such as sample or study characteristics) to explain the heterogeneity may be helpful. However, there is an issue here that future research can address: the presence of between-study heterogeneity implies that the required IV-MASEM assumption checking should be done for each primary study. Chi-square tests, z-tests, and Wald tests of paths from IVs to the predictor, and assessing R^2 values, allow checking the exclusion and relevance assumptions at the aggregate level,

as we have done. However, it is unclear how to check these assumptions in primary studies for meta-analyses. To evaluate this, future research can focus on new methods for detecting individual study violations for IV-MASEM.

Next, it is worth noting that different primary studies may measure the same variable in different ways. For example, Cano, Carrillat, and Jaramillo (2004) quantitatively synthesized the correlations between market orientation and business performance. Regarding the measurement of firm market share, there are three different scales in the primary studies: objective, subjective, and a mixture of these. However, these measures may still have other troubling properties such as range restriction, and well-known corrections for such artifacts are common for univariate meta-analysis (e.g., Hunter & Schmidt, 2004; Røysamb, Nes, Czajkowski, & Vassend, 2018). Future research should explore the impact of measure diversity and artifact corrections in applications of IV-MASEM.


It is also worth noting that meta-analytic datasets can be viewed as having a multi-group structure (i.e., every sample is a group). IV-MASEM is based on a multilevel rather than a multi-group framework, but multi-group IV-MASEM could in theory also be developed. Future research can explore how multi-group SEM can analyze meta-analytic datasets with IV methods. However, these applications may be most useful in specific contexts, such as when there is no missing data, the effects are heterogeneous, and categorical moderation tests involve a small number of primary studies. Specifically, when the number of studies is large, the issue of testing (approximate) invariance across groups will result in difficulty in testing the invariance of any IV-MASEM parameters (e.g., the path from Z to X or from X to Y). In meta-analytic datasets, the number of studies is often large (> 30 or even > 100; e.g., Colquitt et al., 2013; Rupp, Shao, Jones, & Liao, 2014), and thus the IV-MASEM framework we present may be more effective than other approaches when a dataset contains a large number of groups (Muthén & Asparouhov, 2018). Future research could cover this topic in more detail.

Finally, although IV-MASEM has advantages for enhancing the accuracy of inference, it has some limitations of which researchers should be aware, and which could be addressed in future research. First, it is often challenging to find perfect IVs. For example, in research on organizational behavior, personal traits may be appropriate as IVs, but it may be uneasy to identify other types of variables to use as IVs (Antonakis et al., 2010). In particular, the IV-MASEM is developed under the IV-SEM framework. It requires the number of instrumental variables (IVs) to be at least one more than the number of endogenous predictor variables in the theoretical model. Otherwise, the IV-MASEM method will not be able to test whether the exogeneity assumption of the IVs is met based on model fit. However, in empirical research, it is challenging sometimes to identify just one high-quality IV. In this case, it may be challenging to identify two or more suitable IVs. Second, the IV method typically assumes that the relationships between variables are linear and the estimated residuals are homoscedastic, which may be violated (Bascle, 2008; Maydeu-Olivares et al., 2020). Third, our simulations show IV-MASEM can address the endogeneity caused by omitted variables and reverse causality, but other forms of endogeneity are hard to address statistically using IV-MASEM (Hill et al., 2021). For example, endogeneity due to treatment effects needs to be addressed at the stage of research design. To reduce endogeneity caused by design issues, researchers need to assess the quality of primary studies. When screening primary studies for inclusion, researchers could exclude studies that suffer from issues such as treatment effects.

Conclusion

In sum, IV methods are a valuable tool for strengthening the accuracy of inference in non-experimental studies, particularly given the large proportion of such data in management research. By leveraging the benefits of IV methods and MASEM, we have provided researchers with powerful new tools for research synthesis in addressing endogeneity. Our simulations show that modeling effect sizes drawn from non-experimental primary studies with IVs can lead to valid estimates across various conditions. We hope that our paper will stimulate additional work to improve the accuracy of inference in meta-analytic research and beyond.

ORCID iD

Yucheng Zhang  <https://orcid.org/0000-0001-9435-6734>

Notes

1. We focus on OS-MASEM as it is an advance over a related two-stage MASEM developed by Cheung (2015). Both of these two approaches model the effect size data from primary studies hierarchically, using a within-study and between-study framework. The within-study model assumes a vectorized model-implied correlation matrix, true SEM parameters, residual correlations, and sampling error deviations. The between-study model incorporates covariates to explain heterogeneity in SEM parameters.

2. We thank an anonymous reviewer for suggesting a well-organized structure of this simulation study.

3. MASEM estimation did not converge in one simulated dataset in the condition where the number of studies was 10, $\beta_{XZ1} = \beta_{XZ2} = 0.4$, and $\beta_{YZ1} = 0.1$.

References

- Aldrich, J. 1995. Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10: 364-376.
- Angrist, J. D., & Krueger, A. B. 2001. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15: 69-85.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. 2010. On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21: 1086-1120.
- Aryee, S., Sun, L.-Y., Chen, Z. X. G., & Debrah, Y. A. 2008. Abusive supervision and contextual performance: The mediating role of emotional exhaustion and the moderating role of work unit structure. *Management and Organization Review*, 4: 393-411.
- Bascle, G. 2008. Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, 6: 285-327.
- Becker, B. 2009. *Model-based MASEM*. New York, NY: Russell Sage Foundation.
- Bergh, D. D., Aguinis, H., Heavey, C., Ketchen, D. J., Boyd, B. K., Su, P., Lau, C. L., & 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.
- Bergh, D. D., Boyd, B. K., Byron, K., Gove, S., & Ketchen, D. J. 2022. What constitutes a methodological contribution? *Journal of Management*, 48: 1835-1848.
- Bilgili, T. V., Calderon, C. J., Allen, D. G., & Kedia, B. L. 2017. Gone with the wind: A meta-analytic review of executive turnover, its antecedents, and postacquisition performance. *Journal of Management*, 43: 1966-1997.
- Bollen, K. A. 1989. *Structural equations with latent variables*. New York, NY: John Wiley & Sons.
- Bollen, K. A. 2012. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38: 37-72.
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25: 4279-4292.

- Burton, J. P., & Barber, L. K. 2019. The role of mindfulness in response to abusive supervision. *Journal of Managerial Psychology*, 34: 339-352.
- Cano, C. R., Carrillat, F. A., & Jaramillo, F. 2004. A meta-analysis of the relationship between market orientation and business performance: Evidence from five continents. *International Journal of Research in Marketing*, 21: 179-200.
- Chambless, D. L., & Hollon, S. D. 1998. Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, 66: 7-18.
- Chen, S.-C., & Liu, N.-T. 2019. When and how vicarious abusive supervision leads to bystanders' supervisor-directed deviance. *Personnel Review*, 48: 1734-1755.
- Cheung, M. W. 2015. MetaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, 5: 1521.
- Cheung, M. W. 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.
- Cheung, M. W., & Cheung, S. F. 2016. Random-effects models for meta-analytic structural equation modeling: Review, issues, and illustrations. *Research Synthesis Methods*, 7: 140-155.
- 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.
- Cooper, H., Hedges, L. V., & Valentine, J. C. 2009. *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.
- Darlington, R. B., & Hayes, A. F. 2017. *Regression analysis and linear models: Concepts, applications, and implementation*. New York, NY: The Guilford Press.
- Didelez, V., Meng, S., & Sheehan, N. A. 2010. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25: 22-40.
- Friend, S. B., Jaramillo, F., & Johnson, J. S. 2020. Ethical climate at the frontline: A meta-analytic evaluation. *Journal of Service Research*, 23: 116-138.
- Frone, M. R., Russell, M., & Cooper, M. L. 1994. Relationship between job and family satisfaction: Causal or non-causal covariation? *Journal of Management*, 20: 565-579.
- 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.
- Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. 2013. Meta-analytic review of employee turnover as a predictor of firm performance. *Journal of Management*, 39: 573-603.
- Harvey, P., Stoner, J., Hochwarter, W., & Kacmar, C. 2007. Coping with abusive supervision: The neutralizing effects of ingratiation and positive affect on negative employee outcomes. *The Leadership Quarterly*, 18: 264-280.
- Hedges, L. V., & Olkin, I. 2014. *Statistical methods for meta-analysis*. Academic Press.
- Hill, A. D., Johnson, S. G., Greco, L. M., O'boyle, E. H., & Walter, S. L. 2021. Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, 47: 105-143.
- Hünemund, P., & Louw, B. 2023. On the nuisance of control variables in causal regression analysis. *Organizational Research Methods*.
- Hunter, J. E., & Schmidt, F. L. 2004. *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.
- Jak, S., & Cheung, M. W. L. 2020. Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods*, 25: 430-455.
- Jak, S., & Cheung, M. W. L. 2018. Accounting for missing correlation coefficients in fixed-effects MASEM. *Multivariate Behavioral Research*, 53: 1-14.
- Kluemper, D. H., Mossholder, K. W., Ispas, D., Bing, M. N., Iliescu, D., & Ilie, A. 2019. When core self-evaluations influence employees' deviant reactions to abusive supervision: The moderating role of cognitive ability. *Journal of Business Ethics*, 159: 435-453.
- Krippendorff, K. 2018. *Content analysis: An introduction to its methodology* (4th ed.). Los Angeles: Sage.
- Landis, R. S. 2013. Successfully combining meta-analysis and structural equation modeling: Recommendations and strategies. *Journal of Business and Psychology*, 28: 251-261.

- Lian, H., Ferris, D. L., Morrison, R., & Brown, D. J. 2014. Blame it on the supervisor or the subordinate? Reciprocal relations between abusive supervision and organizational deviance. *Journal of Applied Psychology*, 99: 651-664.
- Lian, H., Lance Ferris, D., & Brown, D. J. 2012. Does taking the good with the bad make things worse? How abusive supervision and leader-member exchange interact to impact need satisfaction and organizational deviance. *Organizational Behavior and Human Decision Processes*, 117: 41-52.
- Mackey, J. D., Frieder, R. E., Brees, J. R., & Martinko, M. J. 2015. Abusive supervision: A meta-analysis and empirical review. *Journal of Management*, 43: 1940-1965.
- Mändli, F., & Rönkkö, M. 2023. To omit or to include? Integrating the frugal and prolific perspectives on control variable use. *Organizational Research Methods*.
- Mawritz, M. B., Greenbaum, R. L., Butts, M. M., & Graham, K. A. 2017. I just can't control myself: A self-regulation perspective on the abuse of deviant employees. *Academy of Management Journal*, 60: 1482-1503.
- Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. 2020. Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, 25: 243-258.
- Maydeu-Olivares, A., Shi, D., & Rosseel, Y. 2019. Instrumental variables two-stage least squares (2SLS) vs. maximum likelihood structural equation modeling of causal effects in linear regression models. *Structural Equation Modeling: A Multidisciplinary Journal*, 26: 876-892.
- Michel, J. S., Newness, K., & Duniewicz, K. 2016. How abusive supervision affects workplace deviance: A moderated-mediation examination of aggressiveness and work-related negative affect. *Journal of Business and Psychology*, 31: 1-22.
- Mitchell, M. S., & Ambrose, M. L. 2007. Abusive supervision and workplace deviance and the moderating effects of negative reciprocity beliefs. *Journal of Applied Psychology*, 92: 1159-1168.
- Mitchell, M. S., & Ambrose, M. L. 2012. Employees' behavioral reactions to supervisor aggression: An examination of individual and situational factors. *Journal of Applied Psychology*, 97: 1148-1170.
- Morris, T. P., White, I. R., & Crowther, M. J. 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38: 2074-2102.
- Muthén, B. O., & Asparouhov, T. 2018. Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, 47: 637-664.
- Restubog, S. L. D., Scott, K. L., & Zagenczyk, T. J. 2011. When distress hits home: The role of contextual factors and psychological distress in predicting employees' responses to abusive supervision. *Journal of Applied Psychology*, 96: 713-729.
- Røysamb, E., Nes, R. B., Czajkowski, N. O., & Vassend, O. 2018. Genetics, personality and wellbeing. A twin study of traits, facets and life satisfaction. *Scientific Reports*, 8: 1-13.
- Rupp, D. E., Shao, R. D., Jones, K. S., & Liao, H. 2014. The utility of a multifoci approach to the study of organizational justice: A meta-analytic investigation into the consideration of normative rules, moral accountability, bandwidth-fidelity, and social exchange. *Organizational Behavior and Human Decision Processes*, 123: 159-185.
- Schmidt, F. L., & Hunter, J. E. 2015. *Methods of meta-analysis: Correcting error and bias in research findings*. London: Sage.
- Seibert, S. E., Wang, G., & Courtright, S. H. 2011. Antecedents and consequences of psychological and team empowerment in organizations: A meta-analytic review. *Journal of Applied Psychology*, 96: 981-1003.
- Semadeni, M., Withers, M. C., & Trevis Certo, S. 2014. The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, 35: 1070-1079.
- Tepper, B. J. 2000. Consequences of abusive supervision. *Academy of Management Journal*, 43: 178-190.
- Tepper, B. J. 2007. Abusive supervision in work organizations: Review, synthesis, and research agenda. *Journal of Management*, 33: 261-289.
- Tepper, B. J., Carr, J. C., Breaux, D. M., Geider, S., Hu, C., & Hua, W. 2009. Abusive supervision, intentions to quit, and employees' workplace deviance: A power/dependence analysis. *Organizational Behavior and Human Decision Processes*, 109: 156-167.
- Tepper, B. J., Duffy, M. K., & Shaw, J. D. 2001. Personality moderators of the relationship between abusive supervision and subordinates' resistance. *Journal of Applied Psychology*, 86: 974-983.
- Tepper, B. J., Moss, S. E., Lockhart, D. E., & Carr, J. C. 2007. Abusive supervision, upward maintenance communication, and subordinates' psychological distress. *Academy of Management Journal*, 50: 1169-1180.
- Viswesvaran, C., & Ones, D. S. 1995. Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, 48: 865-885.

- 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.
- Yu, J. J., Downes, P. E., Carter, K. M., & O'Boyle, E. 2018. The heterogeneity problem in meta-analytic structural equation modeling (MASEM) revisited: A reply to Cheung. *Journal of Applied Psychology*, 103: 804-811.
- Zhang, Y., & Liao, Z. 2015. Consequences of abusive supervision: A meta-analytic review. *Asia Pacific Journal of Management*, 32: 959-987.
- Zhang, Y., Liu, X., Xu, S., Yang, L.-Q., & Bednall, T. C. 2019. Why abusive supervision impacts employee OCB and CWB: A meta-analytic review of competing mediating mechanisms. *Journal of Management*, 45: 2474-2497.