Steinmetz, Holger; Block, Jörn

Meta-analytic structural equation modeling (MASEM): new tricks of the trade

Management Review Quarterly

Provided in Cooperation with:
Springer Nature

Suggested Citation: Steinmetz, Holger; Block, Jörn (2022) : Meta-analytic structural equation modeling (MASEM): new tricks of the trade, Management Review Quarterly, ISSN 2198-1639, Springer International Publishing, Cham, Vol. 72, Iss. 3, pp. 605-626, https://doi.org/10.1007/s11301-022-00293-6

Management Review Quarterly (2022) 72:605–626
https://doi.org/10.1007/s11301-022-00293-6

EDITORIAL

Meta-analytic structural equation modeling (MASEM): new tricks of the trade

Holger Steinmetz[1] · **Jörn Block**[1,2,3]

Received: 29 July 2022 / Accepted: 1 August 2022 / Published online: 17 October 2022

Abstract
Meta-analytic structural equation modeling (MASEM) has become a widespread approach to meta-analyze the evidence in a field and to test a (theoretical) multivariate model based on aggregated data. This editorial presents new tricks of the trade and discusses current issues surrounding MASEM that in our view are either insufficiently recognized in the MASEM literature or present new developments. The editorial is organized in three parts, in which we discuss (a) the goals and causal underpinnings of a MASEM, (b) new possibilities to analyze effect size heterogeneities through moderator variables and (c) the use of statistical tests and *p* values.

JEL Classification C51 · C52

1 Introduction

MRQ aggregates empirical evidence about management issues in the form of (systematic) literature reviews, replications and meta-analyses. MRQ also strives to contribute to the methodological development of the management research field by discussing and advancing the state of the art of conducting literature reviews, replications, and meta-analyses. To this end, MRQ published several editorials (Block and Fisch 2020; Block and Kuckertz 2018; Fisch and Block 2018; Kuckertz and Block 2021) and articles (Block et al. 2022a; Clark et al. 2021). This editorial extends the practical guide for a meta-analysis by Hansen et al. (2012) to a more experienced audience. It focuses on new developments regarding meta-analytical structural

✉ Holger Steinmetz
steinmetzh@uni-trier.de

[1] Trier University, Trier, Germany

[2] Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] Wittener Institut für Familienunternehmen, Universität Witten/Herdecke, Witten, Germany

equation modeling (MASEM[1]), which is an important and valuable approach for meta-analysts.

MASEM involves specifying and testing a structural equation model (SEM) based on a meta-analytically derived matrix of all correlations among several model variables (Viswesvaran and Ones 1995). Analogous to the application of SEM to primary data, MASEM has several strengths (Steinmetz et al. 2020). For instance, a MASEM allows representing and testing theories in their breadth (Brown and Peterson 1993), comparing competing models or theories that entail different implications for the data (Harrison et al. 2006; Hom et al. 1992), testing mediators as postulated processes underlying an effect (Shadish 1996), and adjusting for variables that act as confounders (Pearl 1995; Tennant et al. 2020). An important benefit of a MASEM is that it can test models involving variables that were not included in each primary study.

While there is a substantial literature focusing on applications of MASEM in various fields (e.g., Chapman et al. 2005; Gonzalez-Mulé et al. 2017; Harrison et al. 2006; Hom et al. 1992; Murayama and Elliot 2012), most methodological articles on MASEM focus on procedural or statistical issues, for instance, the formation of the meta-analytical correlation matrix (Beretvas and Furlow 2006; Cheung and Chan 2005; Furlow and Beretvas 2005), the investigation of heterogeneity of effect sizes (Cheung 2008, 2018; Yu et al. 2016, 2018), or multilevel procedures addressing effect sizes nested in primary studies (Wilson et al. 2016). Recently, several discussions and new developments evolved in the SEM literature and the broader statistical literature, which are also of interest for MASEM applications. Our editorial discusses what these new developments imply for conducting a MASEM. The editorial is organized in three parts.

The first part concerns the goals of a MASEM as either a descriptive or a causal model. Our perception is that MASEM researchers are often not explicit enough about their underlying perspectives or goals. We discuss the inherent assumptions and challenges underlying a MASEM to be regarded as a causal model. This discussion is inspired by the recent introduction of graph-theoretical approaches to causal inference (Pearl 2009) in the SEM literature stressing the role of causal assumptions inherent in any SEM (Bollen and Pearl 2013; Kline 2016; Thoemmes et al. 2018). So far, this discussion has not been spilled over to the MASEM literature. We provide a short general introduction to causal modeling and its general implications, followed by a discussion of its implications for MASEM. The goal of this part is to help researchers to take a stance when building their model and to judge the degree of support for the overall model as well as specific effects.

The second part concerns the consideration of effect size heterogeneity by analyzing potential moderators. In the case of MASEM, this has traditionally been limited to categorical moderators (analyzed via multigroup MASEM). We discuss recent methodological developments to investigate continuous moderators of MASEM effects. In addition, we briefly introduce two statistical approaches (i.e., location-scale models and generalized additive models), which enrich the meta-analytical toolbox. Such bivariate meta-analytical investigations are often part of a MASEM study.

[1] We use the abbreviation MASEM for meta-analytical structural equation modeling as well as for denoting specific models.

The third part concerns the use of statistical tests and *p* values in a MASEM or an additional bivariate meta-analysis. This part is inspired by discussions in psychology and related disciplines resulting from the replication crisis (Maxwell et al. 2015) about statistical tests of parameters, the need to distinguish statistical from practical significance, the fallacy to conclude zero effects from non-significant test results (Lakens et al. 2018) or whether to use *p* values at all (Amrhein et al. 2019; Lakens 2021; Savalei and Dunn 2015).

2 Part I: Descriptive versus causal perspectives on MASEM

MASEM applies SEM to aggregated data. As with any SEM, its application requires an explicit stance towards either a descriptive/predictive or an explanatory goal. The implications and challenges of choosing one goal over the other should be clear to both the meta-analyst and the recipient: in the predictive mode, a causally correct specification matters less. For instance, predictor-outcome relationships can be substantially confounded or even spurious and nonetheless informative in predictive terms. In addition, the agglomerated nature of any meta-analytical procedure (i.e., the "apples-oranges problem") has fewer substantial implications when the goal of the meta-analysis is simply an overview of the aggregated or average associations. Yet, the interpretation of the results is limited to predictions and researchers should be careful not to interpret the results beyond their descriptive realm. In this regard, it is astonishing that in some fields, meta-analyses of bivariate correlations are regarded as the best available evidence and used when the research question requires a causal answer (Luthans 2011).

Matters differ, in contrast, if the researcher intends to interpret the coefficients causally. The conditions and assumptions, however, under which causal interpretations rest are insufficiently considered. In this respect, MASEM shares the fate of general SEM that over time experienced a shift from a causal towards a descriptive or predictive perspective.[2] This shift has resulted in models regarded as merely representing one of several options to describe a pattern of correlations among the variables (Bollen and Pearl 2013; Pearl 2012).

[2] While the historical roots of SEM clearly focused on causality (Pearl 2012; Wright 1921), SEM has experienced a shift towards data-centered descriptive models. From this perspective, a 'good' (i.e., acceptable) model is one that adequately describes the data (and not the one that correctly represents the causal assumptions). The reasons for this shift are manifold. Among others, they include the dominance of researchers with a primarily statistical perspective, criticisms of the often naive interpretation of parameters, or the long lasting influence of positivist views on the social sciences (see in particular Pearl and MacKenzie 2018). Associated with this shift was a change towards a non-causal language regarding the key concepts (e.g., "covariance structure analysis", "relationships", "regression effect"), goals (i.e., prediction, description, explanation of variance), and evaluation criteria (i.e., "acceptable" data fit). Only recently, causal theorizing has slowly started to re-emerge, strongly influenced by the attempts to incorporate graph theory (Elwert 2013; Pearl 2009; Pearl et al. 2016; Rohrer 2018) into SEM and the publication or update of influential textbooks introducing a change in the perspective on SEM (Kline 2016; Mulaik 2009; Shipley 2004).

Generally, three challenges exist to view a MASEM as a causal model. These concern (a) the potential of conceptual and causal heterogeneity due to aggregating data reflecting different concepts and causal structures, (b) the necessity to reflect upon causal assumptions, expressed by the set of estimated and fixed coefficients, as well as the testable implications of these assumptions, and (c) the necessity and opportunities of generating a causal identification strategy by incorporating control variables and instrumental variables into the model. As a background for the discussion, we briefly introduce graph theory (Pearl 2009) that has gained prominence in the social sciences and is a basis to consider and express causal assumptions.

2.1 Introduction to graph theory

In graph theory, a causal model is represented as a graph consisting of nodes representing constructs and edges reflecting assumed causal effects. By representing a causal model as a graph, the researcher specifies a set of claims about causal effects as well their absence (i.e., constraints). The latter is achieved by leaving out a causal link in the graph. The omission of a link represents the strong assumption (Bollen and Pearl 2013) of a non-existing effect similar to the *exclusion restriction* in econometrics (Angrist and Pischke 2015). Although the set of causal assumptions is not directly testable, the overall model provides implications for the data in terms of covariances and conditional independencies. The latter means that depending on the structure of the model, two variables become uncorrelated once other variables are statistically controlled for. These conditional independencies denote the *testable implications* of the model. A statistical chi-square test summarizes all testable implications and tests whether the data (i.e., the covariance matrix) deviate statistically from the model-implied covariance matrix, which results from the set of estimated effects and constraints. Contrary to predictive regression-type models, this dependence of the estimated parameters on the underlying causal assumptions gives causal meaning to the effects and clarifies the conditional nature of any causal conclusion as well as the relevance of critically testing the implications of the model (Bollen and Pearl 2013; Robles 1996). When it comes to the specification of a meta-analytical SEM, however, the literature lacks a discussion about the applicability of these theoretical notions and the challenges involved in regarding a MASEM as a causal model.

2.2 The challenge of conceptual and causal heterogeneity

Meta-analyses aggregate results for specific concepts to a broader class of constructs guided by some theoretical inclusion rationale. For instance, a meta-analysis focusing on predictors of overall firm performance would aggregate various distinct performance measures such as profit margin, return on assets, or return on investment to an umbrella performance construct. Such an aggregation, however, induces *conceptual heterogeneity* referred to as the "apples-oranges-problem" in the meta-analytical literature (Borenstein et al. 2009). In a MASEM, interpreting the effects resulting from averaging the effects of distinct concepts can at least be difficult or worse, nonsensical.

Another challenge concerns *causal heterogeneity* (Anoke et al. 2019; Athey and Imbens 2016; Pearl 2017), which is especially relevant for MASEM. Causal heterogeneity traditionally means that individuals differ in their responsiveness to a treatment while the average treatment effect represents the average across all individuals. In a MASEM, causal heterogeneity not only results from the aggregation of *populations* with different causal structures but also from the aggregation of *concepts* embedded in varying causal structures. For example, a MASEM may involve studies using a set of variables that reflect a full mediation structure versus a partial mediation versus a confounder or common cause structure. Causal heterogeneity can lead to a model misfit and can result in biased estimates. The bias increases with a larger number of different aggregated causal structures. Causal heterogeneity is a well-known problem in large sample SEM applications (Lubke and Muthén 2005; Muthén 1989) and occurs when a single SEM is estimated in a heterogeneous sample stemming from several populations. To summarize, MASEM researchers should carefully evaluate the extent of conceptual and causal heterogeneity in their sample of primary studies and should critically reflect whether conceptual or causal homogeneity can be justified.

2.3 The challenge of causal assumptions and testable implications

A causal model not only involves the specification of hypothesized effects but also of zero-effects. In addition, researchers routinely fix error covariances between a respective endogenous variable and an outcome variable to zero, which reflects the causal assumption of nonconfounding (i.e., there is no influence of an unmeasured variable on both variables). The role of these constraints is essential as they result in the *testable implications* of a model and allow distinguishing and choosing between alternative models.

For instance, a simple full mediation model implies several constraints that express the causal assumptions of the researcher about the data generating process (e.g., no direct effect, no confounding of all three pairs of variables and no reverse effects). This model has one testable implication, namely, the conditional independence of the explanatory variable and the outcome variable given the mediator. A successful test of this implication represents a contrast to alternative models with a direct effect of the explanatory on the outcome variable (in both directions), confounding of the mediator-outcome link, and a reverse effect of the outcome variable on the mediator. In contrast, a failed test may indicate a failure to meet one or several of these assumptions. Whereas the constraints allow testing the model, an increased saturation of a model by estimating many effects not only implies less (or no) testable implications but also increases the number of statistically equivalent models. For instance, a partial mediation model has no testable implications and hence, none of the discussed alternatives can be differentiated from the target model (Kline 2015; Thoemmes 2015). Even more importantly, while the full mediation structure has only four equivalent

models, the partial mediation structure has several dozens (Kline 2015). Box 1 provides a brief description of the background of testable implications.

Box 1: A Brief Description of the Background on Testable Implications of Causal Models

While the presented mediation model is a simple example, more complex models follow the same principles that are defined by the *d-separation rules* (Pearl, 2009) and the *path tracing rules* (Sewall, 1934). These rules describe characteristics of a path (i.e., any causal or non-causal link between any two variables) in a model—that is, correlations and conditional independencies. The path tracing rules simply state that any path creates a correlation between the variables unless at least one variable lying within the path is a *common effect* of other variables in the path. For instance, assume the path A → B → C ← D. According to the path tracing rules, A and C are correlated but A is uncorrelated with D as C is a common effect (or collider) in this path. Two things should be noted. First, two variables will often be linked by several paths, each probably implying and adding up to the overall correlation. As an example, B and D may be linked by a further mediator Z which, consequently, creates a correlation between A and D (via Z). Hence, each path has to be considered separately with regard to its correlation-inducing characteristics. Second, a variable can only ocurr once in a single path to avoid cycles.

In addition, the d-separation rules define whether two variables suddenly become *uncorrelated* once on or more variables lying on the same path are statisticaly controlled, stratified, or otherwise conditioned on. With regard to the example provided above, A and C become uncorrelated once B is controlled but in contrast, A and D become correlated once C is controlled. Both scenerios are denoted as path *blocking* of an open (i.e., correlation-inducing) path versus path *opening* of a closed (i.e., not correlation-inducing) path. The latter case explains such phenomena as supressor effects (Kim, 2019) or selection bias (Elwert & Winship, 2014). The former scenerio provides the basis for effectively block biasing paths created by confounders which we will discuss in the next section.

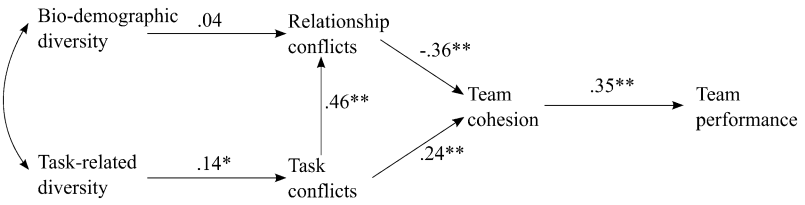
As models consist of a set of paths, a respective model may have several, one, or no testable implication depending on the level of saturation. These implications can either be tested in isolation (e.g., with partial correlations) or with the model chi-square test which provides an overall omnibus test of all implications. Because it is sometimes difficult to manually derive all testable implications of a model, the open source software DAGitty (www.dagitty.net, also contained in the R dagitty package, see Textor et al., 2016) can be used to create a path diagram of the target model and that prints all testable implications of this model.

Adopting such an explicit reflection on constraints and, hence, testable implications contrasts with common practices to discard a model's chi-square test due to the statistical power that is especially substantial when testing a highly-powered meta-analytical model. In this regard, researchers often conduct the "converse error" (McIntosh 2007, p. 1619), whereby the true statement that a trivial specification error leads to a significant test in a large sample is wrongly converted to the belief that a significant test in a large sample reflects a trivial specification error. While we realize that gaining a non-significant chi-square test may be difficult, especially in a MASEM, we likewise recommend to (a) reflect the testable implications and their underlying causal assumptions and (b) reconsider potential specification errors and alternative models that may lead to the observed misfit. For instance, Steinmetz et al. (2020) present the results of a meta-analytical mediation model and conclude that the effect of the mediator on the outcome variable has a more solid basis than the effects within the set of the two

antecedents of the mediator because a reversal of the outcome-mediator effect (e.g., A→B←C instead of A→B→C) would have been in conflict with the fitting full mediation model. Box 2 presents an overview of Steinmetz et al.'s (2020) discussion and illustrates the application of the mentioned principles.

Box 2: An Illustration

As an example, we discuss these issues in reference to an illustration presented by Steinmetz et al. (2020). The authors re-analysed parts of a prior project and specified a MASEM in the field of team diversity. The (recreated) graph is presented in the Figure.



The data underlying the model consisted of 52 published articles containing 211 correlation coefficients (N = 3,388 teams). As aforementioned, the model implied-correlations follow the logic defined by the path tracing rules. For instance, the model structure implies that there must be a correlation between task-related divsersity and team cohesion stemming from the indirect effect mediated by relationship conflicts and task conflicts. As the model does not specify direct effects and because task conflict is a central mediator, controlling for task conflict must eliminate the entire correlation. The latter would be one of the several testable implications. In total, the model has 12 testable implications that refer to the aforementioned d-separation rules. For instance, the model connects task conflicts to team performance via two paths—one via relationship conflicts and team cohesion and the other directly via team cohesion. Hence, the correlation between task conflicts and team performance would diminish, for instance, when team cohesion is controlled. Failure of this implication would suggest a residual relationship between both that may be, among others, due to an additional direct effect of task conflicts or unobserved confounding of the cohesion-performance link.

Steinmetz et al. concluded that the non-significant chi-square value was not *suspicious* which means that a significant test would have reflected an alarm signal potentially pointing to specification errors or lack of causal homogeneity. In the present example, the test was non-significant (χ^2(df) = 15.3(8), p = .05) although at the border of the alpha error level. However, since the effects of the two diversity predictors were weak, absence of direct effects on downstream variables (e.g., cohesion) can hardly by interpreted as causal support. Likewise, as the two conflict variables and cohesion formed a closed triangle, this part of the overall structure has no testable implications (Thoemmes, 2015). As a consequence, alternative permutations of effect directions among these three would not be distinguishable. The only exception refers to the effect of cohesion on team performance. Neither a spurious effect model (caused by an omitted confounder affecting both cohesion and performance) nor a model with a reverse effect would have been in concurrence with the fitting model as these would have implied arrows pointing *towards* cohesion and, due to the d-separation rules, a non-relationship between both conflicts and performance. From an econometric perspective, both conflicts may serve as instruments for cohesion and while the effect sizes of -.36 and .24 would not deliver them the status as strong instruments (Bound et al., 1995), the high power of the chisquare test would probably have resulted in a misfit of the model in case of unobserved confounding or reverse causation.

2.4 The challenge of developing a causal identification strategy

Auxiliary variables in the form of control variables or instruments are important for any empirical model (and thus, MASEM). They represent the researcher's causal identification strategy and are considered "to strip an observed association of all its spurious components" (Elwert 2013, p. 247). Unfortunately, the decision to include specific auxiliary variables into the model is often based on vague impressions about the role of the variables or simply the customs of the respective field. Graph theory (Pearl 2009) offers a theoretical framework to decide whether a particular auxiliary variable should be included into the model. Depending on its role, including a respective control variable may even introduce a bias. The assumed role of the auxiliary variable, thus, is part of the overall set of causal assumptions underlying the estimation of the model parameters. A number of conceptual and empirical articles provide procedures and guidelines for the consideration of control variables (Ferguson et al. 2020; Tennant et al. 2020; Vahratian et al. 2005).[3] Based on this literature, we provide guidelines for MASEM researchers to decide which variables to include into the model, which also serves as basis for discussing the limitations and implications of the study. The latter allows researchers to formulate more precise and informative limitations going beyond the often vague statement that the estimates of the MASEM should not be taken too seriously.

Figure 1 shows a comprehensive generic graph representing all classes of variables with their varying causal links to either the explanatory variable X or the outcome variable Y or both. Depending on its role, a variable has to be used as a control variable or *must not* be controlled to avoid bias of an otherwise unbiased effect. We note, however, that the figure represents a simplistic representation of all classes of variables and practical choices may differ when these variables are embedded in more complex structures.

The first class of variables concern confounders (class C1) that act as common causes of both X and Y (Elwert 2013). Depending on the sign of the target effect and the effects of C1, ignoring C1 will downward or upward bias the target effect. In the extreme case, the existence of a confounder will lead to a spurious target effect where in fact no effect exists. If C1 is not or cannot be measured or coded, there are two options. The first is the possibility to identify variables of class C2 ("surrogate confounders", see Tennant et al. 2020) that mediate the effect of C1 either on Y (as depicted in the figure) or X. Adjusting for C2 will—in graph-theoretical parlance—*block* the confounding path (Pearl 2012, see Box 1).

The second option is rather unknown in the SEM literature but the standard approach in econometrics (Angrist and Krueger 2001; Wooldridge 2012), namely the identification of *instrumental variables*. In Fig. 1, these are the variables of class W1 and W2. Instruments are variables which (a) are strongly related to X (relevance criterion), (b) have no direct effect on Y (exclusion restriction), and

[3] As a practical help, the DAGitty software allows specifying a model including presumed auxiliary variables and their expected role. Therefore, DAGitty will inform the researcher which auxiliary variables actually allow identifying the effect and whether this is possible at all.

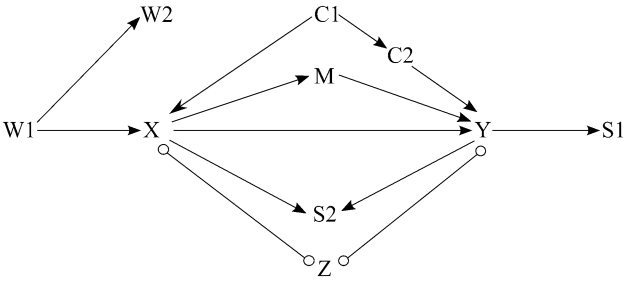


Fig. 1 Classes of potential auxiliary variables. *Note.* X = Independent variables, Y = dependent or outcome variable, C = confounders or surrogate confounders, W = instruments, S = selection factors or colliders, Z = variables with several plausible causal roles

(c) do not correlate with the error term and, hence, with omitted further causes of Y. Whereas econometricians traditionally analyze instruments with a two-stage-least-squares regression, instruments can also be incorporated in a SEM or MASEM (Maydeu-Olivares et al. 2019) and thereby enhance the estimation of the target effect if not all variables of class C1 or C2 can be considered. As Fig. 1 illustrates, being "related" to X can mean that the instrument either is a cause of X (i.e., W1) or shares a common cause (i.e. in the case of W2 sharing W1 as common cause).

Three further issues should be noted: first, simply adjusting for an instrument will lead to an amplification of a confounding bias of the X–Y effect and should be avoided (Ding et al. 2017; Steiner and Kim 2016). Second, even if W1 or W2 has an effect on Y, thus invalidating the exclusion restriction, adjusting for a mediator of this effect will turn the variable into a valid *conditional* instrument (Van Der Zander et al. 2015). Third, considerations of the role of variables as potential instruments are conducted on a theoretical level, not ensuring that the instrument can effectively be used in a finite sample with potentially limited statistical power or weak effect size (Bound et al. 1995). This difference reflects the importance of distinguishing between issues of causal identification and statistical estimation (Morgan and Winship 2007). For instance, W2 may principally be considered an instrument but the sample estimate of the relationship with X may be too low to actually employ it as such.

In contrast to these classes of auxiliary variables that have to be controlled or considered as instruments, the variables belonging to the remaining classes in Fig. 1 should not be included as controls. The variables of class M are mediators that transmit the effect of X on Y. Simply adjusting for M will block the indirect effect and give a biased impression of the total effect of X on Y (overcontrol bias, see Elwert 2013). However, adjusting for M may be a fruitful approach to rule out alternative pathways that are not the focus of the theoretical argument underlying a hypothesized effect. As a prominent example, experimental researchers have to rule out that the treatment effect is due to demand characteristics—that is, cues provided by the context of the experiment that inform participants what the goal of the experiment is (Orne 1962). From an SEM perspective, hence, measuring

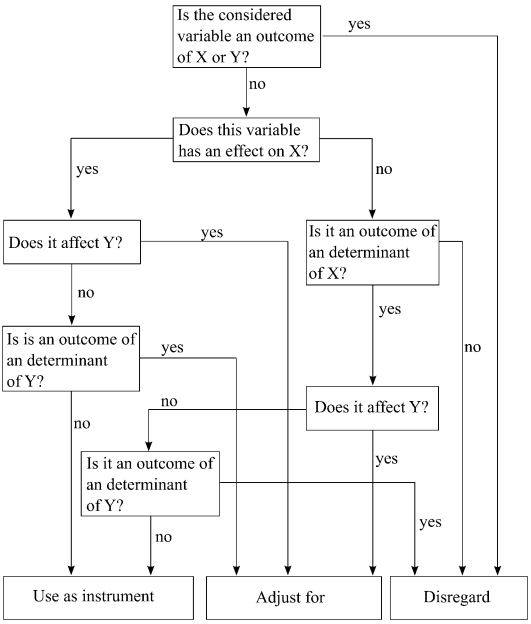


Fig. 2 Decision process underlying the consideration of auxiliary variables

and controlling for demand characteristics would reveal whether there is an indirect as well as a remaining direct effect. It should be noted that including M in a MASEM as an explicit mediator achieves all goals, that is, the estimation of the indirect and direct effects. However, it is the unreflected use of M as a control variable (i.e., specified as a common cause of X and Y or a covariate of X) that is disadvantageous.

The variables of class S1 represent selection factors (Elwert and Winship 2014). While selection effects are a danger for any empirical study and concern survey studies as well as interventions (in the form of attrition bias), controlling for a variable that is caused by the outcome will bias the effect of X on Y. Again, including S1 as a further outcome in the MASEM will not bias the effect as no adjustment of the target effect occurs. Variables of S2 are common outcomes of X and Y and represent colliders. Controlling for a collider will introduce bias.

Finally, the researcher may sometimes be uncertain about how the relationship between a potential control variable Z and either X or Y or both can be causally represented (symbolized as circles instead of arrow heads or origins). A potential avenue in such a scenario could be to run the model with and without controlling for Z. As a practical help, Fig. 2 presents a decision tree containing all critical questions a researcher has to answer when deciding whether to include a variable as a control variable or not. Apart from reflecting on the potential role of an already focused variable, a strategy is to start considering potential causes of X and then move forward from there. It should be noted, however, that the figure represents most but not all scenarios, for instance, situations in which one variable is a confounder as well as a collider.

Box 3: Critical Decisions when Considering Potential Auxiliary Variables

Figure 2 describes the theory-based considerations when deciding whether a variable should be adjusted for (i.e., used as a control variable) or used as an instrumental variable or whether it should be discarded either to avoid a bias or because the variable is irrelevant. As an example, a mediator variable should be discarded as a potential control variable simply as the answer to the first question is "yes" (i.e., the variable is an outcome of X). A second example would be that of a surrogate confounder (C2 in Figure 1): C2 is neither affected by X nor Y; it has no effect on X but its relationship with X results from the common influence of a determinant of X (i.e, C1 in Figure 1). As it has an effect on Y, C2 is involved in a non-causal, biasing path linking X with Y. Hence, C2 has to be adjusted for to block this path. As with Figure 1, we emphasize that the flowchart represents most common but not all possible scenarios. An example for the latter is when a variable is both a confounder and a collider (in different paths). We, hence, recommend to create a full graph involving the relevant variables and their interrelationships.

Once control variables have been identified, the question arises how to incorporate them into the model—especially if the model contains several target variables in a more complex system. One possibility is that the researcher specifies only those effects of the control variable on its expected outcomes variables and fixes all other effects to zero. The other scenario is to estimate all effects of the control variable on all model variables in the SEM. The advantage of the first approach is that the approach is parsimonious and represents a strong match with the researcher's theoretical assumptions. However, the approach adds new constraints to the already existing constraints concerning the target variables. While this is overall beneficial as it increases the testable implications, failure and misfit may be due to specification errors in the structure of the control variables. The second scenario will block the biasing effects of all considered confounders but makes it more difficult to notice specification problems by means of a misfit. A good compromise is to start theoretically considering a full set of effects of the control variables and then disregard those that are theoretically unlikely (Ferguson et al. 2020). We emphasize that this is done in the specification phase of the project and should not be confused with the bad practice to estimate all effects and then eliminate non-significant coeffcients.

3 Part II: Analyses of heterogeneity in MASEM effects and bivariate meta-analyses

The analysis of heterogeneity is an important goal of any meta-analysis. While traditional approaches towards analyzing heterogeneity for complete MASEMs were restricted to subgroup analysis and a multigroup approach where the groups refer to levels of a categorical moderator, the approach by Jak and Cheung (2020) allows to analyze continuous moderators. However, even this approach assumes the *linearity* of the moderator effect (e.g., the effect size of interest linearly decreases or increases with the level of the moderator variable). Recently, two approaches have been introduced to meta-analysis and incorporated in the R package metafor (Viechtbauer 2010), namely the application of nonlinear meta regression analysis by means of generalized additive models (Wood 2017) and location-scale models (Viechtbauer

and López-López 2021). While these two approaches focus on bivariate correlations (and not MASEM parameters), they should help to understand the heterogeneity of target relationships.

3.1 Testing moderators of MASEM effects

When the focus is on singular effect sizes, for instance, on standardized mean differences or correlation coefficients, researchers usually employ a subgroup analysis or meta-regression. In a subgroup analysis, one conducts a separate meta-analysis for each category of a categorical moderator (e.g., student samples vs. employee samples) and the results are compared in a descriptive way or by means of a statistical difference test. In contrast, a meta-regression includes continuous or categorical moderators (as dummies) which allows to estimate their effects while controlling for other moderators and avoiding the reduction in sample size inherent in subgroup analysis (Thompson and Higgins 2002).

While the analysis of moderators is straightforward in these singular-effect-size scenarios, things are more complicated in case of a MASEM. For categorical moderators, one approach is to separate studies according to the levels of the moderator and run the MASEM in each of the sub-samples. For instance, Steinmetz et al. (2021) conducted a MASEM in several sub-samples (e.g., student samples vs. broad samples) and found little differences in the structural effects. Such an approach, however, requires that all possible pairs of correlations are available in the sub-samples, which is often not the case.

Whereas such a sub-sample analysis is possible with categorical moderators, it is not recommended in the case of continuous moderators. Creating artificial subgroups or subsamples along some cut off points (e.g., a dichotomization along a median split) leads to a loss of information and, thus, power resulting from grouping diverse studies or populations into seemingly homogeneous subgroups (MacCallum et al. 2002). A common solution to this problem is to estimate the MASEM for the total sample of studies, and then to analyze moderators in a meta-regression analysis with the bivariate correlation coefficient (instead of the structural effect) as the dependent variable.

In a recent article, Jak and Cheung (2020) presented an approach that allows testing for categorical or continuous moderators of the structural effects of the MASEM instead of the bivariate correlations. Their approach works by integrating study-specific correlation matrices and effects with the estimation of the overall MASEM based on the average correlation matrix. Further, the consideration of the study specific matrices allows testing the interaction between the study-specific effects and the moderator.[4]

[4] To ease the diffusion and applicability of the approach, Jak and Cheung provide an open-source shiny app that can be run online (via https://sjak.shinyapps.io/webMASEM/) or locally on the researcher's computer after downloading and running the code of the software R provided on https://osf.io/x8y7f/. Shiny apps establish an interactive application with a graphical interface that allow researchers to import data sets and run procedures—in our case the overall MASEM and moderator analyses.

A benefit of this approach is its characteristic as a random effect model that allows to reflect true differences between the primary studies and to isolate them from mere sampling errors. Missing variables (and, thus, missing correlations among them) in the primary studies are considered by estimating the model with a full information maximum likelihood estimator (Allison 2003), assuming that the mechanism causing missing correlations is either missing completely at random or missing at random. That is, the missingness of the correlation should not be caused by the size of that correlation (see an easy graph-theoretical introduction to missing data in Thoemmes and Mohan 2015). In such a scenario, the missingness is non-random and the application of the approach induces biases.

The most important limitation of the approach by Jak and Cheung (2020) is the impossibility to consider multiple effect sizes per study (Wilson et al. 2016). Multiple effect sizes occur when several specific effect sizes are coded as representing the same umbrella construct. For instance, there are multiple ways to conceptualize firm performance (e.g., profit margin, return on assets). If a primary study uses several of these variables but the meta-analyst intends to analyze relationships on an aggregate and overall performance level, then the meta-analyst faces the problem of how to treat these multiple effect sizes. In the current debate on how to deal with multiple observations per study, using all available correlations but considering the nested structure of the data (e.g., via multilevel modeling) is favored (see a short discussion in Hansen et al. 2012). The approach by Jak and Cheung so far does not allow integrating a multilevel perspective (but this is planned in the near future).[5]

3.2 Meta regressions with generalized additive models (GAMs)

Generalized additive models (GAMs) are a progression from generalized *linear* models, that again generalize the traditional linear model with Gaussian errors to several types of distributions (e.g., Poisson, Binomial, Gamma). GAMs build on these type of approaches by generalizing the functional relationship between two variables to nonlinear functions of one or several predictor variables. While nonlinear effects can be incorporated into linear models by introducing polynomials, such an approach has limitations (Wood 2017). Most notably, real data often do not concur with the strict functional form implied by the polynomial, which leads to misfit, a lack of power, and bias. GAMs enable fitting nonlinear smoothing functions, whose exact form is estimated from the data and, thus, is not determined ex ante by the researcher.

In a meta-analytical context, the use of a meta-GAM would allow to estimate whether the relationship between the predictors and the effect size is nonlinear or whether a linear relationship would be sufficient. In the case of a nonlinear moderator effect, GAMs have a stronger statistical power to reject the null of a zero

[5] In the meantime, researchers could first estimate an overall MASEM (under the assumption of structural homogeneity) with a multilevel approach (Wilson et al. 2016) and then estimate a moderator model by using within-study composites. While this is clearly not optimal, it allows testing for moderator effects, which will represent the average effect across the multiple effect sizes aggregated. In addition, the comparison of both approaches will allow evaluating the relevance of the multilevel approach.

moderator effect. More importantly, some nonlinear (e.g., curvilinear) relationships will not be detected with a linear model even when the statistical power is strong.

3.3 Location-scale models

Location-scale (LS) models are a "new trick for the trade" recently added to the meta-analyst's toolbox (Viechtbauer and López-López 2021). LS models come into play when one of the core assumptions of the linear model (and thus, meta-regression) fails. Whereas the standard linear model assumes homoscedasticity in order to estimate standard errors, which are the basis for statistical tests and the computation of confidence intervals, LS models are appropriate when heteroscedasticity varies as a function of the same or other predictors of the meta-regression model.

LS models consist of two parts. The location part focuses on the outcome of any linear model—that is, the conditional mean of the outcome variable modelled as a function of the predictors. A nonzero estimate in this part would signal that the effect size constantly increases or decreases with the predictor. The scale-part, in contrast, focuses on the conditional variance. A nonzero estimate in this part would signal that the variance of effect sizes across studies systematically increases or decreases with the predictor. Both components are integrated in one analytical approach and may concern the same or different predictors for both components. A potential (fictitious) example in the meta-analytical field could be a moderator effect of the GDP of a country on both the mean level of a relationship between two target variables as well as their variance across studies. The location part of the LS model would answer the question whether the relationship constantly increases or decreases with increasing GDP—the scale part would reveal whether studies in countries with lower GDP have a stronger variation compared to countries with higher GDP (e.g., due to resource constraints). Hence, LS models can be a fruitful approach to enlarge the tools available to analyse the heterogeneity of study results.

4 Part III: Statistical tests

4.1 Statistical significance versus practical significance

Apart from the test of the overall model, the individual effect sizes matter. Researchers typically report the average estimated effect size, the associated standard error, and test whether the estimate is statistically different from zero. Yet, due to the high-powered meta-analytic data, solely focusing on statistical significance creates the problem of statistically significant albeit practically trivial estimates. Consequently, relying on null-hypothesis testing as a test of a theoretically based hypothesis and practically relevant effect is only partially informative, particularly with large samples of primary studies. In one of our own recent MASEMs (Block et al. 2022b), involving sample sizes of over $N = 100{,}000$ firms, for instance, resulted in standardized effects of $\beta = 0.01$ being significant on an alpha level of 0.05.

In addition to issues of triviality and practically insignificant results, there is a large literature on typical problems associated with using significance tests. These issues include, among others, the difficulties to correctly interpret the *p* value (see next section), the application of a mindless ritual including using a fixed alpha level or the tendencies to interpret very small *p* values as a form of effect size in itself (Gigerenzer 2004). Based on these criticisms and despite the value of null-hypothesis testing (Lakens 2021; Savalei and Dunn 2015), statisticians have recommended placing more emphasis on the confidence intervals. Likewise, scientific associations such as the American Psychological Association have recommended their use explicitly for decades (APA 2002; Wilkinson 1999). Acknowledging the widespread emphasis that researchers routinely also misinterpret confidence intervals (Hoekstra et al. 2014) as the interval including the true parameter with a 0.95 probability,[6] a recent study by Amrhein et al. (2019) recommended interpreting and using it as a *compatibility interval,* that is, the range of parameters being compatible with the data. Such an interpretation and its increased emphasis especially in meta-analyses has the following advantages.

First, confidence intervals reflect the implicit characteristics of any estimation process resulting in an estimate surrounded by tremendous uncertainty. The confidence interval clearly communicates this uncertainty and may create caution on the side of the researcher when interpreting the results. We realize that existing studies investigating the interpretation of confidence intervals present mixed results for this claim (Savalei and Dunn 2015) but nonetheless recommend using the interval in this regard. Second, relying on the confidence interval increases the value of meta-analyses and their updates (Lakens et al. 2016), as increasing the empirical basis for the target coefficient will increase its precision and reduce uncertainty. As meta-analysts, we have experienced several times a rejection of a meta-analysis study, as reviewers doubted its usefulness in cases where primary studies had found support for their hypothesis. Focusing on a confidence interval would clarify the substantial increase in precision when meta-analysing results of primary studies.

Finally, while we still think that null hypothesis testing has its merits, having a highly powered meta-analysis will lead to small coefficients becoming significant. Hence, issues of practical versus statistical significance become prevalent and require the interpretation of coefficients with regard to the application of choice. For instance, some small coefficients may be practically insignificant with regard to their causal or predictive role but can gain significance on a time-related or hierarchically aggregated level. An example for a time-related aggregation is that estimated effects always reflect some timely interval. Even if these effects are small, aggregating several of the time spans can result in a substantial long-term effect in cases where the effects accumulate. Such an aggregation, however, depends on the dynamic characteristics of both cause and effect and makes longitudinal studies and meta-analyses

[6] Such an interpretation contradicts frequentist statistics and would regard the interval as fixed but the parameters as a randomly (from sample to sample) varying variable. However, the opposite is true and the parameter is a fixed entity while the estimated interval is the randomly varying entity. Hence, it is not the parameter which has a probability to be in the interval but the interval which has a probability of including the parameter.

valuable. An example for a hierarchical aggregation is a study showing estimated effects that represent the role of a variable for the average individual unit. While this estimated effect on the unit level may be negligible, the implications for a set of many units (e.g., a whole industry) can be substantial due to the aggregation across these units. This is comparable with a low effect of a drug on a person but a substantial relevance across a large number of individuals in a society. The key issue is however, the difference between the levels of analysis and researchers should be careful not to commit an atomistic or ecological fallacy.

4.2 Testing the null hypothesis (somewhat)

An ongoing problem when interpreting the results of statistical tests in primary studies or meta-analyses is how to interpret a non-significant coefficient. The most prevalent interpretation is that the effect is zero or there is no mean difference between the analysed groups (Lakens 2017) whereas the adequate perspective would be that the study only could not find enough evidence to reject the H0 of a null effect. Likewise, it sometimes occurs that the researcher explicitly states a null hypothesis. For instance, when analyzing the relation between management rank, knowledge, and performance, Hunton et al. (2000) formulated the hypothesis that "among manager-level managerial accountants, there will be no relationship between technical (managerial accounting) knowledge and job performance" (p. 754). Testing for exact values of a coefficient, however, is impossible as the *p* values of significance always reflect the integral of the sampling distribution *equal or larger* a test criterion (e.g., the critical z value of 1.968 that reflects a probability of 0.05 being equal or larger than this value under the true effect of $z = 0$).

Nonetheless, concluding a null effect based on empirical analyses is theoretically and practically fruitful (Stanton 2020). A practical solution to the aforementioned impossibility to test an exact zero-effect is to test its *equivalence with zero* (Lakens 2017; Lakens et al. 2018), which means that the researcher specifies an interval around zero that can be regarded as equivalent to zero from a theoretical or practical perspective (Wellek 2002). Such an interval requires the researcher to clarify what counts as meaningful. An equivalence test is conducted with the TOST (two-one-sided-tests) procedure initially developed by Schuirmann (1987). In this approach, the researcher specifies the aforementioned interval around the null effect, with the lower (θ_{low}) and upper bound (θ_{high}) marking the boundaries of this interval that contain irrelevant values of θ. Then, two compound one-sided H0 tests are performed; one testing the alternative hypothesis that the true effect size θ is statistically smaller than θ_{high} (H0: $\theta \geq \theta_{high}$) and the other that the observed effect size is statistically larger than θ_{low} (H0: $\theta \leq \theta_{low}$). If the compound test results in a *p* value smaller than the decision criterion (e.g., $p = 0.05$), then the true effect size must lie between both boundaries and can be considered equivalent to zero with an overall type I error probability of 0.05.

Lakens notes that equivalence tests that test some pre-described area lead to more substantial hypotheses than classical hypothesis tests, in which the conclusion is merely "nonzero". It should, however, be noted that the practical usefulness of equivalence tests rises and falls with the sample sizes and their consequential

statistical power and will result in a non-significant result in situations where the true effect size is very close to zero. This fact however, highlights the TOST approach as an ideal approach for meta-analyses, where, due to the large sample size, trivial mean effect sizes may become statistically significant. However, even if the procedure is conducted in studies with a moderate sample size, non-significant results will increase the researcher's sensitivity towards the role of power for interpreting classical non-significant effects. Consequently, erroneous conclusions of a zero effect based on a non-significant classical test will be avoided.

Box 4: The TOSTER Package in R

Lakens (2017) has created the R package "TOSTER" that contains a number of functions for different kinds of effect sizes (e.g., correlations, Cohen's D) in which the measured effect size, standard error, and boundaries of the interval are fed in. The functions perform a classical null-hypothesis test as well as an equivalence test and report p-values and confidence intervals of both. Naturally, the choice of the boundary is critical and has to be informed by practical and theoretical criteria (see Stanton, 2020, for a discussion). Often, however, theoretical and practical issues can depart because theoretically small and irrelevant effect sizes can nonetheless be practically relevant, especially when considering a timely or hierarchically aggregated form (as discussed in the former section). One disadvantage--especially from the perspective of conducting a MASEM--is that the package does not allow to test the null equivalence of a regression or structural coefficient. This, however, can be bypassed by calculating the partial correlation coefficient from the meta-analytic correlation coefficients and applying TOSTER to this coefficient. Partial correlation coefficients can be computed in the R software by means of the partial.r-function in the psych-package (Revelle, 2022). Knowledge which variables to include in the conditioning set, again, can be based on the d-separation rules discussed earlier (Shipley, 2004).

4.3 Bayesian approaches towards meta-analyses

While null-hypothesis significance testing (NHST) is the standard in most disciplines, the use of Bayesian statistics, either as an addition or as an alternative, is on the rise. And whereas statisticians arguing for one or the other approach routinely engaged in heated debates, practical researchers were either unaware of these discussions or, due to the lack of easily applicable software solutions, refrained from applying such approaches. Since the latter has considerably changed in the last years, applying a Bayesian approach to regression (Kruschke et al. 2012; Winter and Bürkner 2021) or meta-analysis (Anderson and Maxwell 2016; Williams et al. 2018) has become possible for applied researchers. One suitable, comprehensive, and thus powerful approach has been provided in form of the brms package that is run within the open source software R (Bürkner 2017; Nalborczyk et al. 2019).

Bayesian meta-analyses can be applied as a means to estimate the subjective probability resulting from updating prior beliefs with new data for a range of effect sizes. Bayesian methods rely on Bayes' theorem in probability theory (Bayes 1763) and proceed in three steps: First, a priori beliefs (from theory or prior empirical research) about the relationship of interest are formulated (the prior). Next, a

probability of occurrence of the data given the prior is assumed (the likelihood function). Then, data are used to update the prior. The result is the posterior distribution, which is a probability density function of the effect size of interest. Hence, Bayesian meta analysis does not simply provide a point estimate or a confidence interval but rather an entire distribution function showing how likely an effect is. With Bayesian analysis, it is thus possible to make statements in terms of likely or unlikely effects, which is not possible with NHST but very intuitive for practitioners. In addition, especially when it comes to updates of former meta-analyses, Bayesian approaches have specific advantages. The results of the former meta-analysis can be used as a prior and the result of the Bayesian meta-analysis with the updated data tells the researcher whether the estimated relationships have changed (substantially) and an update was necessary. It should be noted, however, that even though focusing on the posterior distribution and defined ranges of values and their probability is attractive, such a focus comes at the cost of sacrificing a clearly defined error rate that the researcher is willing to accept.

5 Conclusion

In this editorial, we discussed several new developments in the meta-analytical field and the broader applies statistical literature with a focus on MASEM and accompanying bivariate meta-analytical investigations. As with any empirical approach, clarifying the epistemological goals of the analysis is most central, in particular whether the goal of the MASEM is descriptive, predictive or causal. With a causal goal, a meta-analysis of experimental primary studies are the optimal approach and golden standard. With non-experimental studies, a MASEM has advantages over the mere aggregation of correlation coefficients. For this case, our editorial presents a number of challenges, opportunities, and practical guidelines along the "new tricks of the trade" to improve the validity and quality of the MASEM.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors have not disclosed any funding.

Declarations

Competing interests The authors have not disclosed any competing interests.

References

Allison PD (2003) Missing data techniques for structural equation modeling. J Abnorm Psychol 112(4):545–557. https://doi.org/10.1037/0021-843X.112.4.545

Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. Nature 567(7748):305–307. https://doi.org/10.1038/d41586-019-00857-9

Anderson SF, Maxwell SE (2016) There's more than one way to conduct a replication study: beyond statistical significance. Psychol Methods 21(1):1–12. https://doi.org/10.1037/met0000051

Angrist JD, Krueger AB (2001) Instrumental variables and the search for identification: from supply and demand to natural experiments. J Econ Perspect 15(4):69–85. https://doi.org/10.1257/jep.15.4.69

Angrist JD, Pischke J-S (2015) Mastering metrics. Princeton University Press. https://doi.org/10.1017/CBO9781107415324.004

Anoke SC, Normand SL, Zigler CM (2019) Approaches to treatment effect heterogeneity in the presence of confounding. Stat Med 38(15):2797–2815

APA (2002) Publication manual of the american psychological association, 6th edn. American Psychological Association

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci 113(27):7353–7360. https://doi.org/10.1073/pnas.1510489113

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 53:370–418

Beretvas SN, Furlow CF (2006) Evaluation of an approximate method for synthesizing covariance matrices for use in meta-analytic SEM. Struct Equ Model 13(2):153–185

Block J, Fisch C (2020) Eight tips and questions for your bibliographic study in business and management research, vol 70. Springer, pp 307–312

Block J, Kuckertz A (2018) Seven principles of effective replication studies: strengthening the evidence base of management research, vol 68. Springer, pp 355–359

Block J, Fisch C, Kanwal N, Lorenzen S, Schulze A (2022a) Replication studies in top management journals: an empirical investigation of prevalence, types, outcomes, and impact. Manag Rev Q. https://doi.org/10.1007/s11301-022-00269-6

Block J, Hansen C, Steinmetz H (2022b) Are family firms doing more innovation output with less innovation input? A replication and extension. Entrep Theory Pract. https://doi.org/10.1177/10422587221084249

Bollen KA, Pearl J (2013) Eight myths about causality and structural equation modeling. In: Morgan SL (ed) Handbook of causal analysis for social research. Springer, pp 301–328

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Introduction to meta-analysis. Wiley

Bound J, Jaeger DA, Baker RM (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc 90(430):443–450. https://doi.org/10.1080/01621459.1995.10476536

Brown SP, Peterson RA (1993) Antecedents and consequences of salesperson job satisfaction: meta-analysis and assessment of causal effects. J Mark Res 30:63–77

Bürkner P-C (2017) Advanced Bayesian multilevel modeling with the R package brms. arXiv preprint arXiv:1705.11123

Chapman DS, Uggerslev KL, Carroll SA, Piasentin KA, Jones DA (2005) Applicant attraction to organizations and job choice: a meta-analytic review of the correlates of recruiting outcomes. J Appl Psychol 90(5):928–944

Cheung MWL (2008) A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. Psychol Methods 13(3):182–202

Cheung MW-L (2018) Issues in solving the problem of effect size heterogeneity in meta-analytic structural equation modeling: a commentary and simulation study on Yu, Downes, Carter, and O'Boyle (2016). J Appl Psychol. https://doi.org/10.1037/apl0000284

Cheung MWL, Chan W (2005) Meta-analytic structural equation modeling: a two-stage approach. Psychol Methods 10(1):40–64

Clark WR, Clark LA, Raffo DM, Williams RI (2021) Extending Fisch and Block's (2018) tips for a systematic review in management and business literature. Manag Rev Q 71(1):215–231

Ding P, VanderWeele T, Robins J (2017) Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika 104(2):291–302. https://doi.org/10.1093/biomet/asx009

Elwert F (2013) Graphical causal models. In: Morgan SL (ed) Handbook of causal analysis for social research. Springer, pp 245–273
Elwert F, Winship C (2014) Endogenous selection bias: the problem of conditioning on a collider variable. Ann Rev Sociol 40:31–53. https://doi.org/10.1146/annurev-soc-071913-043455
Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, Lewsey JD (2020) Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. Int J Epidemiol 49(1):322–329
Fisch C, Block J (2018) Six tips for your (systematic) literature review in business and management research. Manag Rev Q 68(2):103–106
Furlow CF, Beretvas SN (2005) Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. Psychol Methods 10(2):227–254
Gigerenzer G (2004) Mindless statistics. J Socio-Econ 33:587–606
Gonzalez-Mulé E, Carter KM, Mount MK (2017) Are smarter people happier? Meta-analyses of the relationships between general mental ability and job and life satisfaction. J Vocat Behav 99:146–164. https://doi.org/10.1016/j.jvb.2017.01.003
Hansen C, Steinmetz H, Block J (2012) How to conduct a meta-analysis in eight steps: a practical guide. Manag Rev Q 72:1–19
Harrison DA, Newman DA, Roth PL (2006) How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. Acad Manag J 49(2):305–325
Hoekstra R, Morey RD, Rouder JN, Wagenmakers E-J (2014) Robust misinterpretation of confidence intervals. Psychon Bull Rev 21(5):1157–1164
Hom PW, Caranikas-Walker F, Prussia GE, Griffeth RW (1992) A meta-analytical structural equations analysis of a model of employee turnover. J Appl Psychol 77:890–909. https://doi.org/10.1037/0021-9010.77.6.890
Hunton JE, Wier B, Stone DN (2000) Succeeding in managerial accounting. Part 2: a structural equations analysis. Account Organ Soc 25(8):751–762
Jak S, Cheung MW-L (2020) Meta-analytic structural equation modeling with moderating effects on SEM parameters. Psychol Methods 25(4):430–455. https://doi.org/10.31234/osf.io/ce85j
Kim Y (2019) The causal structure of suppressor variables. J Educ Behav Stat. https://doi.org/10.3102/1076998619825679
Kline RB (2015) The mediation myth. Basic Appl Soc Psychol 37(4):202–213. https://doi.org/10.1080/01973533.2015.1049349
Kline RB (2016) Principles and practice of structural equation modeling, vol 156, 4 edn. The Guilford Press
Kruschke JK, Aguinis H, Joo H (2012) The time has come: Bayesian methods for data analysis in the organizational sciences. Organ Res Methods 15(4):722–752
Kuckertz A, Block J (2021) Reviewing systematic literature reviews: ten key questions and criteria for reviewers, vol 71. Springer, pp 519–524
Lakens D (2017) Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Pers Sci 8(4):355–362. https://doi.org/10.1177/1948550617697177
Lakens D (2021) The practical alternative to the p value is the correctly used p value. Perspect Psychol Sci 16(3):639–648. https://doi.org/10.1177/1745691620958012
Lakens D, Hilgard J, Staaks J (2016) On the reproducibility of meta-analyses: six practical recommendations. BMC Psychology 4(1):24
Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psychol Sci 1(2):259–269. https://doi.org/10.1177/2515245918770963
Lubke G, Muthén B (2005) Investigating population heterogeneity with factor mixture models. Psychol Methods 10(1):21–39
Luthans F (2011) Organizational behavior: an evidence-based approach. McGraw-Hill, Inc.
MacCallum RC, Zhang S, Preacher KJ, Rucker DD (2002) On the practice of dichotomization of quantitative variables. Psychol Methods 7(1):19–40
Maxwell SE, Lau MY, Howard GS (2015) Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am Psychol 70(6):487
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. Struct Equ Model Multidiscip J 26(6):876–892
McIntosh C (2007) Rethinking fit assessment in structural equation modeling: a commentary and elaboration on Barrett (2007). Person Individ Differ 42(5):859–867

Morgan SL, Winship C (2007) Counterfactuals and causal inference: methods and principles for social research (analytical methods for social research). Cambridge University Press
Murayama K, Elliot AJ (2012) The competition-performance relation: a meta-analytic review and test of the opposing processes model of competition and performance. Psychol Bull 138(6):1035–1070. https://doi.org/10.1037/a0028324
Muthén BO (1989) Latent variable modeling in heterogeneous populations. Psychometrika 54(4):557–585
Nalborczyk L, Batailler C, Lœvenbruck H, Vilain A, Bürkner P-C (2019) An introduction to Bayesian multilevel models using brms: a case study of gender effects on vowel variability in standard Indonesian. J Speech Lang Hear Res 62(5):1225–1242
Orne MT (1962) On the social psychology of the psychological experiment with particular reference to demand characteristics and their implications. Am Psychol 17:776–783
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82(4):669–688. https://doi.org/10.1093/biomet/82.4.669
Pearl J (2009) Causality: models, reasoning, and inference. Cambridge University Press
Pearl J (2012) The causal foundations of structural equation modeling. In: Hoyle RH (ed) Handbook of structural equation modeling. Guilford Press, pp 68–91
Pearl J (2017) Detecting latent heterogeneity. Sociol Methods Res 46(3):370–389. https://doi.org/10.1177/0049124115600597
Revelle WR (2022).psych: procedures for personality and psychological research. https://CRAN.R-project.org/package=psych Version 2.2.3. Northwestern University, Evanston, Illinois, USA
Robles J (1996) Confirmation bias in structural equation modeling. Struct Equ Model 3(1):73–83. https://doi.org/10.1080/10705519609540031
Savalei V, Dunn E (2015) Is the call to abandon p-values the red herring of the replicability crisis? [Opinion]. Front Psychol. https://doi.org/10.3389/fpsyg.2015.00245
Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15(6):657–680
Sewall W (1934) The method of path coefficients. Ann Math Stat 5:161–215
Shadish WR (1996) Meta-analysis and the exploration of causal mediating processes: a primer of examples, methods, and issues. Psychol Methods 1(1):47–65
Shipley B (2004) Cause and correlation in biology. A user's guide to path analysis, structural equations and causal inference. Cambridge University Press. https://doi.org/10.1007/s13398-014-0173-7.2
Stanton JM (2020) Evaluating equivalence and confirming the null in the organizational sciences. Organ Res Methods 24:491–512
Steiner PM, Kim Y (2016) The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. J Causal Inference 4(2):1–22
Steinmetz H, Bosnjak M, Isidor R (2020) Meta-analytische Strukturgleichungsmodelle: Potenziale und Grenzen illustriert an einem Beispiel aus der Organisationspsychologie [Meta-analytical structural equation modelling: potentials and limitations illustrated with an example from organizational psychology]. Psychol Rundsch 71:111–118
Steinmetz H, Isidor R, Bauer C (2021) Gender differences in the intention to start a business: an updated and extended meta-analysis. Zeitschrift Für Psychologie 229(1):70–84. https://doi.org/10.1027/2151-2604/a000435
Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, Ellison GTH (2020) Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. https://doi.org/10.1093/ije/dyaa213
Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT (2016) Robust causal inference using directed acyclic graphs: the R package 'dagitty.' Int J Epidemiol 45(6):1887–1894
Thoemmes F (2015) Reversing arrows in mediation models does not distinguish plausible models. Basic Appl Soc Psychol 37(4):226–234. https://doi.org/10.1080/01973533.2015.1049351
Thoemmes F, Mohan K (2015) Graphical representation of missing data problems. Struct Equ Model 22(4):631–642
Thoemmes F, Rosseel Y, Textor J (2018) Local fit evaluation of structural equation models using graphical criteria. Psychol Methods 23(1):27–41. https://doi.org/10.1037/met0000147
Thompson SG, Higgins JPT (2002) How should meta-regression analyses be undertaken and interpreted? Stat Med 21:1559–1573. https://doi.org/10.1002/sim.1187)
Van Der Zander B, Textor J, Liškiewicz M (2015) Efficiently finding conditional instruments for causal inference. In: IJCAI International Joint Conference on Artificial Intelligence (IJCAI), pp 3243–3249

Vahratian A, Siega-Riz AM, Savitz DA, Zhang J (2005) Maternal pre-pregnancy overweight and obesity and the risk of cesarean delivery in nulliparous women. Ann Epidemiol 15(7):467–474. https://doi.org/10.1016/j.annepidem.2005.02.005

Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw 36(3):1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer W, López-López JA (2021) Location-scale models for meta-analysis. In: Research synthesis methods

Viswesvaran C, Ones DS (1995) Theory testing: Combining psychometric meta-analysis and structural equations modeling. Pers Psychol 48:865–885

Wellek S (2002) Testing statistical hypotheses of equivalence. Chapman and Hall

Wilkinson L (1999) Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54(8):594

Williams DR, Rast P, Bürkner P-C (2018) Bayesian meta-analysis with weakly informative prior distributions. https://psyarxiv.com/7tbrm/; PsyArXiv

Wilson SJ, Polanin JR, Lipsey MW (2016) Fitting meta-analytic structural equation models with complex datasets. Res Synth Methods 7(2):121–139. https://doi.org/10.1002/jrsm.1199

Winter B, Bürkner P-C (2021) Poisson regression for linguists: a tutorial introduction to modeling count data with brms. Lang Linguist Compass. https://doi.org/10.1111/lnc3.12439

Wood SN (2017) Generalized additive models: an introduction with R. Chapman and Hall

Wooldridge JM (2012) Introductory econometrics. Cengage Learning

Yu JJ, Downes PE, Carter KM, O'Boyle EH (2016) The problem of effect size heterogeneity in meta-analytic structural equation modeling. J Appl Psychol 101(10):1457–1473. https://doi.org/10.1037/apl0000141

Yu JJ, Downes PE, Carter KM, O'Boyle E (2018) The heterogeneity problem in meta-analytic structural equation modeling (MASEM) revisited: a reply to Cheung. J Appl Psychol. https://doi.org/10.1037/apl0000328