# Practical challenges in mediation analysis: a guide for applied researchers

Megan S. Schuler[1] · Donna L. Coffman[2] · Elizabeth A. Stuart[3,4,5] · Trang Q. Nguyen[4] · Brian Vegetabile[6] · Daniel F. McCaffrey[7]

## Abstract

Mediation analysis is a statistical approach that can provide insights regarding the intermediary processes by which an intervention or exposure affects a given outcome. Mediation analyses rose to prominence, particularly in social science research, with the publication of Baron and Kenny's seminal paper and is now commonly applied in many research disciplines, including health services research. Despite the growth in popularity, applied researchers may still encounter challenges in terms of conducting mediation analyses in practice. In this paper, we provide an overview of conceptual and methodological challenges that researchers face when conducting mediation analyses. Specifically, we discuss the following key challenges: (1) Conceptually differentiating mediators from other "third variables," (2) Extending beyond the single mediator context, (3) Identifying appropriate datasets in which measurement and temporal ordering support the hypothesized mediation model, (4) Selecting mediation effects that reflect the scientific question of interest, (5) Assessing the validity of underlying assumptions of no omitted confounders, (6) Addressing measurement error regarding the mediator, and (7) Clearly reporting results from mediation analyses. We discuss each challenge and highlight ways in which the applied researcher can approach these challenges.

---

✉ Megan S. Schuler
    mschuler@rand.org

1   RAND Corporation, Arlington, VA, USA

2   Department of Psychology, University of South Carolina, Columbia, SC, USA

3   Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

4   Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

5   Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

6   RAND Corporation, Santa Monica, CA, USA

7   ETS, Princeton, NJ, USA

🙏 Springer

## 1 Introduction

Mediation analysis is a statistical approach that can provide insights regarding *how* an intervention or exposure affects a given outcome by investigating potential intermediate variables (aka "mediators") that are influenced by the exposure and, in turn, influence the outcome. Specifically, for an a priori proposed mediator, mediation analysis can evaluate to what extent the effect of the exposure on the outcome occurs through a mechanistic pathway involving the mediator, compared to all other pathways that do not involve the mediator.

Mediation analysis has primarily been applied in the fields of health and social science research to investigate potential mediating pathways in the context of both randomized interventions (e.g., a medication regimen or parenting skill-building program) or non-randomized exposures (e.g., depressive disorder or COVID-19 related school closures). Many behavioral interventions (e.g., a nutrition education program) are designed to impact the outcome of interest (e.g., cardiovascular health) by addressing an intermediate factor (e.g., diet). Mediation analysis can provide insights as to what extent the intervention impacted the outcome through the hypothesized mediating pathway(s), which can facilitate design of future interventions that are more effective and/or more parsimonious, for example by identifying the mediators that are particularly important for improving outcomes. Some agencies funding health and social science research–e.g., the US National Institute of Mental Health (NIMH)–now encourage the use of mediation analysis for investigating mechanisms of intervention effects (Cashin et al. 2019; Nguyen et al. 2020). In observational studies, mediation analyses can help to identify novel mediating pathways linking a non-randomized exposure and outcome (e.g., through what processes did the COVID-19 pandemic impact depression?), which may be targeted by subsequent intervention development.

We note that although mediation analysis has seen increased application in the field of program evaluation, it is rarely applied, to date, in the context of policy analysis (Keele et al. 2015; Ludwig et al. 2011). However, mediation analysis can similarly offer insights as to the means through which local, state, or federal policies may impact population outcomes. Having a mechanistic understanding of how certain policies work can help policy-makers to focus resources on the key elements of the policy that are driving outcomes as well as to consider alternative policy strategies if policies are found to have minimal effect on the desired intermediary pathways.

Despite the growing popularity of mediation analysis methods, applied health services and health policy researchers may still encounter challenges in terms of implementing such analyses in practice. In this paper, we review 7 key challenges, ranging from conceptualizing scientific questions regarding mediation to analytic challenges, such as how to address measurement error in mediation analyses. We hope this review provides pragmatic guidance for researchers implementing mediation analyses.

## 2 Background on mediation analysis

Figure 1 depicts the traditional mediator model, linking exposure *A,* mediator *M,* and outcome *Y.* Specifically, all relationships depicted in the mediation conceptual model—namely, the effect of *A* on *M*, the effect of *M* on *Y*, and the effect of *A* on *Y*—are
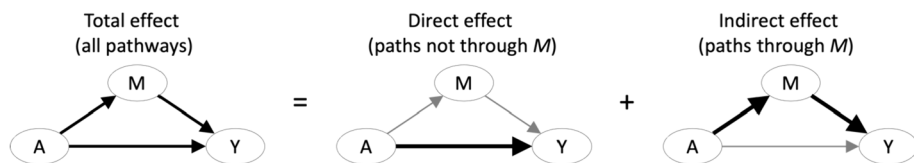
**Fig. 1** Mediation: Decomposition of the total effect into the direct effect and indirect effect

hypothesized to be causal. There are two possible causal paths from *A* to *Y*: (1) the path from *A* to *Y* that goes through the mediator *M* (referred to as the *indirect effect* of *A* on *Y*) and (2) the path from *A* to *Y* that does not go through *M* (referred to as the *direct effect)*. The indirect effect is generally of primary interest, as this represents the mediated pathway. The direct effect quantifies the relative magnitude of the effect of *A* on *Y*, through all other pathways that do not involve the mediator.

## 2.1 Traditional approach

Mediation analyses rose to prominence, particularly in social science research, with the publication of Baron and Kenny's (Baron and Kenny 1986) seminal paper, which has been cited nearly 125,000 times to date. Traditional mediation analysis required a continuous mediator and outcome and was estimated using linear regression. Under the parametric models specified in the traditional mediation framework the sum of the indirect effect and the direct effect exactly equals the *total effect* of *A* on *Y.* Baron and Kenny originally proposed a 3-step approach comprised of: (1) confirming the relationship between the exposure and outcome by regressing the outcome on the exposure, (2) confirming the relationship between the exposure and the mediator by regressing the mediator on the exposure and (3) confirming the relationship between the mediator and the outcome by regressing the outcome on both the exposure and the mediator. Conceptually, in Step (3) if mediation is present the coefficient on the mediator would be significant and the coefficient on the exposure would be attenuated relative to the coefficient estimated in Step 1. The estimate of the total effect is the exposure coefficient from the Step 1 model (outcome~exposure) and the estimate of the direct effect is the exposure coefficient from the Step 3 model (outcome~exposure+mediator). The estimate of the indirect effect is calculated as the difference between the estimates for the total effect and the direct effect. Additionally, Sobel's test can be used to test for the presence of a significant indirect effect, namely presence of significant mediation (Sobel 1982). We note that it has now been recognized that mediation can exist without Step 1 being satisfied (i.e., mediation in the absence of a significant total effect) (O'Rourke and MacKinnon 2015).

## 2.2 Modern causal inference approach

Methodological work subsequently highlighted certain key limitations of the traditional approach to mediation analysis, relating both to the analytic approach (e.g., strong reliance on linear parametric models) as well as the study design (e.g., lack of attention to confounding and temporality). A major advancement in more recent years is the incorporation of a causal inference framework and methodology (Imai et al. 2010; Pearl 2001; Robins and Greenland 1992; VanderWeele 2015). The causal mediation analysis framework

distinguishes between three distinct steps: (1) causal effect definitions, (2) causal effect identification, and (3) causal effect estimation. In this framework, causal effects are defined as the difference between two potential outcomes (rather than based on quantities from parametric models) (Holland 1986). There are several different types of effect definitions (to be introduced under Challenge 4 later in the paper), which are more general than those permitted by the traditional approach. Resultantly, mediation analyses can be conducted in settings with binary variables and allowing exposure-mediator interactions (Hong et al. 2015; Moerkerke et al. 2015; Rijnhart et al. 2021a, b; Valeri and Vanderweele 2013). Multiple recent papers have detailed the relationship between traditional mediation analysis and causal mediation analysis, including describing under which conditions and assumptions both approaches yield the same estimates (MacKinnon et al. 2020; Rijnhart et al. 2017, 2019, 2023).

In this framework, estimation of the causal effects can be conducted using either parametric or nonparametric models. Notably, causal mediation is a field of active methodological development, with many distinct estimation methods proposed for various causal mediation effects and particular recent growth in nonparametric and machine learning approaches. Broadly, estimation approaches primarily entail regression, weighting, or simulation (Nguyen et al. 2023). Estimation methods may differ not only with respect to which specific causal mediation effects they estimate, but also on factors such as bias/variance trade-offs and robustness to model misspecification. Additional considerations may include the nature of the outcome—e.g., alternative methods are needed in the context of a survival outcome measured as a time-to-event variable (Fairchild et al. 2019; Tchetgen Tchetgen 2011; VanderWeele 2015; Vansteelandt et al. 2019; Vo et al. 2022). We note that a full discussion of the specific estimation approaches in causal mediation analysis is beyond the scope of this paper—see (Nguyen et al. 2023) for a relatively accessible introduction to several common strategies used in effect estimation, as well as citations for many important papers on this topic. Finally, the causal inference framework has served to clarify the set of underlying assumptions required to estimate specific mediation effects.

We now turn to a series of challenges in conducting mediation analyses that have not generally received sufficient attention in the methodological or applied literatures.

## 3 Challenge 1: conceptually differentiating mediators from other "third variables"

Fundamentally, a mediation analysis must specify a presumed causal model that forms the basis for the proposed analyses. Indeed, mediation analysis properly falls under the umbrella of causal inference methods, as mediation analysis inherently examines causal effects (Coffman 2015; Nguyen et al. 2020), a point consistently highlighted in the mediation literature (e.g., Baron and Kenny 1986; MacKinnon 2008; Preacher 2015). That is, mediators are variables that lie in the causal pathway between exposure and outcome, and the mediator is understood to be part of a pathway that conveys (some of) the effect of the exposure to the outcome.

However, in practice, when one is constructing a causal model, it can be challenging to conceptually identify the mediating variable of interest. Generally, the exposure (treatment) and outcomes are readily identifiable, yet numerous "third variables" may be considered as potential mediators (Coffman 2015). As we review here, a given third variable may represent a mediator, or it may more appropriately be conceptualized as a moderator,

confounder, or collider. As we detail below, these 4 types of third variables—mediators, moderators, confounders, and colliders—are distinguished by their specific relationships to the exposure and the outcome (Fig. 2).

A mediator, commonly denoted *M*, is the only type of third variable that lies in the causal path between the exposure (denoted *A*) and outcome (denoted *Y*). Essentially, the relationships between *A, M,* and *Y* (i.e., the effect of *A* on *M*, the effect of *M* on *Y*, and the effect of *A* on *Y*)—are hypothesized to be causal. As an illustrative example, one mediator in the path between socioeconomic status (SES) and cardiovascular disease may be diet. An individual's diet may be directly influenced by their SES and may in turn impact their cardiovascular health.

A moderator (commonly denoted *Z*), shown in Panel II of Fig. 2, represents another type of third variable. A moderator is a variable that alters the effect of the exposure on the outcome, resulting in effect heterogeneity. Note that a moderator is not directly on the causal path between *A* and *Y*, but rather the level of the moderator variable determines the magnitude of the *A-Y* effect. That is, moderation questions seek to determine *for whom* and *under what conditions* a given exposure affects the outcome (Hong 2015). For example, if a certain type of medication has a larger effect on blood pressure for women compared to men, then sex is a moderator of the medication–blood pressure effect. Moderating variables may include individual characteristics (e.g., sex may impact the pharmacological effects of a medication), contextual characteristics (e.g., a clinician's level of training may impact how effectively they provide psychotherapy), and prior or concurrent treatments (e.g., whether a patient previously received a specific type of chemotherapy may impact how effective current radiation therapy is). While some have asserted that moderators must temporally precede the exposure whose effect they moderate, e.g., (Kraemer et al. 2008), Hong clarifies that moderators neither need to occur prior to the exposure nor be independent of the exposure (see (Hong 2015) for a detailed discussion of moderated treatment effects).
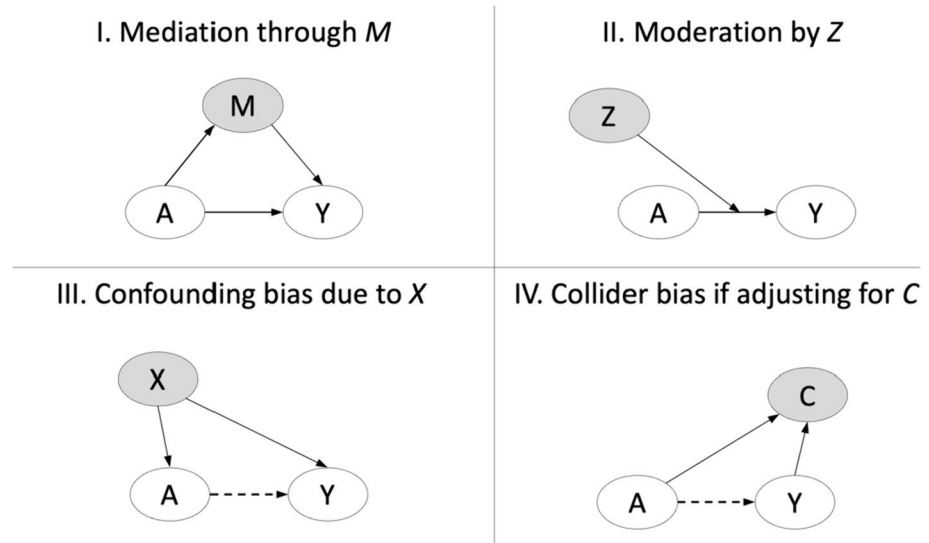


**Fig. 2** Distinguishing third variables: mediators, moderators, confounders, and colliders

An additional type of third variable is a confounder (commonly denoted *X*), as depicted in Panel III of Fig. 2. A confounder represents a common cause of both *A* and *Y*. For example, in the context of investigating the effect of Alzheimer's disease on mortality, age would represent a confounder, as age is causally associated with both Alzheimer's disease and mortality. Comparing confounders and mediators, we note that a confounder both precedes and causes *A* whereas a mediator is caused by *A* and thus occurs subsequent to *A*. Note that the effect of *A* on *Y* is shown as a dashed line to highlight that a confounder can distort the true effect of *A* on *Y*—indeed, the presence of a confounder may induce the appearance of an effect between *A* and *Y* when none truly exists. To avoid confounding bias, confounding must be addressed on the design side via randomization or on the analytic side via statistical adjustment or stratification (Lipsky and Greenland 2022).

A final type of third variable is a collider (denoted *C*) as shown in Panel IV of Fig. 2. A collider variable is one that has two variables causally leading into it. We note that while the figure depicts *A* and *Y* as these 2 causal variables, definitionally it does not need to be *A* and *Y* (e.g., it could be *A* and *M*). In some sense a collider is the opposite of a confounder, as a collider is a common outcome of *A* and *Y* as opposed to a common cause. For example, consider an individual's socioeconomic status during childhood as the exposure of interest and their high school GPA as the outcome of interest. Both of these factors likely influence one's educational status at age 25, which would represent a collider in this context. Collider bias occurs when an observational study either disproportionately samples individuals based on the collider variable (e.g., only includes individuals with at least a college degree) or stratifies on the collider variable in the analysis—in both scenarios, the observed effect of *A* on *Y* will represent a distortion of the true effect of *A* on *Y*. To avoid collider bias, sampling or stratification based on the collider should be avoided (Holmberg and Andersen 2022) (Table 1).

It is imperative to remember that statistical analyses generally cannot empirically differentiate whether a given third variable is "truly" one of the above four types (McGowan et al. 2023). Rather, the applied researcher must a priori specify a third variable's role (based on their contextual knowledge) and then subsequently determine the most appropriate way to analytically model the variable. An additional point of confusion is that the same "variable" could represent either a moderator or a mediator in different contexts or measured at different points in time (Coffman 2015). For example, consider the relationship between depressive symptoms and suicidality. Evidence suggests that social support acts as a moderator, in that higher levels of social support can attenuate the relationship between depressive symptoms and suicidality (Rubio et al. 2020). Social support is not on the hypothesized causal path between depressive symptoms and suicidality, and so is not a mediator in this context. In contrast, consider a community-building intervention for older adults that seeks to reduce loneliness by improving social support. In this context, level of social support is a hypothesized mediator in the causal path linking the intervention and loneliness. Thus, a given factor—e.g., social support—is not inherently a mediator or moderator; rather, the hypothesized causal relationships determine the role of a specific variable in a given analytic context. Furthermore, in the context of mediation, statistical analyses cannot provide direct evidence of causality. Analytic results provide insights regarding observed associations, given the hypothesized conceptual model. As highlighted by Holland (1986), the challenges of causal inference lie in inferring evidence of causality from these observed associations, which generally reflect a mix of causal effects and various non-causal components, and in assessing the validity of the assumptions required to interpret relationships as causal.

**Table 1** Summary of different types of "third variables"

| | Definition | Illustrative example |
|---|---|---|
| Mediator | A mediator is the only type of third variable that lies in the causal path between the exposure and outcome | One mediator in the path between socioeconomic status (SES) and cardiovascular disease may be diet. SES may directly impact one's diet, which may in turn impact their cardiovascular health |
| Moderator | A moderator is a variable that alters the effect of the exposure on the outcome, resulting in effect heterogeneity | If a certain type of medication has a larger effect on blood pressure for women compared to men, then sex is a moderator of the medication's effect |
| Confounder | A confounder represents a common cause of both the exposure and outcome | If age is causally associated with both Alzheimer's disease and mortality, then age would confound the relationship between Alzheimer's disease and mortality |
| Collider | A collider variable is one that has two variables causally leading into it (i.e., it is a common outcome of two variables) | If both an individual's childhood socioeconomic status and their high school GPA influence one's educational status at age 25, then age 25 education status represent a collider with respect to childhood SES and high school GPA |

*3 different indirect effects through $M_1$ and/or $M_2$*
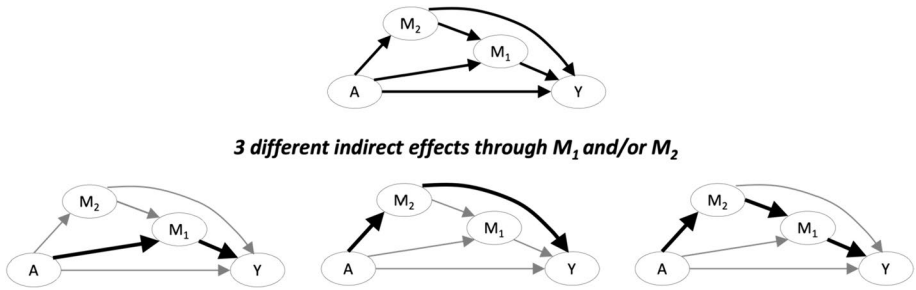


**Fig. 3** Conceptual model with 2 causally ordered mediators, $M_1$ and $M_2$

## 4 Challenge 2: extending beyond the single mediator context

While mediation is often discussed with respect to a single mediator, in practice there are often multiple potential mediators.

*Illustrative example:* In an intervention context, consider a parenting intervention program that seeks to reduce substance use among parents of young children. The program specifically targets three constructs that may drive substance use behaviors: depression, parenting stress, and household poverty (Massarwi et al. 2021). When evaluating effectiveness of this intervention, examining the mediating roles of all three constructs may be of interest. The following figure depicts the relationship between the intervention program, the 3 mediators, and the outcome (assuming independence of the 3 mediators).
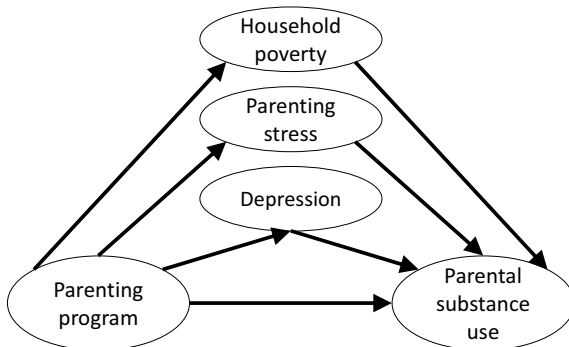


To date, the mediation literature has largely focused on applications with a single mediator, or the straightforward extension in which a set of mediators are treated as single *en bloc* composite variable, allowing application of traditional single mediator methods (Daniel et al. 2015). When multiple mediators are present, a host of conceptual and analytic questions arise: Are the mediators causally ordered, such that earlier mediators may affect later mediators? If the mediators are not causally ordered, are they assumed to be independent or allowed to be correlated? A growing literature on multiple mediator analysis has emerged, highlighting the complexities (and required assumptions) of identifying path-specific effects through multiple mediators. As shown in Fig. 3, in the setting in which there are 2 causally ordered mediators, $M_1$ and $M_2$, there are 3 different indirect effects—one strictly through $M_1$, one strictly through $M_2$, and one through both. For references on the setting in which mediators causally affect each other see: (Daniel et al. 2015; Gao and
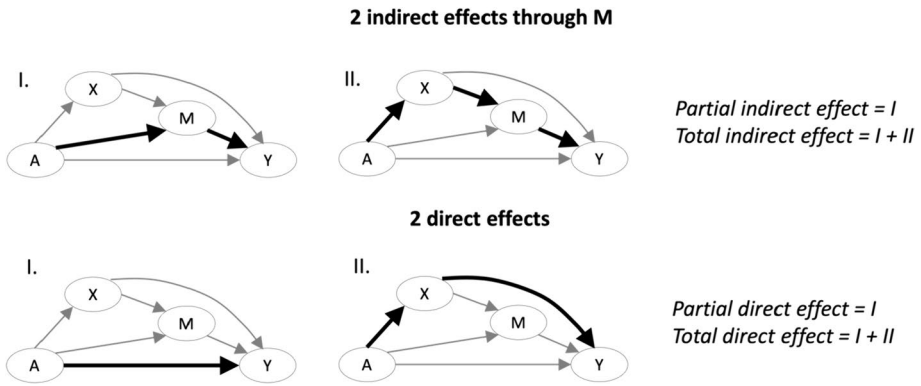
**Fig. 4** Illustration of the additional direct and indirect paths between the exposure *A* and outcome *Y* in the presence of a post-treatment confounder *X*

Albert 2019; Imai and Yamamoto 2013; Steen et al. 2017; Tai et al. 2021); for the setting in which multiple mediators do not causally affect each other see: (Jerolon et al. 2020; Lange et al. 2013; MacKinnon 2015; Preacher and Hayes 2008; Taguri et al. 2018).

Alternatively, it may be the case that an additional intermediary variable—here denoted as *X*—is not of substantive interest as a mediator, but rather represents a potential post-treatment confounder (Coffman 2015). Specifically, *X* occurs after the exposure *A* (and is influenced by *A*) and precedes *M,* functioning as a confounder of the mediator and outcome (see Fig. 4). Notably, *X* is on the causal path between *A* and *Y*, and thus is a candidate mediator. As shown in Fig. 4, the presence of *X* results in two indirect effects (both through and independent of *X*) as well as two direct effects (both through and independent of *X*). In this context, we can define additional effects of interest, namely the *partial indirect effect, total indirect effect, partial direct effect,* and *total direct effect*—each of which may be of interest depending on the scientific question under investigation. The *partial indirect effect* consists only of the pathway that goes through the mediator but not through the post-treatment confounder, whereas the *total indirect effect* consists of all pathways that go through *M*. Similarly, the *partial direct effect* consists only of the path between the exposure and the outcome that does not involve either the mediator or the post-treatment confounder, whereas the *total direct effect* consists of all pathways that do not go through *M*. It is essential to appropriately account for post-treatment confounding, as treating a post-treatment confounder like a baseline confounder (e.g., using regression adjustment) will result in biased estimates of the direct effect. See (Coffman and Zhong 2012; De Stavola et al. 2015; Hong et al. 2022; Miles et al. 2020; Moerkerke et al. 2015; Valente et al. 2017) for more discussion.

Another extension of the traditional single mediator model is the time-varying context, in which any of the exposure, mediator, outcome, and third variables may be dynamically changing across time. In this complex and challenging context, previous assessments of the time-varying exposures, mediators and outcomes act as post-treatment confounders or additional mediators. Multiple methods have been proposed for mediation analysis in this context, using approaches such as latent growth models, multilevel structural equation models, autoregressive models, and sequential mediation models. See (Gunzler et al. 2014), (VanderWeele and Tchetgen Tchetgen 2017), (Lin et al. 2017a, b), (Berli et al.

2021), and (Cai et al. 2022) for more details. This area remains an active field of methodological development.

---

*Illustrative example*: A pharmacological medication for smoking cessation may impact smoking outcomes by reducing cravings for cigarettes. However, the degree of cravings that a person experiences may decrease as the time since quitting smoking increases. Hence, craving may act as a time-varying mediator. See (Cai et al. 2022) for further examples.

---

## 5 Challenge 3: identifying appropriate datasets in which measurement and temporal ordering supports the hypothesized mediation model

As originally included in Hill's criteria, temporality is a core aspect of causality (Hill 1965), in that the treatment or exposure of interest must come before the outcome in order to be deemed a causal factor – i.e., "cause precedes effect." In the mediation context, this temporal ordering requirement applies to multiple paths, as the treatment should temporally precede the mediator and the mediator should precede the outcome. For a mediation analysis to be credible, the temporal ordering of the treatment, mediator, and outcome variables must support the temporal ordering of the proposed mediation model. When planning longitudinal data collection for mediation analyses it is crucial for researchers to consider optimal measurement timing, accounting for the hypothesized timing of the effects of interest. In the context of secondary data analysis, in some cases, it may be possible to conduct mediation analyses post hoc if the measurement timing reflects the temporal ordering of the hypothesized model. In settings where repeated waves of data were collected, it may be possible to specify a mediation model for which the baseline covariates were measured before the exposure, the exposure was measured before the mediator(s) of interest, and the mediator(s) were measured before the outcomes of interest. However, sometimes the nature of available data simply may not support analysis of the proposed mediation model. If the way the exposure, potential mediator, and outcome variables were measured does not clearly establish temporality—as is often the case in cross-sectional data—it is impossible to analytically identify temporal ordering. We also highlight that establishing temporality of covariates—relative to the exposure, mediator, and outcome—is crucial, as this differentiates baseline (pre-treatment) confounders from post-treatment confounders.

Conducting mediation analysis using cross-sectional data in which temporal ordering is not well-defined has been shown to often result in biased, misleading results (Lindenberger et al. 2011; Maxwell and Cole 2007; Maxwell et al. 2011; O'Laughlin et al. 2018; Selig and Preacher 2009). Fundamentally, one of the key assumptions regarding a mediational pathway is the temporal ordering, namely the passage of time that elapses between the exposure, mediator, and outcome. As such, a mediation analysis should examine variables that are measured sequentially, rather than concurrently (O'Laughlin et al. 2018). Generally, use of cross-sectional measures of the exposure, mediator, and outcome (all assessed at the same time) would implausibly suggest that effects are instantaneous (Gollob and Reichardt 1987). We note that, in some cases, the temporal ordering can be assumed with cross-sectional data—e.g., when using retrospective questions that refer to different time periods. However, the validity of the temporal ordering may still be questionable because of factors such as recall bias. A second key limitation of using cross-sectional data is that it prevents controlling for baseline measures of the mediator(s) and outcome(s), which has

been shown to be important for reducing bias when estimating mediational effects (Selig and Preacher 2009).

Finally, we note that temporal ordering becomes even more important—and more complex—when considering time-varying exposures, mediation processes, and/or outcomes. In such a case it is important to carefully consider—and appropriately account for—the relative temporal ordering of variables. Recent methodological developments in this area have allowed for time-varying exposures and mediators (Lin et al. 2017a, b; VanderWeele and Tchetgen Tchetgen 2017), time-varying mediators and outcomes (Bind et al. 2016; Zeng et al. 2021), and time-varying mediation effects with time-varying mediators and outcomes (Cai et al. 2022; Chakraborti et al. 2022). Notably, the latter two references allow the mediated effect itself to vary as a function of time in addition to allowing values of the mediator and/or outcome to vary over time. A central challenge regarding estimation of mediational effects in the presence of time-varying exposures and mediators is time-varying confounding. Specifically, a time-varying exposure serves as a "post-treatment" confounder regarding the mediator-outcome effect (similar to Fig. 4 but replacing $A$ and $X$ with $A_t$ and $A_{t+1}$, respectively). As we discuss in the next section, not all mediational effects are able to be estimated in this context (i.e., see section on *interventional effects*).

# 6 Challenge 4: selecting mediation effects that reflect the scientific question of interest

The incorporation of a causal inference perspective in mediation analysis reveals that causal effects can be defined in different ways depending on which potential outcomes are contrasted. The well-known effect types include natural (in)direct effects (Pearl 2001; Robins and Greenland 1992), controlled direct effects, and interventional (in)direct effects (Lok 2016; Lok and Bosch 2021; Vanderweele et al. 2014; Vansteelandt and Daniel 2017). Also, other effects can also be defined flexibly depending on the specific scientific question (Nguyen et al. 2020). The discussion of these various effects in the literature generally focuses on the issue of identification and estimation, offering little guidance to applied researchers regarding how to select which type of mediation effect is most appropriate for the research question at hand. In this section, we give a brief introduction to these effects, describing the motivation for using each effect type.

## 6.1 Natural direct and indirect effects

Natural direct and indirect effects are motivated by the desire to explain the total causal effect. They are defined to decompose the total effect, i.e., the combination of the natural direct effect and the natural indirect effect is equal to the total effect. Natural effects are popular because they speak to the original motivation for mediation analysis, namely, to explain the total causal effect through effect decomposition.

These effects are defined based on potential outcomes. Consider the case in which $A$ is a binary exposure variable, indicating the exposed ($A = 1$) or the comparison ($A = 0$) condition. Each individual has a potential outcome had they received the exposure (denoted $Y_1$) <u>and</u> a potential outcome had they received the comparison condition (denoted $Y_0$). While both potential outcomes exist for all individuals, we can observe only one of the two. Formally, $Y = Y_A$, i.e., the observed outcome reveals the potential outcome corresponding to

the individual's actual exposure condition; this is the *consistency* assumption. For an individual, the effect of the exposure on the outcome (which in mediation analysis is called the total effect) is taken to be the difference between these two potential outcomes, $= Y_1 - Y_0$. Generally, one takes the average of this difference across individuals to obtain a population average effect.

The definition of natural direct and indirect effects relies on a "nested" type of potential outcomes that involves the mediator $M$. Note that the mediator is an intermediate outcome of the exposure and itself has potential values, $M_1$ and $M_0$. Let $a$ and $a\prime$ be two indices that could be either 1 or 0. Let $Y_{aM_{a\prime}}$ denote the potential outcome for the case when the exposure is equal to $a$ and the mediator is equal to $M_{a\prime}$ (the potential value of the mediator under condition $a\prime$). Crossing both exposure conditions with both potential mediator values yields four nested potential outcomes: $Y_{1M_1}$, $Y_{0M_0}$, $Y_{1M_0}$ and $Y_{0M_1}$. The latter two of these correspond to *cross-world* conditions, namely hypothetical (i.e., unobservable) conditions in which $a \neq a\prime$. Natural direct and indirect effects are contrasts of these potential outcomes. There are two natural indirect effects, $NIE_1 = Y_{1M_1} - Y_{1M_0}$ and $NIE_0 = Y_{0M_1} - Y_{0M_0}$, each contrasting a change of mediator from $M_0$ to $M_1$ while holding the exposure constant. There are two natural direct effects $NDE_0 = Y_{1M_0} - Y_{0M_0}$ and $NDE_1 = Y_{1M_1} - Y_{0M_1}$, each contrasting a change of exposure from 0 to 1 while holding the mediator constant. These effects form two pairs, the $[NDE_0, NIE_1]$ pair and the $[NDE_1, NIE_0]$ pair, each of which sum to the overall total effect, $TE = Y_{1M_1} - Y_{0M_0}$.

We highlight that, when reporting natural effects, the researcher must select which decomposition(s) of the total effect to report—a choice that has generally received little attention in the literature. Notably, these two decompositions become equivalent if there is no exposure-by-mediator interaction. Methodological papers tend to either present both decompositions or to present only the decomposition into $NDE_0$ and $NIE_1$ (without discussing the motivation for this choice). Recent work by Nguyen et. al. (2020) provides suggestions on which natural effect decomposition to select. They propose using the $[NDE_0, NIE_1]$ pair if the research question is "Is there a mediated effect?" or "Is the causal effect (partly) mediated by the proposed mediator?" Alternatively, they propose using the $[NDE_1, NIE_0]$ pair if the question is "In addition to the mediated effect, is there a direct effect?" or "Does the exposure influence the outcome in other ways, not through this mediator?" Finally, they recommend presenting both decompositions if there is no prior assumption or preferred question regarding direct or indirect effects. We note that there are alternatives to the two decompositions discussed above that may be of interest in particular settings; see (Hong et al. 2015; Nguyen et al. 2022; VanderWeele 2015) for more details.

Additionally, we highlight that these are conceptual, non-parametric definitions for the NDE and NIE that do not require a specific estimation approach for modeling $M$ and $Y$. Thus, after selecting the NDE and NIE as the mediation effects of interest, the applied researcher needs to choose a specific estimation approach (e.g., inverse probability weighting, g-computation). Analytic considerations regarding the estimation approach are beyond the scope of this paper—please see (Nguyen et al. 2023) for a detailed discussion.

In light of the call for "pragmatic epidemiological research" (i.e., focusing on real-life exposures or interventions applied in routine (rather than optimal) community settings), estimands that rely on hypothetical "cross-world" counterfactuals may seem impractical. However, these cross-world counterfactuals are a means to an end to obtain estimates for the NDE and NIE, which can indeed offer meaningful and pragmatic insights. In particular, mediation analyses often focus on identifying potential (actionable) mediators in

health-related processes, with the intention of developing / refining interventions that target mediators.

## 6.2 Interventional effects

Interventional effects refer to a broad class of causal contrasts that involve hypothetical conditions where the exposure and/or mediator is intervened on and set to either a specific value or a specific distribution. Most of these effects are defined in the aggregate rather than individual levels. Since these effects are about hypothetical situations (including hypothetical modifications of exposure or hypothetical interventions on the mediator), they should be interpreted as "prescriptive" effects. A key difference between interventional effects and the natural (in)direct effects above is that, in principle, one could conduct an experiment contrasting the interventional effect conditions of interest, whereas natural (in) direct effects are not experimentally testable (Robins 2003). Additionally, interventional effects do not decompose the total effect on the outcome while natural (in)direct effects do.

### 6.2.1 Controlled direct effects

The simplest effects in the class of interventional effects are controlled direct effects, which contrast conditions where both exposure and mediator are set to specific, known, values. A controlled direct effect (for mediator level $m$) is the causal effect of the exposure on the outcome if the mediator were held constant at level $m$ for the entire population (across both treatment conditions) (Robins and Greenland 1992). That is, $\text{CDE}(m) = \text{E}[Y_{1m}] - \text{E}[Y_{0m}]$, where $\text{E}[\bullet]$ is notation for the population mean. (A controlled direct effect can also be defined at the individual level as $Y_{1m} - Y_{0m}$.) While this contrast compares conditions with and without the intervention, it is defined in a context where regardless of intervention condition, the mediator is fixed to a specific value. Controlled direct effects are thus only of interest in very specific situations where an anticipated/imagined external manipulation is believed to have this mediator-fixing impact.

---

*Illustrative example:* Nguyen et al. (2020) give the example of a childhood burn prevention campaign, which sought to raise parental awareness of common causes of childhood burns, including excessively high settings on household water heaters. It was determined that one mechanism of the intervention's impact on the rate of childhood burns (the outcome) was through reductions in water heater temperatures (the mediator). A new city ordinance will soon require that household water heaters must not be set higher than 120°F (the $m$ value). City officials want to know how much impact continuing the burn prevention campaign will have, considering the new ordinance that fixes the mediator level – this effect can be estimated by the controlled direct effect CDE($m$).

---

We note that a more generalized version, referred to as a *generalized direct effect,* holds the mediator constant at a given distribution, allowing some variation in mediator level across individuals, rather than a fixed level $m$. Also, we highlight that there is no corresponding notion of a "controlled indirect effect." This is not surprising because we can define controlled direct effects without conceptualizing mediation at all; a controlled direct effect is simply the effect of the exposure in a specific setting (where variable $M$ is controlled). Thus, estimation of controlled direct effects will only provide insight regarding direct effects, not indirect effects.

### 6.2.2 Interventional (in)direct effects

Historically, one of the motivations for defining interventional (in)direct effects is that these effects—unlike natural (in)direct effects—are identified in the presence of mediator-outcome confounders that are influenced by exposure (i.e., post-treatment confounders). The *interventional direct effect* (also called *randomized, stochastic, or organic direct effect*) is a specific type of generalized direct effect in which the mediator distribution is held constant to the distribution of potential mediator value $M_1$ (or $M_0$), given a set of pre-exposure covariates $C$ (Didelez et al. 2006; Lok 2016; Vanderweele et al. 2014). The key difference between interventional direct effects and natural direct effects is that the latter involve setting the mediator to individual-specific potential mediator values $M_1$ (or $M_0$), whereas interventional effects shift the mediator distribution to be equal to a conditional distribution of the potential mediator. Also, interpretation of natural effects does not invoke the notion of a hypothetical intervention; rather, natural effects seek to decompose the observed total effect.

In order to formally define interventional (in)direct effects, we need additional notation: let $\mathcal{M}_{a|C}$ (with a script letter $\mathcal{M}$) denote the distribution of potential mediator $M_a$ given $C$ and let $Y_{a\mathcal{M}_{a'|C}}$ denote the outcome that would arise if the exposure were set to $a$ and the mediator distribution was intervened on and set to the distribution of $M_{a'}$ given $C$. The two interventional direct effects are defined as: $IDE_0 = \mathrm{E}[Y_{1\mathcal{M}_{0|C}}] - \mathrm{E}[Y_{0\mathcal{M}_{0|C}}]$ and $IDE_1 = \mathrm{E}[Y_{1\mathcal{M}_{1|C}}] - \mathrm{E}[Y_{0\mathcal{M}_{1|C}}]$. Analogously there are two interventional indirect effects: $IIE_0 = \mathrm{E}[Y_{0\mathcal{M}_{1|C}}] - \mathrm{E}[Y_{0\mathcal{M}_{0|C}}]$ and $IIE_1 = \mathrm{E}[Y_{1\mathcal{M}_{1|C}}] - \mathrm{E}[Y_{1\mathcal{M}_{0|C}}]$, each of which contrasts shifting the mediator distributions from $\mathcal{M}_{0|C}$ to $\mathcal{M}_{1|C}$, while holding the exposure fixed. Intuitively, shifting the mediator distribution is equivalent to assigning each individual a random draw from that specific distribution (Nguyen et al. 2020). Interventional (in)direct effects do not decompose the total effect, as they are not designed to do so.

In practice, researchers may choose to estimate interventional effects in settings where natural effects are not identified (i.e., the presence of post-treatment confounders) – in this sense, interventional effects might serve as an approximation for the unidentified natural (in)direct effects. However, the motivation of an analysis estimating interventional effects is generally different than that of an analysis estimating natural effects. As detailed by Nguyen et al. (2020), interventional effects seek to answer "what if" questions about what magnitude of change could be expected at the population level if a hypothetical intervention was implemented.

---

*Illustrative example:* Consider a high school that currently has a substance use prevention program that comprised several different components, including self-efficacy building, physical exercise, and mindfulness. Another school in the district is considering implementing this program, yet due to budget constraints they would drop the mindfulness component (here, our potential mediator M). An estimate of the level of substance use-related incidents (here, Y) this school could anticipate if they implemented this modified program (here, A = 1) would be provided by $IDE_0$. Here, $IDE_0$ contrasts the effect of the proposed modified program which would not impact the mindfulness distribution ($\mathrm{E}[Y_{1\mathcal{M}_{0|C}}]$) to a hypothetical control condition (e.g., no prevention program), assuming a similar lack of impact on student mindfulness ($\mathrm{E}[Y_{0\mathcal{M}_{0|C}}]$).
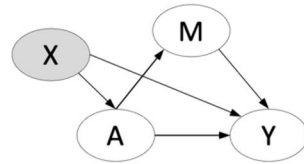
---

### 6.2.3 Defining interventional effects more flexibly

Rather than defaulting to interventional (in)direct effects, Nguyen et al. (2022) advocates for defining interventional effects flexibly to best match the scientific research question.
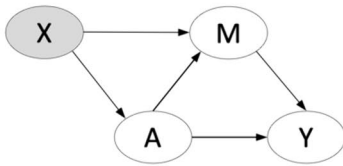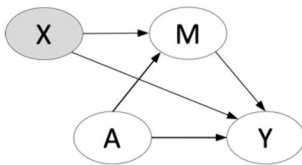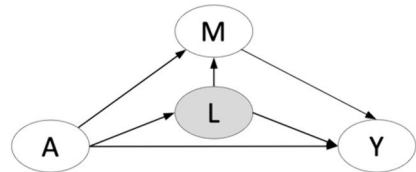
**Fig. 5** The four types of confounding that are assumed *not* to exist for identification of natural direct and indirect effects

This requires being thoughtful about (a) which active intervention condition is of interest, and given this choice, (b) which condition is the appropriate comparison condition. In the prior example of modifying the substance use prevention program, a more relevant comparison condition may be the true status quo $E[Y_0]$, without imposing the distributional assumption about mindfulness $E[Y_{0 \mathcal{M}_{0|c}}]$. As such, a more meaningful interventional effect may be given by $E[Y_{1 \mathcal{M}_{0|c}}] - E[Y_0]$. This more general framework for interventional effects provides flexibility to define causal effects that better accommodate a much broader range of research questions. For an example of how flexibly defining effects readily applies to research on health/social disparities, see (Nguyen et al. 2020). As such research involves imagining alternative worlds where social and structural elements that contribute to disparities were mitigated or neutralized, there is no standard pair (or set) of contrasts. Rather, each imagined alternative world calls for an appropriate contrast that can be accommodated within this flexible framework.

## 7 Challenge 5: assessing the validity of underlying assumptions required for estimation

This section will discuss the necessary assumptions for identification (i.e., identifying the potential outcomes, which are unobserved, from the observed data) for each of the mediational effects defined in the previous section—natural effects, interventional effects, and controlled direct effects. Importantly, all of these effects require several assumptions regarding confounding—referred to as "ignorability assumptions" or "sequential randomization assumption"—in order to be interpreted as causal. We note that our focus is on the single mediator case—the context of multiple mediators may require additional assumptions (which may vary depending on the proposed interrelationship between the mediators).

Natural effects require the most stringent assumptions. Specifically, to identify natural direct and indirect effects, four assumptions regarding confounding are needed: no unobserved confounding of the (1) exposure-mediator relationship, (2) exposure-outcome relationship, or (3) mediator-outcome relationship, as well as (4) no observed or unobserved confounders of the mediator-outcome relationship that have been influenced by the exposure (also referred to as "post-treatment confounders" or "intermediate confounders" and commonly denoted by $L$). These confounding assumptions are depicted in Fig. 5.

We note that certain modern estimation methods for the NIE/NDE may relax some of these assumptions in some contexts, and that certain types of potential confounding may be able to be addressed analytically through techniques such as regression adjustment or propensity score weighting (see (Nguyen et al. 2022) for more details). If there is random assignment to the intervention or exposure variable then the assumptions of no unmeasured confounders of the exposure-mediator and exposure-outcome relationships should hold; however, the other two assumptions regarding confounding of the mediator-outcome relationship are not guaranteed to hold even when there is random assignment to the exposure variable. The no intermediate confounders assumption is particularly stringent. It means there are no time-varying (i.e., post-treatment) confounders and that any additional mediators must be conditionally independent of the mediator of interest. Yet this assumption has been traditionally required to identify mediation effects, starting with Baron and Kenny (1986), Robins and Greenland (1992), and Pearl (2001). Subsequent work identified approaches and required assumptions to assess mediation in the presence of an exposure-mediator interaction, including (Preacher et al. 2007), (Vansteelandt and Vanderweele 2012), (Coffman and Zhong 2012), (Valeri and Vanderweele 2013), and (Tchetgen Tchetgen and Vanderweele 2014). The development of interventional effects by (Vanderweele et al. 2014) was motivated by a desire to identify causal mediation effects that do not require the assumption of no intermediate confounding; this approach was later extended to allow for multiple mediators by (Vansteelandt and Daniel 2017).

Interventional effects require fewer assumptions to identify than the natural effects; yet they do not decompose the total effect like the natural direct and indirect effects do. Identification of the interventional direct and indirect effects requires the assumption of no unobserved confounding of the (1) exposure-mediator relationship, (2) exposure-outcome relationship, or (3) mediator-outcome relationship. If, additionally, there are no observed or unobserved confounders of the mediator-outcome relationship that have been influenced by the exposure, then the interventional direct and indirect effects are equal to the natural direct and indirect effects (VanderWeele and Tchetgen Tchetgen 2017).

Identification of the controlled direct effect requires the fewest confounding assumptions, namely that there are no unmeasured confounders of the (1) exposure-outcome relationship and (2) mediator-outcome relationship. If there is random assignment to the exposure, then the first confounding assumption should be satisfied, although it is still a concern in observational studies. As randomization to the exposure does not imply randomization to levels of the mediator, the second assumption will be relevant in both randomized and observational studies. No assumptions about confounders of the exposure-mediator relationship or intermediate confounders are necessary. Although it is easier to identify the controlled direct effect, it unfortunately does not have a corresponding indirect effect as mentioned in the previous section.

Additionally, identification of natural effects, interventional effects, and controlled direct effects all also require the assumption of positivity, that is, that individuals have some positive probability of receiving each level of the treatment and each level of the mediator. See

**Table 2** Summary of required assumptions regarding unobserved confounding and positivity required for estimation of natural effects, interventional effects, and controlled direct effects

| | No *unobserved* confounders of the: | | | No *observed or unobserved* confounders of the: mediator-outcome relationship that have been influenced by the exposure | Positivity |
|---|---|---|---|---|---|
| | Exposure-mediator relationship | Exposure-outcome relationship | Mediator-outcome relationship | | |
| Natural effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Interventional effects | ✓ | ✓ | ✓ | (if true, equivalent to natural effects) | ✓ |
| Controlled direct effect | | ✓ | ✓ | | ✓ |

(Nguyen et al. 2022) for further discussion of the positivity assumption for each effect type (Table 2).

Notably, the previously described assumptions regarding no unmeasured confounding cannot be empirically assessed, as they fundamentally relate to unobserved variables. Rather, sensitivity analyses have been developed that assess the potential bias that would arise from different types of unmeasured confounding. These sensitivity analyses can be used to determine the magnitude of unobserved confounding that would qualitatively change the results of a given mediation analysis. The specific choice of sensitivity analysis depends on which mediation effects are being estimated. For the natural direct and indirect effects, (Imai et al. 2010) proposed a sensitivity analysis to the no-unmeasured-confounding assumptions, which was then extended for the presence of a post-treatment confounder (Imai and Yamamoto 2013) and implemented in the *mediation* R package. VanderWeele proposed a sensitivity analysis for the no-unmeasured-confounding assumptions used in identifying controlled direct effects (VanderWeele 2010) as well as sensitivity analyses for natural direct and indirect effects regarding mediator-outcome confounders affected by the exposure (i.e., post-treatment confounders; (VanderWeele and Chiba 2014). VanderWeele's sensitivity analyses are implemented in the *CMAverse* R package (Shi et al. 2021). Additional work has proposed sensitivity analyses for natural effects to simultaneously assess unmeasured confounding in the mediator-outcome, exposure-outcome, and exposure-mediator relationships using a Bayesian approach (McCandless and Somers 2019). Hong et. al. introduced a weighting-based sensitivity analysis (an extension of the ratio-of-mediator-probability weighting (RMPW) method) to assess potential bias arising from omitted pre-treatment or post-treatment covariates; this method has minimal functional form assumptions and accommodates a broad range measurement scales for the mediator, outcome, and omitted covariates (Hong et al. 2018).

More recently, more comprehensive approaches to sensitivity tests have been proposed, referred to as "multiverse approach" or a "multi-model analysis" (Steegen et al. 2016; Young and Holsteen 2017). These approaches seek to quantify how robust findings are to *all* decisions and specifications made throughout the data analytic processes (Steegen et al. 2016). Recent work has detailed how to take a multiverse approach in the context of mediation analysis (Rijnhart et al. 2022). Specifically, the authors highlight that mediation analysis, by virtue of the addition of an intermediate mediation variable, entail more analytic decisions regarding operationalization of variables and model specification compared to bivariate (i.e., non-mediation) models that simply regress an outcome on an exposure. As such, the multiverse examined in this type of sensitivity analysis will be larger and more complex than multiverses corresponding to bivariate models (Rijnhart et al. 2022). Multiverse methods can generate specification curves that quantify the impact of analytical decisions on the magnitude and statistical significance of the direct, indirect, and total effect estimates (Rijnhart et al. 2022). Overall, we note that most of the sensitivity analyses described above are for the context of a single mediator – to date, there are few sensitivity analysis methods for the context of multiple mediators or time-varying mediation, although this is an area of ongoing methodological work (Daniel et al. 2015; Wickramarachchi et al. 2023).

In addition to identification assumptions specific to mediation, mediation analyses are also subject to the standard assumptions related to the specific statistical modeling approach used. Unlike identification assumptions, it may be possible to empirically assess the plausibility of modeling assumptions (e.g., assess concerns regarding small-sample bias of logistic regression by examining prevalence of the outcome in the data). We highlight that identification and modeling assumptions can become intertwined in the context

of mediation analysis. For example, one way around the assumption of no post-treatment confounders of the mediator and outcome is to assume that there are no interactions between the treatment and the mediator; this parametric modeling assumption, when combined with an assumption of linear relationships among all the variables, serves to identify the direct and indirect effects in the traditional mediation model. Additionally, if there is no exposure-mediator interaction, the controlled direct effect of the treatment will be equal to the natural direct effect. The analyst may then employ inverse-probability of treatment weighting to adjust for post-treatment confounders (Austin and Stuart 2015; Coffman and Zhong 2012).

## 8 Challenge 6: addressing measurement error regarding the mediator

While potential bias arising from violations of the assumptions related to confounding (discussed in the prior section) has received much attention in the mediation literature, potential bias arising due to measurement error has received much less discussion to date. To ensure plausibility of the underlying consistency assumption in mediation analysis, it is imperative that the variables used in the analysis – i.e., treatment variable, mediator(s), outcome(s), and other covariates – are measured without error. In many fields of study, however, mediators are likely measured with error. For example, psychological variables are not observed directly but measured imprecisely through multiple indicators. Educational tests provide indirect estimates of the latent abilities of interest, as do survey measures of attitudes and beliefs. Self-reports of substance use and other behaviors are also known to be measured with error; for examples see: (Cole and Preacher 2014; Del Boca and Noll 2000; Johnson and Fendrich 2005; Morral et al. 2000; Natarajan et al. 2010).

---

*Illustrative example:* Consider again the high school substance use prevention program comprised of self-efficacy building, physical exercise, and mindfulness components. As part of the program evaluation, students completed surveys with scales measuring their mindfulness, measured with multiple items or indicator variables. The scores on the mindfulness scales contain measurement error and using them in mediation analysis would result in bias. For example, if the reliability of the scale score is 0.80 (often considered an acceptable level for reliability) and the models for the mediator and outcome are linear, then using the formula for the bias in Valeri et al. (2014) the NIE would be biased toward zero by about 20 percent, i.e., a positive NIE would be 20% too small and a negative NIE would be too large by 20%. Such bias could lead to inaccurate conclusions about the importance of mindfulness in the substance use prevention program. Analytic methods that account for measurement error should be used.

---

Much work regarding mediation measurement error has been in the context of the regression-based approach to mediation analysis, in which one estimates parameters from both an outcome and mediator regression model, and then calculates direct and indirect causal effects as functions of those regression parameters (Valeri and Vanderweele 2014). It is well-known that measurement errors can create bias in regression analysis (Carroll et al. 2006; Fuller 1987). As early as 1981, (Judd and Kenny, 1981a) discussed that mediator measurement error has the potential to bias estimates of the mediated effects. Baron and Kenny (1986) postulated that because measurement error in predictor variables in linear regression tends to attenuate the regression coefficients, measurement error in a mediator would be expected to bias the coefficient on the mediator in the outcome model toward zero (i.e., leading to indirect effects biased toward zero). Given the additive relationship between the direct and indirect effect, mediator measurement error would also be expected

to bias the coefficient for the treatment in the outcome model, leading to direct effects biased away from zero.

One method of addressing measurement error in this context is to apply a correction factor, based on the theorized magnitude of the bias. Ogburn and VanderWeele show that, when a single dichotomous mediator is measured with error, (1) the NIE of the error-prone mediator will generally be between zero and the true NIE for the error-free mediator and (2) the NDE will be between the true NDE and the total effect (Ogburn and VanderWeele 2012). These results are nonparametric and require no assumptions of a linear or generalized linear model for the outcome or mediator. In the context of linear, logistic regression, and generalized linear models for the outcome and mediator, numerous authors have given formulas for the bias caused by measurement errors in dichotomous and continuous mediators as well as the bias in the direct and indirect effects (Fritz et al. 2016; le Cessie et al. 2012; Valeri et al. 2014; Valeri and Vanderweele 2014; VanderWeele et al. 2012). Specifically, le Cessie et al. (2012) only considers direct effects; (VanderWeele et al. 2012) suggest how to extend the results to indirect effects. (Valeri et al. 2014; Valeri and Vanderweele 2014) further extend this work by developing correction approaches that allow for exposure–mediator interaction for both binary and continuous mediators.

An alternative approach to addressing measurement error in mediation analysis is use of structural equation models (SEM), which have been used for over 30 years for mediation analysis when there are multiple (error-prone) indicators for the latent variable (Brown 1997; Judd and Kenny, 1981b). For example, depression can be considered a latent construct, in that is not directly measured, but rather typically assessed via responses to multiple questions that are all related to depression. Using a SEM, one can specify a measurement model linking the multiple indicators to a latent variable and then specify models for the exposure, latent variables, and outcomes to study the indirect effects of an exposure through a latent variable such as depression (Hayes and Preacher 2010).

Alternatively, when the reliability of the error-prone measure is known, methods for directly correcting measurement error have been proposed, including the simulation and extrapolation method (SIMEX) (Cook and Stefanski 1994) or regression calibration (Carroll et al. 2006). In some fields, information about the measurement error may be available—e.g., in the context of educational or psychological measurement, in which psychometric properties of tests are explicitly measured. In other applications, such as self-reported substance use, the distribution of errors is generally unknown, although widely considered to exist. In these cases, direct adjustment will be impossible unless information about the measurement errors can be obtained.

Finally, as highlighted by (Ledgerwood and Shrout 2011), in some contexts it may be more desirable to use the biased estimates rather than correcting for measurement error, even when corrections are possible. They note that unadjusted estimated indirect effects using the observed error-prone mediator variables may be more precise than those obtained using latent regression or SEM approaches. Given that statistical power for estimation of indirect effects may be limited in some contexts, a more precise, but biased, estimate may provide more powerful hypothesis tests regarding the direct and indirect effects compared to a bias-corrected estimate. Overall, whether correcting for measurement error is desirable will depend on the details of any application, yet it is important that standard errors be calculated appropriately, and the relative precision be considered when correcting for measurement error in mediation analysis.

# 9 Challenge 7: clearly reporting results from mediation analyses

As noted earlier, mediation analyses have become increasingly common in recent years and some recent NIH funding mechanisms require investigation of mediating mechanisms. However, as we have highlighted in Challenges 1–6, applied researchers face numerous methodological challenges when conducting mediation analyses. This is reflected in the literature: multiple methodological reviews of the applied mediation literature have found that the execution and presentation of applied mediation studies is of varying quality and often lacking (e.g., Cashin et al. 2020; Gelfand et al. 2009; Liu et al. 2016; Rijnhart et al. 2021a, b; Stuart et al. 2022; Vo et al. 2020)). For example, while (Rijnhart et al. 2021a, b) found that approximately 75% of the 174 mediation studies they reviewed included a figure showing the proposed mediation model, Cashin et al. (2020) found that fewer than 15% of studies reviewed actually presented a conceptual justification for examining the proposed mediation pathway. Many studies performed mediation analyses using variables with murky temporal ordering – (Vo et al. 2020) found that the mediator and outcome were measured simultaneously in over 50% of mediation studies they reviewed and (Rijnhart et al. 2021a, b) found that nearly half of the studies reviewed used a cross-sectional design. Focusing on the psychology and psychiatry literature, Stuart et al. (2022) found that only about 25% of the 206 mediation analyses reviewed had full temporal ordering of exposure, mediator, and outcome, and that fewer than half of the papers controlled for confounders. Currently, the majority of mediation studies use a traditional mediation approach (e.g., Barron and Kenny), with approximately 5–15% of studies using a causal mediation framework (Rijnhart et al. 2021a, b; Stuart et al. 2022). Both (Stuart et al. 2022; Vo et al. 2020) found that a minority of studies stated the underlying assumptions of their mediation analysis and both (Rijnhart et al. 2021a, b; Stuart et al. 2022) found that less than 5% of studies performed any type of sensitivity analysis.

In order to advance the state of the science regarding mediation, reporting guidelines—i.e., AGReMA or "**A G**uideline for **Re**porting **M**ediation **A**nalyses"—have recently been proposed by Cashin et al. (2020) and (Lee et al. 2021). While reporting guidelines have existed for randomized trial study designs (see Consolidated Standards of Reporting Trials (CONSORT)) and for observational studies (see Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)), formal guidance regarding mediation analyses was previously quite limited. AGReMA was developed via a systematic, expert consensus process that focused on identifying essential elements that should be reported in mediation studies using experimental or observational data. AGReMA emphasizes that studies should clearly state and justify the causal question(s) being investigated, as highlighted by the following recommendations: (1) clearly state the mediation question of interest; (2) provide rationale for investigating the specified mediation pathway(s); (3) present a figure depicting the underlying causal model (including confounders); and (4) state which specific mediation effects are of interest. Additionally, AGReMA highlights key recommendations regarding measurement and analysis, including: (1) specify how and when all variables (e.g., exposures, mediators, outcomes, confounders)—were measured; (2) report details of the statistical methods used to estimate causal effects, including how potential confounding was addressed; and (3) report and interpret estimated effects, (including uncertainty estimates). AGReMA also recommends that studies: (1) specify all assumptions underlying the analyses reported; (2) discuss the plausibility of underlying assumptions; (3) report results from sensitivity analyses assessing the robustness of study findings; and (4) discuss study limitations, additional potential sources of bias, and generalizability of findings.

There is both a long-form AGReMA checklist for studies that are conducting mediation as a primary analysis, as well as a short-form checklist for when mediation is a secondary study aim (Lee et al. 2021).

We highlight that AGReMA provides guidance on *reporting* of mediation analyses, rather than guidelines regarding the *analytic execution* of mediation analyses. By encouraging greater transparency and more detailed methodological description when reporting mediation studies, the AGReMA guidelines seek to make it easier to assess the methodological quality of mediation studies. While these guidelines provide a comprehensive overview of best reporting practices, the ultimate success of these endeavors ultimately lies with journal editors, reviewers, and granting agencies, and researchers themselves to ensure that standards are met in applied studies.

In additional to AGReMA, other guidance for reporting mediation analyses has been offered in the literature. For example, concerns have been raised about the utility of an often-reported mediation quantity, the "proportion mediated" which is generally calculated as the ratio between the NIE and total effect. Specifically, this quantity may be ill-defined in the presence of exposure-mediator interactions (i.e., the proportion mediated may vary across levels of the mediator) or "inconsistent mediation" when the total effect is near zero or of the opposite sign as the NIE (VanderWeele 2013). VanderWeele alternatively proposes the reporting of the proportion of the total effect that could be eliminated by an intervention that set the mediator to a fixed level (VanderWeele 2013).

Finally, conducting systematic reviews and meta-analyses of mediation studies is an emerging area, given the growing number of published mediation studies. However, as we have highlighted in this paper, any single mediation analysis entails a myriad of decision points regarding variable measurement, hypothesized causal structure, mediation effects of interest, and estimation methods, including strategies to mitigate potential confounding and assess robustness to violations of underlying assumptions or model misspecification. In practice, mediation analyses investigating related mediating mechanisms may vary so substantially that it is infeasible to pool any information across studies. A recent paper outlined some key difficulties regarding synthesizing mediation analyses for systematic reviews or meta-analyses falling under 3 broad challenges: (1) identification of eligible studies; (2) assessing the presence of different sources of biases in eligible studies, and (3) synthesizing quantitative findings across eligible studies (Vo and Vansteelandt 2022). Ongoing methodological work is needed in this area.

## 10 Conclusion

This paper provides a summary of 7 key challenges encountered when conducting mediation analysis as well as a discussion of how applied researchers can address each challenge. Mediation analysis entails an examination of causal mechanisms—this fundamental motivation should be reflected in the design and interpretation of mediation analyses. When designing mediation studies, it is important to clearly identify the objectives of mediation analysis and how these differ from objectives of other types of "third variable" analyses. Additionally, in the literature, more attention is needed regarding the assumptions inherently underlying mediation analysis. The field of causal mediation has much to offer in this regard, as the application of a causal inference framework has served both to differentiate specific types of mediation effects that can be estimated as well as to formalize the necessary assumptions underlying each effect type. Mediation analyses generally require more

assumptions than traditional bivariate exposure-outcome analyses, underscoring the importance of assessing the robustness of study findings to assumption violations via sensitivity analyses. To promote adoption of mediation sensitivity analyses, methodologists should strive to disseminate both knowledge of and accessible software to implement these methods. Finally, there is a need for better reporting of mediation analysis—the development of AGReMA represents a very promising direction for the field. Overall, we hope that this paper will help advance the practice of mediation analysis, leading to publication of more robust mediation studies in the health services and health policy fields.

## Declarations

**Conflict of interest** No authors reported any financial or other conflicts of interest in relation to the work described.

## References

Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. **34**(28), 3661–3679 (2015). https://doi.org/10.1002/sim.6607

Baron, R.M., Kenny, D.A.: The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. **51**(6), 1173–1182 (1986). https://doi.org/10.1037//0022-3514.51.6.1173

Berli, C., Inauen, J., Stadler, G., Scholz, U., Shrout, P.E.: Understanding between-person interventions with time-intensive longitudinal outcome data: longitudinal mediation analyses. Ann. Behav. Med. **55**(5), 476–488 (2021). https://doi.org/10.1093/abm/kaaa066

Bind, M.A., Vanderweele, T.J., Coull, B.A., Schwartz, J.D.: Causal mediation analysis for longitudinal data with exogenous exposure. Biostatistics **17**(1), 122–134 (2016). https://doi.org/10.1093/biostatistics/kxv029

Brown, R.L.: Assessing specific mediational effects in complex theoretical models. Struct. Equ. Modeling **4**, 142–156 (1997)

Cai, X., Coffman, D.L., Piper, M.E., Li, R.: Estimation and inference for the mediation effect in a time-varying mediation model. BMC Med. Res. Methodol. **22**(1), 113 (2022). https://doi.org/10.1186/s12874-022-01585-x

Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M.: Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC (2006)

Cashin, A.G., Lee, H., Lamb, S.E., Hopewell, S., Mansell, G., Williams, C.M., Kamper, S.J., Henschke, N., McAuley, J.H.: An overview of systematic reviews found suboptimal reporting and methodological

limitations of mediation studies investigating causal mechanisms. J. Clin. Epidemiol. **111**, 60–68 (2019). https://doi.org/10.1016/j.jclinepi.2019.03.005

Cashin, A.G., McAuley, J.H., Lamb, S.E., Hopewell, S., Kamper, S.J., Williams, C.M., Henschke, N., Lee, H.: Development of a guideline for reporting mediation analyses (AGReMA). BMC Med. Res. Methodol. **20**(1), 19 (2020). https://doi.org/10.1186/s12874-020-0915-5

Chakraborti, Y., Coffman, D.L., Piper, M.E.: Time-varying mediation of pharmacological smoking cessation treatments on smoking lapse via craving, cessation fatigue, and negative mood. Nicotine Tob. Res. (2022). https://doi.org/10.1093/ntr/ntac068

Coffman, D.L., Zhong, W.: Assessing mediation using marginal structural models in the presence of confounding and moderation. Psychol. Methods **17**(4), 642–664 (2012). https://doi.org/10.1037/a0029311

Coffman, D. L.: Third variables: scientific meanings and modeling in non-randomized studies. In S. J. Henley (Ed.), Routledge international handbook of advanced quantitative methods in nursing research. routledge international (2015)

Cole, D.A., Preacher, K.J.: Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error. Psychol. Methods **19**(2), 300–315 (2014). https://doi.org/10.1037/a0033805

Cook, J.R., Stefanski, L.A.: Simulation-extrapolation estimation in parametric measurement error models. JASA **89**(428), 1314–1328 (1994)

Daniel, R.M., De Stavola, B.L., Cousens, S.N., Vansteelandt, S.: Causal mediation analysis with multiple mediators. Biometrics **71**(1), 1–14 (2015). https://doi.org/10.1111/biom.12248

De Stavola, B.L., Daniel, R.M., Ploubidis, G.B., Micali, N.: Mediation analysis with intermediate confounding: Structural equation modeling viewed through the causal inference lens. Am. J. Epidemiol. **181**(1), 64–80 (2015). https://doi.org/10.1093/aje/kwu239

Del Boca, F.K., Noll, J.A.: Truth or consequences: the validity of self-report data in health services research on addictions. Addiction **95**(3), S347-360 (2000). https://doi.org/10.1080/09652140020004278

Didelez, V., Dawid, A. P., Geneletti, S.: Direct and indirect effects of sequential treatments. Proceedings of the 22nd conference on uncertainty in artificial intelligence (2006)

Fairchild, A.J., Cai, C., McDaniel, H., Shi, D., Gottschall, A., Masyn, K.E.: Evaluating a method to estimate mediation effects with discrete-time survival outcomes. Front. Psychol. **10**, 740 (2019). https://doi.org/10.3389/fpsyg.2019.00740

Fritz, M.S., Kenny, D.A., MacKinnon, D.P.: The combined effects of measurement error and omitting confounders in the single-mediator model. Multivariate Behav Res **51**(5), 681–697 (2016). https://doi.org/10.1080/00273171.2016.1224154

Fuller, W.A.: Measurement error models. John Wiley & Sons (1987)

Gao, T., Albert, J.M.: Bayesian causal mediation analysis with multiple ordered mediators. Stat Modelling **19**(6), 634–652 (2019). https://doi.org/10.1177/1471082x18798067

Gelfand, L.A., Mensinger, J.L., Tenhave, T.: Mediation analysis: a retrospective snapshot of practice and more recent directions. J. Gen. Psychol. **136**(2), 153–176 (2009). https://doi.org/10.3200/GENP.136.2.153-178

Gollob, H.F., Reichardt, C.S.: Taking account of time lags in causal models. Child Develop. **58**(1), 80–92 (1987)

Gunzler, D., Tang, W., Lu, N., Wu, P., Tu, X.M.: A class of distribution-free models for longitudinal mediation analysis. Psychometrika **79**(4), 543–568 (2014). https://doi.org/10.1007/s11336-013-9355-z

Hayes, A.F., Preacher, K.J.: Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. Multivar. Behav. Res. **45**(4), 627–660 (2010). https://doi.org/10.1080/00273171.2010.498290

Hill, A.B.: The environment and disease: association or causation? Proceed. Royal Soc. Med. **58**, 295–300 (1965)

Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986). https://doi.org/10.2307/2289069

Holmberg, M.J., Andersen, L.W.: Collider bias. JAMA (2022). https://doi.org/10.1001/jama.2022.1820

Hong, G.: Causality in a social world: moderation, mediation and spill-over. Wiley (2015)

Hong, G., Deutsch, J., Hill, H.D.: Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. J Educ Behav Stat **40**(3), 307–340 (2015)

Hong, G., Qin, X., Yang, F.: Weighting-based sensitivity analysis in causal mediation studies. J Educ Behav Stat **43**(1), 32–56 (2018)

Hong, G., Yang, F., Qin, X.: Posttreatment confounding in causal mediation studies: a cutting-edge problem and a novel solution via sensitivity analysis. Biometrics (2022). https://doi.org/10.1111/biom.13705

Imai, K., Yamamoto, T.: Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. Polit. Anal. **21**(2), 141–171 (2013). https://doi.org/10.1093/pan/mps040

Imai, K., Keele, L., Tingley, D.: A general approach to causal mediation analysis. Psychol. Methods **15**(4), 309–334 (2010). https://doi.org/10.1037/a0020761

Jerolon, A., Baglietto, L., Birmele, E., Alarcon, F., Perduca, V.: Causal mediation analysis in presence of multiple mediators uncausally related. Int. J. Biostat. **17**(2), 191–221 (2020). https://doi.org/10.1515/ijb-2019-0088

Johnson, T., Fendrich, M.: Modeling sources of self-report bias in a survey of drug use epidemiology. Ann. Epidemiol. **15**(5), 381–389 (2005). https://doi.org/10.1016/j.annepidem.2004.09.004

Judd, C.M., Kenny, D.A.: Estimating the effects of social intervention. Cambridge University Press, Cambridge (1981a)

Judd, C.M., Kenny, D.A.: Process analysis–estimating mediation in treatment evaluations. Eval. Rev. **5**, 602–619 (1981b)

Keele, L., Tingley, D., Yamamoto, T.: Identifying mechanisms behind policy interventions via causal mediation analysis. J. Policy Anal. Manage. **34**(4), 937–963 (2015)

Kraemer, H.C., Kiernan, M., Essex, M., Kupfer, D.J.: How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. Health Psychol. **27**(2S), S101-108 (2008). https://doi.org/10.1037/0278-6133.27.2(Suppl.).S101

Lange, T., Rasmussen, M., Thygesen, L.C.: Assessing natural direct and indirect effects through multiple pathways. Am. J. Epidemiol. **179**(4), 513–518 (2013). https://doi.org/10.1093/aje/kwt270

le Cessie, S., Debeij, J., Rosendaal, F.R., Cannegieter, S.C., Vandenbroucke, J.P.: Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. Epidemiology **23**(4), 551–560 (2012). https://doi.org/10.1097/EDE.0b013e318254f5de

Ledgerwood, A., Shrout, P.E.: The trade-off between accuracy and precision in latent variable models of mediation processes. J. Personal. Soc. Psychol. **101**(6), 1174–1188 (2011)

Lee, H., Cashin, A.G., Lamb, S.E., Hopewell, S., Vansteelandt, S., VanderWeele, T.J., MacKinnon, D.P., Mansell, G., Collins, G.S., Golub, R.M., McAuley, J.H.: A guideline for reporting mediation analyses of randomized trials and observational studies: the AGReMA statement. JAMA **326**(11), 1045–1056 (2021). https://doi.org/10.1001/jama.2021.14075

Lin, S.H., Young, J., Logan, R., Tchetgen Tchetgen, E.J., VanderWeele, T.J.: Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. Epidemiology **28**(2), 266–274 (2017a). https://doi.org/10.1097/EDE.0000000000000609

Lin, S.H., Young, J.G., Logan, R., VanderWeele, T.J.: Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. Stat. Med. **36**(26), 4153–4166 (2017b). https://doi.org/10.1002/sim.7426

Lindenberger, U., von Oertzen, T., Ghisletta, P., Hertzog, C.: Cross-sectional age variance extraction: what's change got to do with it? Psychol. Aging **26**(1), 34–47 (2011). https://doi.org/10.1037/a0020525

Lipsky, A.M., Greenland, S.: Causal directed acyclic graphs. JAMA **327**(11), 1083–1084 (2022). https://doi.org/10.1001/jama.2022.1816

Liu, S.H., Ulbricht, C.M., Chrysanthopoulou, S.A., Lapane, K.L.: Implementation and reporting of causal mediation analysis in 2015: a systematic review in epidemiological studies. BMC. Res. Notes **9**, 354 (2016). https://doi.org/10.1186/s13104-016-2163-7

Lok, J.J.: Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. Stat. Med. **35**(22), 4008–4020 (2016). https://doi.org/10.1002/sim.6990

Lok, J.J., Bosch, R.J.: Causal organic indirect and direct effects: closer to the original approach to mediation analysis, with a product method for binary mediators. Epidemiology **32**(3), 412–420 (2021). https://doi.org/10.1097/EDE.0000000000001339

Ludwig, J., Kling, J.R., Mullainathan, S.: Mechanism experiments and policy evaluations. J. Econ. Perspect. **25**, 17–38 (2011)

MacKinnon, D.P.: Introduction to statistical mediation analysis. Taylor & Francis (2008)

MacKinnon, D.P.: Contrasts in multiple mediator models. In: Rose, J.S., Chassin, L., Presson, C.C., Sherman, S.J. (eds.) Multivariate applications in substance use research: new methods for new questions, pp. 141–160. Psychology Press (2015)

MacKinnon, D.P., Valente, M.J., Gonzalez, O.: The correspondence between causal and traditional mediation analysis: the link is the mediator by treatment interaction. Prev. Sci. **21**(2), 147–157 (2020). https://doi.org/10.1007/s11121-019-01076-4

Massarwi, A.A., Cluver, L., Meinck, F., Doubt, J., Lachman, J.M., Shenderovich, Y., Green, O.: Mediation pathways for reduced substance use among parents in South Africa: a randomized controlled trial. BMC Public Health **21**(1), 1656 (2021). https://doi.org/10.1186/s12889-021-11651-6

Maxwell, S.E., Cole, D.A.: Bias in cross-sectional analyses of longitudinal mediation. Psychol. Methods **12**(1), 23–44 (2007). https://doi.org/10.1037/1082-989X.12.1.23

Maxwell, S.E., Cole, D.A., Mitchell, M.A.: Bias in cross-sectional analyses of longitudinal mediation: partial and complete mediation under an autoregressive model. Multivar. Behav. Res. **46**(5), 816–841 (2011). https://doi.org/10.1080/00273171.2011.606716

McCandless, L.C., Somers, J.M.: Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. Stat. Method. Med. Res. **28**(2), 515–531 (2019). https://doi.org/10.1177/0962280217729844

McGowan, L. D. A., Gerke, T., Barrett, M.: Causal inference is not just a statistics problem (2023)

Miles, B.C.H., Shpitser, I., Kanki, P., Meloni, S., Tchetgen, E.J.T.: On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. Biometrika **107**(1), 159–172 (2020). https://doi.org/10.1093/biomet/asz063

Moerkerke, B., Loeys, T., Vansteelandt, S.: Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. Psychol. Method. **20**(2), 204–220 (2015). https://doi.org/10.1037/a0036368

Morral, A.R., McCaffrey, D., Iguchi, M.Y.: Hardcore drug users claim to be occasional users: drug use frequency underreporting. Drug Alcohol Depend. **57**(3), 193–202 (2000). https://doi.org/10.1016/s0376-8716(99)00048-4

Natarajan, L., Pu, M., Fan, J., Levine, R.A., Patterson, R.E., Thomson, C.A., Rock, C.L., Pierce, J.P.: Measurement error of dietary self-report in intervention trials. Am. J. Epidemiol. **172**(7), 819–827 (2010). https://doi.org/10.1093/aje/kwq216

Nguyen, T.Q., Schmid, I., Stuart, E.A.: Clarifying causal mediation analysis for the applied researcher: defining effects based on what we want to learn. Psychol. Method. **26**(2), 255–271 (2020). https://doi.org/10.1037/met0000299

Nguyen, T.Q., Schmid, I., Ogburn, E.L., Stuart, E.A.: Clarifying causal mediation analysis: effect identification via three assumptions and five potential outcomes. J. Causal Inference **10**(1), 246–279 (2022)

Nguyen, T.Q., Ogburn, E.L., Schmid, I., Sarker, E.B., Greifer, N., Koning, I.M., Stuart, E.A.: Causal mediation analysis: from simple to more robust strategies for estimation of marginal natural (in) direct effects. Stat. Surv. **17**, 1–41 (2023)

Ogburn, E.L., VanderWeele, T.J.: Analytic results on the bias due to nondifferential misclassification of a binary mediator. Am. J. Epidemiol. **176**(6), 555–561 (2012). https://doi.org/10.1093/aje/kws131

O'Laughlin, K.D., Martin, M.J., Ferrer, E.: Cross-sectional analysis of longitudinal mediation processes. Multivar. Behav. Res. **53**(3), 375–402 (2018). https://doi.org/10.1080/00273171.2018.1454822

O'Rourke, H.P., MacKinnon, D.P.: When the test of mediation is more powerful than the test of the total effect. Behav. Res. Method. **47**(2), 424–442 (2015). https://doi.org/10.3758/s13428-014-0481-z

Pearl, J.: Direct and indirect effects. Proceedings of the 17th conference on uncertainty in artificial intelligence 411–420 (2001)

Preacher, K.J.: Advances in mediation analysis: A survey and synthesis of new developments. Annu. Rev. Psychol. **66**, 825–852 (2015)

Preacher, K.J., Hayes, A.F.: Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav. Res. Method. **40**(3), 879–891 (2008). https://doi.org/10.3758/brm.40.3.879

Preacher, K.J., Rucker, D.D., Hayes, A.F.: Addressing moderated mediation hypotheses: theory, methods, and prescriptions. Multivar. Behav. Res. **42**(1), 185–227 (2007). https://doi.org/10.1080/00273170701341316

Rijnhart, J.J.M., Twisk, J.W.R., Chinapaw, M.J.M., de Boer, M.R., Heymans, M.W.: Comparison of methods for the analysis of relatively simple mediation models. Contemp Clin Trials Commun **7**, 130–135 (2017). https://doi.org/10.1016/j.conctc.2017.06.005

Rijnhart, J.J.M., Twisk, J.W.R., Eekhout, I., Heymans, M.W.: Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. BMC Med. Res. Methodol. **19**(1), 19 (2019). https://doi.org/10.1186/s12874-018-0654-z

Rijnhart, J.J.M., Lamp, S.J., Valente, M.J., MacKinnon, D.P., Twisk, J.W.R., Heymans, M.W.: Mediation analysis methods used in observational research: a scoping review and recommendations. BMC Med. Res. Methodol. **21**(1), 226 (2021a). https://doi.org/10.1186/s12874-021-01426-3

Rijnhart, J.J.M., Valente, M.J., MacKinnon, D.P., Twisk, J.W.R., Heymans, M.W.: The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction. Struct. Equ. Modeling **28**(3), 345–355 (2021b). https://doi.org/10.1080/10705511.2020.1811709

Rijnhart, J.J.M., Twisk, J.W.R., Deeg, D.J.H., Heymans, M.W.: Assessing the robustness of mediation analysis results using multiverse analysis. Prev. Sci. **23**(5), 821–831 (2022). https://doi.org/10.1007/s11121-021-01280-1

Rijnhart, J.J.M., Valente, M.J., Smyth, H.L., MacKinnon, D.P.: Statistical mediation analysis for models with a binary mediator and a binary outcome: the differences between causal and traditional mediation analysis. Prev. Sci. **24**(3), 408–418 (2023). https://doi.org/10.1007/s11121-021-01308-6

Robins, J.M.: Semantics of causal DAG models and the identification of direct and indirect effects. In: Green, P., Hjort, N., Richardson, S. (eds.) Highly structured stochastic systems, pp. 70–81. Oxford University Press (2003)

Robins, J.M., Greenland, S.: Identifiability and exchangeability for direct and indirect effects. Epidemiology **3**(2), 143–155 (1992). https://doi.org/10.1097/00001648-199203000-00013

Rubio, A., Oyanedel, J.C., Cancino, F., Benavente, L., Cespedes, C., Zisis, C., Paez, D.: Social support and substance use as moderators of the relationship between depressive symptoms and suicidal ideation in adolescents. Front. Psychol. **11**, 539165 (2020). https://doi.org/10.3389/fpsyg.2020.539165

Selig, J.P., Preacher, K.J.: Mediation models for longitudinal data in developmental research. Res. Hum. Dev. **6**(2–3), 144–164 (2009). https://doi.org/10.1080/15427600902911247

Shi, B., Choirat, C., Coull, B.A., VanderWeele, T.J., Valeri, L.: CMAverse: a suite of functions for reproducible causal mediation analyses. Epidemiology **32**(5), e20–e22 (2021). https://doi.org/10.1097/EDE.0000000000001378

Sobel, M.E.: Asymptotic intervals for indirect effects in structural equations models. In: Leinhart, S. (ed.) Sociological methodology 1982, pp. 290–312. Jossey-Bass (1982)

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. **11**, 702–712 (2016). https://doi.org/10.1177/1745691616658637

Steen, J., Loeys, T., Moerkerke, B., Vansteelandt, S.: Flexible mediation analysis with multiple mediators. Am. J. Epidemiol. **186**(2), 184–193 (2017). https://doi.org/10.1093/aje/kwx051

Stuart, E.A., Schmid, I., Nguyen, T., Sarker, E., Pittman, A., Benke, K., Rudolph, K., Badillo-Goicoechea, E., Leoutsakos, J.M.: Assumptions not often assessed or satisfied in published mediation analyses in psychology and psychiatry. Epidemiol. Rev. **43**(1), 48–52 (2022). https://doi.org/10.1093/epirev/mxab007

Taguri, M., Featherstone, J., Cheng, J.: Causal mediation analysis with multiple causally non-ordered mediators. Stat. Methods Med. Res. **27**(1), 3–19 (2018). https://doi.org/10.1177/0962280215615899

Tai, A.S., Huang, Y.T., Yang, H.I., Lan, L.V., Lin, S.H.: G-computation to causal mediation analysis with sequential multiple mediators-investigating the vulnerable time window of HBV activity for the mechanism of HCV induced hepatocellular carcinoma. Front. Public Health **9**, 757942 (2021). https://doi.org/10.3389/fpubh.2021.757942

Tchetgen Tchetgen, E.J.: On causal mediation analysis with a survival outcome. Int. J. Biostat. (2011). https://doi.org/10.2202/1557-4679.1351

Tchetgen Tchetgen, E.J., Vanderweele, T.J.: Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Epidemiology **25**(2), 282–291 (2014). https://doi.org/10.1097/EDE.0000000000000054

Valente, M.J., Pelham, W.E., Smyth, H., MacKinnon, D.P.: Confounding in statistical mediation analysis: what it is and how to address it. J. Couns. Psychol. **64**(6), 659–671 (2017). https://doi.org/10.1037/cou0000242

Valeri, L., Vanderweele, T.J.: Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychol. Method. **18**(2), 137–150 (2013). https://doi.org/10.1037/a0031034

Valeri, L., Vanderweele, T.J.: The estimation of direct and indirect causal effects in the presence of misclassified binary mediator. Biostatistics **15**(3), 498–512 (2014). https://doi.org/10.1093/biostatistics/kxu007

Valeri, L., Lin, X., VanderWeele, T.J.: Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. Stat. Med. **33**(28), 4875–4890 (2014). https://doi.org/10.1002/sim.6295

VanderWeele, T.J.: Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology **21**(4), 540–551 (2010). https://doi.org/10.1097/EDE.0b013e3181df191c

VanderWeele, T.J.: Policy-relevant proportions for direct effects. Epidemiology **24**(1), 175–176 (2013). https://doi.org/10.1097/EDE.0b013e3182781410

VanderWeele, T.J.: Explanation in causal inference: methods for mediation and interaction. Oxford University Press, Oxford (2015)

VanderWeele, T.J., Chiba, Y.: Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. Epidemiol. Biostat. Public Health (2014). https://doi.org/10.2427/9027

VanderWeele, T.J., Tchetgen Tchetgen, E.J.: Mediation analysis with time varying exposures and mediators. J. R. Stat. Soc. Ser. B Stat Methodol. **79**(3), 917–938 (2017). https://doi.org/10.1111/rssb.12194

VanderWeele, T.J., Valeri, L., Ogburn, E.L.: The role of measurement error and misclassification in mediation analysis: Mediation and measurement error. Epidemiology **23**(4), 561–564 (2012). https://doi.org/10.1097/EDE.0b013e318258f5e4

Vanderweele, T.J., Vansteelandt, S., Robins, J.M.: Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology **25**(2), 300–306 (2014). https://doi.org/10.1097/EDE.0000000000000034

Vansteelandt, S., Daniel, R.M.: Interventional effects for mediation analysis with multiple mediators. Epidemiology **28**(2), 258–265 (2017). https://doi.org/10.1097/EDE.0000000000000596

Vansteelandt, S., Vanderweele, T.J.: Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. Biometrics **68**(4), 1019–1027 (2012). https://doi.org/10.1111/j.1541-0420.2012.01777.x

Vansteelandt, S., Linder, M., Vandenberghe, S., Steen, J., Madsen, J.: Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. Stat. Med. **38**(24), 4828–4840 (2019). https://doi.org/10.1002/sim.8336

Vo, T.T., Vansteelandt, S.: Challenges in systematic reviews and meta-analyses of mediation analyses. Am. J. Epidemiol. **191**(6), 1098–1106 (2022). https://doi.org/10.1093/aje/kwac028

Vo, T.T., Superchi, C., Boutron, I., Vansteelandt, S.: The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. J. Clin. Epidemiol. **117**, 78–88 (2020). https://doi.org/10.1016/j.jclinepi.2019.10.001

Vo, T.T., Davies-Kershaw, H., Hackett, R., Vansteelandt, S.: Longitudinal mediation analysis of time-to-event endpoints in the presence of competing risks. Lifetime Data Anal. **28**(3), 380–400 (2022). https://doi.org/10.1007/s10985-022-09555-7

Wickramarachchi, D.S., Lim, L.H.M., Sun, B.: Mediation analysis with multiple mediators under unmeasured mediator-outcome confounding. Stat. Med. **42**(4), 422–432 (2023). https://doi.org/10.1002/sim.9624

Young, C., Holsteen, K.: Model uncertainty and robustness: a computational framework for multimodel analysis. Soc. Method. Res. **46**, 3–40 (2017)

Zeng, S., Rosenbaum, S., Alberts, S.C., Archie, E.A., Li, F.: Causal mediation analysis for sparse and irregular longitudinal data. Ann. Appl. Stat. **15**(2), 747–767 (2021)