Recommendations for improving causal inference of mediation analysis

Wen Wei Loh¹, and Dongning Ren²

¹ Department of Data Analysis, Ghent University, Gent, Belgium

² Department of Social Psychology, Tilburg University, Tilburg, The Netherlands

Draft version 2.0, 12 September, 2021. This article has not been peer reviewed.

Please do not copy or cite without first author's permission.

Abstract

Mediation analysis is an indispensable tool for investigating how a treatment causally affects an outcome via intermediate variables. Recently, there have been increased concerns about the validity of causal inferences in conclusions drawn using mediation analysis. However, the discussions are limited to a single mediator, and importantly, there is a lack of guidelines on substantiating causal inferences. In this article, we first provide a thorough examination of the causal assumptions underpinning mediation analysis. We pay particular attention to the practice of exploring mediated effects along various paths linking several mediators and the stringent – yet often overlooked – assumptions that predicate valid inference. To mitigate the risk of invalid inference, we introduce an alternative approach focusing on mediator-specific indirect effects. An appealing feature of this approach is that valid causal inference of mediation analysis with multiple mediators does not necessitate assuming a (correct) causal structure among the mediators. Finally, we provide a practical guide to improve the research practice of mediation analysis. We clarify when mediation analysis is (in)appropriate; when appropriate, we recommend that researchers preregister (i) justifications of the asserted causal relations in their mediation analysis, (ii) pre-treatment confounders, and (iii) the either path- or mediator-specific indirect effects to be investigated. Confounding adjustment when estimating the preregistered indirect effects and sensitivity analyses for unmeasured confounding can fortify causal inferences in the conclusions. We hope this article will encourage explication, justification, and reflection of the causal assumptions underpinning mediation analysis to improve the validity of causal inferences in psychology research.

Keywords: Causal diagrams; Confounding; Interventional indirect effects; Parallel mediation models; Serial mediation models

Recommendations for improving causal inference of mediation analysis

Mediation analysis is an invaluable etiologic research tool for investigating the mechanisms driving the effect of a treatment on an outcome. It is regularly used by social and personality psychologists to understand the causal mechanisms of a treatment effect across a broad span of topics. For example, recent publications in the *Journal of Personality and Social Psychology* use mediation analysis to understand: how experience of awe (treatment) affects life satisfaction (outcome) via appraisal of vastness in relation to the self, and changes in daily stress levels (Bai et al., 2021); how either support- or oppose-framed attitudes (treatment) affects sharing (outcome) via value expressiveness and impression management (Catapano & Tormala, 2021); how perceived target's extraversion (treatment) affects target's reputation of having leadership potential (outcome) via target's reputation of extraversion (Costello & Srivastava, 2021); how helping opportunities (treatment) affect moral outrage (outcome) via perceived reputation benefits of punishment, and general reputation concerns (Jordan & Rand, 2020); how friendship jealousy (treatment) affects intent to engage in friend guarding via jealousy, neutrality, happiness, sadness, and anger (Krems et al., 2021); and how search for meaning in life (non-randomized treatment) affects support for political violence (outcome) via sensation seeking and willingness to self-sacrifice (Schumpe et al., 2020).

Because mediation analyses assess causal relations using statistical methods, their validity are susceptible to violations of either causal or statistical assumptions. Psychologists have recently started to raise concerns about causal inferences in mediation analyses (Bullock et al., 2010; Cuartas & McCoy, 2021; Fiedler et al., 2018; Kline, 2015; Pek & Hoyle, 2016; Rohrer et al., 2021; Tate, 2015; Trafimow, 2015). However, the existing literature is limited in two ways. First, although researchers commonly hypothesize multiple possible mediators (see the examples provided in the previous paragraph), the current discussions of the methodological issues with mediation analysis have only focused on settings with a single mediator, while leaving the issues with multiple mediators

unaddressed. Second, and just as importantly, the current discussions have focused on criticisms of causal claims in conclusions drawn from mediation analysis and advocated understatement of causality, rather than offer concrete suggestions to substantiate and fortify the causal inferences.

The goal of the current article is to improve causal inferences of mediation analysis in psychology research, by developing a set of practical guidelines for rigorous research practice when conducting mediation analysis in randomized experiments. To do so, we provide a thorough and coherent discussion of the causal assumptions underpinning mediation analysis. In our discussion, we draw on modern causal concepts of, and statistical methods for, mediation analysis that are currently dispersed across the behavioral, epidemiological, and social sciences. References to additional resources (including articles, books, and software) for further details are provided throughout. We first explicate the causal assumptions for both single and multiple mediator settings. In particular, we address the practice of exploring mediated effects along various paths traversing (causally) linked mediators using serial mediation models. We describe the (often implicit) stringent causal assumptions necessary for valid causal inferences when presupposing causally ordered mediators. When these assumptions fail to hold – as is likely in practice – invalid conclusions about the causal mechanisms follow. To address these challenges, we introduce a state-of-the-art framework focusing on mediator-specific indirect effects that relies on fewer causal assumptions than conventional approaches for valid causal inferences. In practice, researchers adopting this framework can investigate indirect effects via each possible mediator without concern about invalid inferences or misleading conclusions due to incorrectly assuming the underlying (and usually unknown) causal structure among the mediators. Finally, we provide a practical guide for improving the research practice when using mediation analysis in randomized experiments. We offer clear recommendations at each study phase, from study design to data analysis and reporting, to improve readers' and researchers' confidence in the credibility of causal inferences in mediation analysis.

Drawing causal relations using causal diagrams

To facilitate the discussions of mediation analyses considered in this article, we briefly introduce key features of the *causal diagram* approach to causality (Pearl, 2009). Mediation analyses are inherently based on assumptions about the underlying data-generating processes that must hold true for the substantive conclusions to be valid. Causal diagrams, such as those presented in Figures 1 and 2, are intuitive graphical representations of assumed causal relations, and thus extremely well-suited for researchers planning mediation analyses to “draw their assumptions before their conclusions” (Hernán & Robins, 2020). A causal diagram is a *causal graphical model*, or *causal directed acyclic graph* (DAG), that represents causal relations among a set of variables. In a causal diagram, vertices (or *nodes*) represent variables, and a directed edge (or *arrow*) from one node to another signifies that the variable the arrow emanates from causes the variable that the arrow enters. Causal diagrams may resemble traditional path diagrams in the structural equation modeling (SEM) framework, but there are key differences between them (Bollen & Pearl, 2013; Pearl, 2012). First, unlike path diagrams, causal diagrams encode causal relations that do not depend on statistical assumptions about the nature of the relationships between the variables (Elwert, 2013). It is thus unnecessary to specify the functional dependence of each variable on its causally antecedent variables, or to display path coefficients and error terms. Second, unlike traditional path diagrams which are (typically) drawn based on the observed data, causal diagrams are drawn based on theoretical knowledge and external empirical evidence to accurately represent data-generating processes.¹

To illustrate specific concepts of causal diagrams in the context of mediation analysis, suppose that a researcher is planning to investigate whether the effect of a randomized treatment X on an outcome of interest Y is mediated by two intervening variables, M_1 and

¹In addition to the aforementioned references, broader presentations of causal diagrams are given by e.g., Hayduk et al. (2003), and Rohrer (2018), while metaphysical and philosophical concepts of causality and cause are discussed by Morgan & Winship (2015) and Pearl & MacKenzie (2018).

M_2 .² These theorized causal relations can be drawn using causal diagrams such as those in Figure 1. For example, the arrow drawn directly from X to Y represents the treatment’s direct influence on the outcome via unknown mechanisms which did not involve either mediator. The absence of an arrow between the mediators, such as in Figure 1(a), precludes the possibility of a direct causal effect between them. The path $X \rightarrow M_1 \rightarrow Y$ describes the causal effect that is transmitted first from the treatment X to the mediator M_1 , then from the mediator M_1 to the outcome Y . This is an example of a *directed path* connecting a sequence of variables using arrows that are all oriented in the same direction, with no recurring variables on that path. Note that causal effects along different directed paths linking two variables, possibly intersecting other variables, can be combined using *path analysis* (Duncan, 1966; Wright, 1934). For example, the treatment’s *total* effect on the outcome in Figure 1(a) is the combined effect along three different directed paths: the path via M_1 ($X \rightarrow M_1 \rightarrow Y$), the path via M_2 ($X \rightarrow M_2 \rightarrow Y$), and the path via neither M_1 or M_2 ($X \rightarrow Y$).

More generally, paths can link variables using arrows pointing in different directions. For example, in Figure 1, the covariate(s) C simultaneously influence the mediators and outcome, as depicted by arrows from C entering M_1 , M_2 , and Y . Each mediator is therefore linked via C to either another mediator, e.g., along the path $M_1 \leftarrow C \rightarrow M_2$; or the outcome, e.g., along the path $M_1 \leftarrow C \rightarrow Y$. We denote such common causes, henceforth termed *confounders*, of the mediator-mediator, or mediator-outcome, relations collectively by C . Confounders in C that are unadjusted for, because they are either unmeasured, or measured but omitted from the statistical analysis, generate non-causal (i.e., “spurious”) associations that can render statistical biases in causal effect estimates. Confounding is one of the key issues that can severely undermine causal inferences of mediation analysis. We elaborate on this point in the next section.

²In this article, the subscripts indexing the mediators are used simply to distinguish the mediators, and need not represent or imply an assumed causal ordering; i.e., M_1 need not causally precede M_2 .

Figure 1. Causal diagrams with two mediators where either (a) M_1 and M_2 do not causally influence each other directly, or (b) M_1 causally precedes M_2 , or (c) M_1 and M_2 do not affect each other but share an unmeasured common cause U , or (d) M_2 causally precedes M_1 . Rectangular nodes denote observed variables, while round nodes denote unmeasured (or “hidden”) variables. For visual clarity, common causes C are represented by a single node, and arrows emanating from C are drawn in gray, although the effects of C on each variable are allowed to have different strengths.

Causal assumptions underlying mediation analyses

In this section, we use the causal diagram approach to discuss fundamental causal assumptions of mediation analyses. In a randomized experiment, the causal effects of the treatment, such as $X \rightarrow M_1$, $X \rightarrow M_2$, and $X \rightarrow Y$ in Figure 1, can be well-justified using the study design (by randomizing treatment) and the temporal order (by measuring mediator(s) and outcome after implementing treatment). Below we focus on causal paths that require rigorous conceptual and methodological justifications: mediator-outcome, and

mediator-mediator, causal relations. To better illustrate the issues addressed and demonstrate their relevance, below we will use a hypothetical experiment from social psychology as a running example. In this hypothetical experiment, researchers are investigating how social exclusion (randomized treatment) increases social withdrawal (outcome) via negative views of the self (poor self-esteem), and negative views of others (cynicism), as two possible mediators.

Mediator-outcome causal relations

Because mediators are typically not experimentally manipulated, the mediator-outcome causal relations (e.g., $M_1 \rightarrow Y$, and $M_2 \rightarrow Y$, in Figure 1) need to be carefully justified. An empirical (statistical) association between a mediator and the outcome could just as likely resulted from e.g., reverse causality (where Y actually causes the mediator), or non-causal correlations (where Y and the mediator share an unmeasured confounder). Crucially, a mediator-outcome causal relation cannot be empirically verified from the data alone. The practice of reversing the direction of the arrow between the mediator and the outcome, and comparing the empirical associations before and after such post-hoc reversals provides no support for either causal structure (Lemmer & Gollwitzer, 2017; Thoemmes, 2015). Significance-based tests based on empirical correlations of separate “components” of the indirect effect (Yzerbyt et al., 2018) alone cannot reveal whether each correlation was either driven by a true causal effect, or borne of their mutual association with an unmeasured confounder.

Continuing with our example above, existing literature must provide theoretical or empirical support that intervening on each posited mediator (self-esteem and cynicism) can potentially lead to a change in the outcome of social withdrawal. It is important to note that the causal effect of each mediator on the outcome is of an unknown magnitude and may even be absent empirically. In fact, a central component of mediation analysis is to statistically assess the extent to which the mediator causally affects the outcome. Hence, it

is paramount to first establish causality between each posited mediator (self-esteem and cynicism) and outcome (social withdrawal) before empirically estimating the indirect effects. In a study where the possibility of reverse causation, e.g., either self-esteem or cynicism (mediators) being causally affected by social withdrawal (outcome), cannot be plausibly ruled out, the subsequent mediation analysis becomes logically fallible and is rendered meaningless.

Mediation analyses (implicitly) endow the mediator-outcome relation with a causal interpretation that necessitates thoughtful consideration of confounding.³ We describe two possible scenarios where unmeasured (or “hidden”) confounding of the mediator-outcome causal relation can arise. One straightforward possibility is when the mediator and outcome are merely different manifestations of the same underlying latent construct, thus generating a non-causal empirical association between the mediator and outcome (Fiedler et al., 2011). The other more complex possibility is illustrated using a single mediator (M) as depicted in Figure 2(a). Let M_0 and Y_0 denote pretest (or pre-treatment) measurements of the mediator and outcome respectively that are taken before implementing the treatment and thus unaffected by treatment. Both M_0 and Y_0 are confounders of the $M \rightarrow Y$ causal relation (as indicated by arrows pointing to both M and Y from each variable). Failing to adjust for both M_0 and Y_0 induces (statistical) bias in the effect estimate due to confounding. In fact, “two-wave mediation models” (Cole & Maxwell, 2003; MacKinnon, 2008; Valente & MacKinnon, 2017) commonly used to assess mediation in pretest-posttest randomized studies already condition on pretest measurements of the mediator and outcome. However, M_0 and Y_0 need not causally influence both M and Y directly to generate confounding. Consider the causal diagram in Figure 2(b) adapted from Figure 2 of Mayer et al. (2014). Here, M_0 affects M only (and not Y directly), and Y_0

³The causal assumption that there be no unmeasured confounding of the mediator-outcome relation has been formally stated in the literature, and is a component of what are often termed the “sequential ignorability” assumptions (Imai et al., 2010; Pearl, 2014; VanderWeele & Vansteelandt, 2009).

184 affects Y only (and not M directly). Because M_0 and Y_0 are correlated due to the
 185 unmeasured confounder U , there is a path linking M and Y via U ; i.e.,
 186 $M \leftarrow M_0 \leftarrow U \rightarrow Y_0 \rightarrow Y$. This path generates a non-causal association between M and Y
 187 that can be eliminated by adjusting for either M_0 , or Y_0 , or both.

Figure 2. Causal diagrams for a two-wave setting with a single mediator (M). (a) The pretest measurements of the mediator (M_0) and outcome (Y_0) each causally influence the posttest measurements of both the mediator and outcome, or (b) The pretest measurements M_0 and Y_0 are correlated because they share an unmeasured confounder U . Edges that differ between the two settings are drawn in red. Rectangular nodes denote observed variables, while round nodes denote unmeasured (or “hidden”) variables. For visual clarity, common causes C are represented by a single node, and arrows emanating from C are drawn in gray, although the effects of C on each variable are allowed to have different strengths.

188 Mediator-mediator causal relations

189 In addition to mediator-outcome causal relations, researchers need to carefully
 190 consider the causal structure among the mediators when multiple possible mediators are
 191 present. Two common approaches are parallel and serial mediation models (Rucker et al.,
 192 2011; Hayes, 2018): see e.g., Catapano & Tormala (2021), Jordan & Rand (2020), and

Krems et al. (2021) for recent use of parallel mediation models; and see e.g., Bai et al. (2021), and Schumpe et al. (2020) for recent use of serial mediation models. Notwithstanding the prevalence of these models, mediated effects defined using these conventional approaches focus on specific assumed causal paths. Hence, their interpretation and validity are predicated on stringent causal assumptions about the asserted (lack of) mediator-mediator causal relations as we discuss next.

Parallel mediation models. In a parallel mediation model, the mediators are a priori assumed not to causally affect each other, such as in Figures 1(a) and 1(c). Under such a model, the indirect effect via each mediator is interpreted to be along the path from X to Y that intersects only the mediator in question (MacKinnon, 2000; Preacher & Hayes, 2008). The absence of posited causal effects between the mediators in a parallel mediation model *prima facie* implies a stringent assumption that there are no underlying causal effects between the mediators. Such narrowly defined interpretations of the indirect effects therefore rule out any (underlying) causal effects among the mediators. Nevertheless, (conceptually distinct) mediators may be empirically correlated with each other, not only because they may be manifestations of an unknown latent variable or process, or because they might share hidden common causes, but because there are unknown causal mechanisms flowing among them.

Continuing with our example above, to estimate a parallel mediation model, researchers need to assume that there is no underlying causal effect of self-esteem on cynicism, or vice versa. Under such an assumption, the indirect effect via self-esteem is the effect of treatment on the outcome transmitted along the causal path intersecting self-esteem alone; i.e., social exclusion $\rightarrow$ self-esteem $\rightarrow$ social withdrawal. Similarly, the indirect effect via cynicism corresponds to the causal path intersecting cynicism alone; i.e., social exclusion $\rightarrow$ cynicism $\rightarrow$ social withdrawal. But the assumption that the mediators are causally independent of each other may not hold in reality. For example, one of the mediators, self-esteem, may have affected the other mediator, cynicism, in the underlying

(causal) data-generating process. When there are (underlying but unknown) causal effects among the mediators, the separate indirect effects defined in parallel mediation models no longer accurately represent the true causal effects traversing the connected mediators.

Serial mediation models. Alternatively, researchers may consider a serial mediation model, where causal effects between the mediators are assumed, such as in Figures 1(b) and 1(d). The causally ordered mediators allow for “compound” paths from treatment to outcome that traverse several (causally) linked mediators. The effect of treatment on the outcome transmitted along any directed path passing through at least one of the mediators can then be separately assessed, offering a variety of mediated effects corresponding to different paths. But valid causal inference of mediated effects under serial mediation models is predicated on (implicit) stringent causal assumptions.

First, the mediated effects are well-defined exclusively when the posited (directions of the) causal effects among the mediators are correct. The causal structure cannot be uncovered using only empirical associations between the mediators. Competing (mediation) models can codify conflicting causal relations (e.g., Figures 1(b), 1(c), and 1(d)) that represent divergent theories about the mediators, but they are nonetheless compatible with the same joint (probability) distribution of the observed variables. Furthermore, when the models are saturated (with zero degrees of freedom), they will have identical goodness-of-fit measures (MacCallum et al., 1993), and be statistically indistinguishable using routine model selection methods that rely exclusively on empirical associations. Continuing our example above, suppose that one of the mediators, self-esteem, was incorrectly specified to affect the other mediator, cynicism, when the converse was true in the underlying causal process; i.e., cynicism affected self-esteem instead. Then the mediated effect along the (incorrectly) asserted compound path, social exclusion $\rightarrow$ self-esteem $\rightarrow$ cynicism $\rightarrow$ social withdrawal, does not comport with the true mediated effect via both connected mediators. In other words, mediated effects defined in serial mediation models are accurate and relevant only when the causal ordering of the

mediators is correctly specified; otherwise they have no meaningful causal interpretation.

Second, even when the underlying causal relations among the mediators are correctly assumed, (statistically) unbiased estimation of each individual mediator-mediator causal effect demands adjusting for all confounders. For example, consider the causal effect $M_1 \rightarrow M_2$ in Figure 1(b), or the causal effect $M_2 \rightarrow M_1$ in Figure 1(d). If both mediators concurrently share a confounder, e.g., $M_1 \leftarrow U \rightarrow M_2$, failing to adjust for U will generate a spurious correlation between the mediators that induces (statistical) bias in the mediator-mediator causal effect estimate. The bias will propagate to estimates of any mediated effect which includes that causal path, such as $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$, or $X \rightarrow M_2 \rightarrow M_1 \rightarrow Y$, respectively. Continuing our example above, suppose that one of the mediators, self-esteem, was correctly specified to affect the other mediator, cynicism, in a serial mediation model. Unmeasured common causes of these two mediators, such as negative childhood experiences, can nonetheless generate spurious correlations between the mediators. Statistically unbiased estimation of the path coefficient for the self-esteem $\rightarrow$ cynicism path, and subsequently the mediated effect that includes this causal path, demands measuring and adjusting for all confounders of both mediators.

Third, a further complication arises when mediators are (correctly) causally connected to each other, so that one mediator adopts the role of a so-called *post-treatment* (or *treatment-induced*) confounder of another mediator-outcome relation. For example, consider the causal diagram in Figure 1(b). The variable M_1 is a confounder of the $M_2 \rightarrow Y$ relation that is concurrently affected by treatment (as depicted by the $X \rightarrow M_1$ arrow). The conflicting roles that M_1 plays along the different paths $X \rightarrow M_1 \rightarrow Y$ (where it is a mediator) and $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ (where it is a confounder) preclude statistically unbiased estimation of separate indirect effects along compound paths traversing the mediators, without strong (empirically untestable) parametric assumptions about the joint distribution of the mediators (Avin et al., 2005; Imai & Yamamoto, 2013; Shpitser, 2013; Andrews & Didelez, 2021). Post-treatment confounding can arise in studies with partial

compliance, also termed “noncompliance,” where participants who select not to adhere to the assigned treatment (control) can be considered to have experienced the control (treatment) condition. Compliance to treatment (M_1) is a post-treatment confounder because it is affected by treatment and concurrently influences both the mediator (M_2) and outcome (Y) of interest. Adjusting for post-treatment confounding is complicated, and valid causal inference may be limited to only the direct effect (Coffman & Zhong, 2012; De Stavola et al., 2014; Kisbu-Sakarya et al., 2020; Loh et al., 2020; Moerkerke et al., 2015; Tchetgen Tchetgen & VanderWeele, 2014), or certain definitions of indirect effects that combine path-specific effects along different (correctly) determined causal pathways (Daniel et al., 2015; Imai & Yamamoto, 2013; Steen et al., 2017a; Zhou & Yamamoto, 2020).

To summarize, defensibly justifying a determined (lack of) causal relations among the mediators requires strong theoretical knowledge or clear experimental evidence, which are often unavailable in the existing literature. Moreover, even when the mediators’ causal structure can be correctly specified based on existing literature, unbiased estimation of mediated effects traversing causally linked mediators demands measuring all possible confounders of the mediators. Recording all possible confounders is frequently impractical or unfeasible, e.g., when mediators are different facets of a shared latent construct. Therefore, researchers are put in a position to carry out a mediation analysis using conventional mediation methods with high risks of invalid inference and erroneous conclusions. In the next section, we introduce a solution for this conundrum.

Focusing on mediator-specific indirect effects in mediation analyses with multiple mediators

In this section, we introduce an approach for valid causal inference of mediation analysis with multiple mediators that eschews (correctly) specifying the mediators’ causal structure. We encourage researchers to consider assessing indirect effects specific to each distinct mediator, instead of all possible fine-grained compound paths. Our

recommendation is based on a framework of so-called “interventional indirect effects” from the causal inference and epidemiology literature (Didelez et al., 2006; Geneletti, 2007; Lok & Bosch, 2021; VanderWeele et al., 2014) that has recently been extended to multiple mediators (Loh et al., 2021; Moreno-Betancur & Carlin, 2018; Vansteelandt & Daniel, 2017). Indirect effects defined using this framework have the benefit of being well-defined and possessing the same scientifically meaningful causal interpretations, even when there are underlying causal dependencies among the mediators, or the mediators share hidden confounders, or both. Moreover, the causal interpretations are not dependent on (or implied by) the statistical model used for estimation. In fact, estimators of interventional direct and indirect effects under a certain class of linear (and additive) regression models are mathematically equivalent to estimates from conventional parallel mediation models (MacKinnon, 2000; Preacher & Hayes, 2008). Researchers need only replace the conventional (path-focused) interpretation of the estimates with the introduced (mediator-focused) interpretation of interventional effects to be endowed with attractive causal properties.

To illustrate, recall the social exclusion experiment above with self-esteem and cynicism as potential mediators. Researchers adopting the interventional effects framework may estimate the indirect effects using a parallel mediation model (without assuming that the mediators are causally independent). The *product-of-coefficients* (Alwin & Hauser, 1975; MacKinnon et al., 2002) estimators of the indirect effect via each mediator (e.g., self-esteem) is the product of (i) the coefficient for the treatment effect on that mediator (the so-called “*a*” path) and (ii) the main effect of that mediator on the outcome (the so-called “*b*” path). As previously described, conventional methods using a parallel mediation model strictly interpret this indirect effect as the effect of treatment (social exclusion) on the specific mediator (self-esteem) alone, and then the mediator to outcome (social withdrawal). No accommodation is permitted for (underlying) causal effects among the mediators under such a narrow definition. In contrast, the interventional indirect effect

should be interpreted as the combined effect along all underlying – and likely unknown – causal paths linking treatment (social exclusion) to the specific mediator (self-esteem), and then the mediator to outcome (social withdrawal). For example, suppose that cynicism affected self-esteem in the underlying (and unknown) data-generating process. The interventional indirect effect via self-esteem would then be the combined effect along the causal paths: (i) social exclusion $\rightarrow$ self-esteem $\rightarrow$ social withdrawal, and (ii) social exclusion $\rightarrow$ cynicism $\rightarrow$ self-esteem $\rightarrow$ social withdrawal. Then the “ a ” path for the indirect effect via self-esteem describes the “overall” effect of social exclusion (treatment) on social withdrawal, including the effect via the other mediator (cynicism)⁴. This causal interpretation of the interventional indirect effect is maintained regardless of how self-esteem may be causally (un)connected to the other mediators, despite using a parallel mediation model that is simply a statistical tool for estimation. A detailed illustration using regression formulae for the simplest possible example with two mediators M_1 and M_2 as depicted in Figure 1 is provided in the Supplemental Materials. Finally, we provide an example for interpreting interventional indirect effects as follows.

As preregistered, there was insufficient evidence in the literature to justify a causal structure among the possible mediators (self-esteem and cynicism). We therefore tested the interventional indirect effects (Loh et al., 2021) of social exclusion on social withdrawal via the two possible mediators using a parallel mediation model. For self-esteem, the overall effect of social exclusion on self-esteem, not limited to this specific causal path, was $b = \dots, SE = \dots, 95\%CI = \dots$; the effect of self-esteem on social withdrawal, after adjusting for possible confounders, was $b = \dots, SE = \dots, 95\%CI = \dots$. Hence the indirect effect of social exclusion on social withdrawal via self-esteem was $b = \dots, SE = \dots, 95\%CI = \dots$. For cynicism, the overall effect of social

⁴We elected not to use the label “total” effect of treatment on each mediator to avoid potential confusion with the total effect of treatment on outcome.

exclusion on cynicism, not limited to this specific causal path, was
 $b = \dots, SE = \dots, 95\%CI = \dots$; the effect of cynicism on social withdrawal,
after adjusting for possible confounders, was $b = \dots, SE = \dots, 95\%CI = \dots$.
Hence the indirect effect of social exclusion on social withdrawal via cynicism
was $b = \dots, SE = \dots, 95\%CI = \dots$.

It is important to note that we are not advocating researchers rule out serial
mediation models. We acknowledge that interventional indirect effects reveal less
information about causal mechanisms than mediated effects along directed (compound)
paths in serial mediation models that correctly represent the underlying data-generating
causal structure. Valid inference of causal effects using causal diagrams that carefully and
accurately represent theoretical models and data-generating processes is invaluable toward
advancing knowledge (Grosz et al., 2020). The abundant compound paths linking
treatment with outcome via multiple linked mediators are an attractive feature of serial
mediation models because of the variety of mediated effects that they generate. However,
the promise of using serial mediation models to elevate understanding of causal
mechanisms is secured only when the causal assumptions are met. Therefore, the benefits
of narrowly focusing on particular mediated effects traversing several mediators need to be
carefully weighed against their susceptibility to misleading conclusions when the causal
assumptions fail to hold. Any plausible ambiguity that admits the possibility of either
switching the direction of an arrow along a directed path between the treatment and
outcome, or misconstruing an empirical association due to unmeasured confounding as a
causal effect between two variables on that path, would render its causal interpretation
meaningless. In summary, we believe researchers investigating mediational processes will
benefit from a broader and more robust assessment of indirect effects via individual
mediators using interventional indirect effects.

Figure 3. Proposed guidelines for improving causal inferences of mediation analysis in randomized experiments.

Proposed guidelines to improve causal inference of mediation analysis

In this section, we propose a set of practical guidelines for researchers to improve the quality and validity of causal inference from mediation analysis in randomized experiments. As discussed in the preceding sections, the validity of substantive conclusions drawn using mediation analysis are predicated on causal assumptions being met. Therefore, we suggest that researchers carefully evaluate – preferably at the initial stages of a research project – whether mediation analysis is appropriate before data collection or empirical analysis. Mediation analysis is inappropriate when causality between the mediator(s) and outcome cannot be justified or the relevant confounders cannot be recorded. Hence, mediation analysis should not be used for routine exploratory or post-hoc data analysis to uncover empirical correlation patterns without attention to the causal assumptions. When mediation analysis is deemed to be appropriate, we propose the following steps in Figure 3 as a guide to help researchers plan their mediation analysis towards strengthening causal inferences from mediation analysis in randomized experiments.

Preregister the mediators and indirect effects, and justify the posited causal relations

We propose that researchers declare possible mediators of the treatment and outcome of interest prior to data collection, as part of study preregistration. Moreover, we recommend visualizing the causal relations between the treatment, mediator(s), and outcome using a causal diagram. Such an exercise encourages explicit evaluation and judicious justification of the assumed causal relations, thereby reducing the risks of reverse causality. Delineating the separate roles of stating causal assumptions and conducting statistical estimation can potentially avoid (ad-hoc and) post-hoc attribution of possibly spurious correlations as causations (e.g., in an exploratory analysis), and improve research practice.

The presence (or absence) of each causal effect should be clearly justified using

well-established subject matter theory, previous empirical research, or thoughtful study designs. For example, the treatment-mediator(s) and treatment-outcome causal relations are well-justified using the study design (by randomizing treatment) and the temporal order (by measuring mediator(s) and outcome after implementing treatment). For each mediator-outcome causal relation, longitudinal study designs where the mediator is measured at an earlier occasion than the outcome can offer temporal justification to rule out reverse causation. Nonetheless, temporal precedence may not suffice to establish causality because the mediator and outcome may be correlated only due to their mutual associations with an unmeasured pre-treatment common cause, such as a latent construct. Establishing the mediator-outcome causal relation would thus require further reasoning based on substantive theory. Another possibility is to experimentally manipulate the mediator (if feasible) to lend support for the posited mediator-outcome causal relation (Imai & Yamamoto, 2013; MacKinnon & Pirlott, 2015; Pirlott & MacKinnon, 2016; Spencer et al., 2005). However, a challenge of such experimental approaches is that it can be complicated to manipulate mediators that are psychological processes. When there are multiple mediators, the conceptual and methodological justifications of the mediator-outcome causal relation for a single mediator must then be extended to each of the possible mediators in turn.

When interest is in further assessing mediated effects along compound paths connecting several mediators in serial mediation models, each separate causal effect between the mediators must be similarly meticulously justified. Researchers should therefore seek to establish a theoretically sound and defensible causal structure for all the directed paths leading from treatment to the outcome that possibly intersect the mediator(s). Moreover, the either path- or mediator-specific indirect effects of interest should be included in the preregistration. For example, when using a serial mediation model, the compound paths linking the treatment and outcome via different connected mediators should be stated. Unless researchers can (i) correctly specify the mediators'

causal structure with certainty, and (ii) measure all pre-treatment confounders of the mediators when using serial mediation models, we recommend adopting the interventional effects framework instead. When there is inadequate prevailing knowledge to unambiguously assert the (lack of) mediator-mediator causal relations, researchers may simply leave the mediators' causal structure unspecified. It can then be stated that interest is in investigating mediator-specific interventional indirect effects.

Continuing with our example above, where the effect of social exclusion (treatment) on social withdrawal (outcome) is possibly mediated by self-esteem and cynicism (mediators), we first visualize the assumed causal relations using a causal diagram. We provide an example in Figure 4. Next, we carefully consider each individual causal effect that contributes to the causal interpretations of the mediation analysis: whether the causal effect can be justified using available knowledge, and how to design the study to strengthen the justification (such as temporal precedence). These considerations can be formalized as part of study preregistration. We offer an example in Table 1 of how to present the justifications using relevant theory, existing empirical evidence, and current study design. Readers and reviewers are then empowered to judge for themselves whether the causal assumptions can be plausibly met based on the arguments put forth. Because interventional indirect effects are the focus of the mediation analysis in this example, only the mediator-outcome causal relations need to be justified. Had we considered a serial mediation model instead, the causal effect between self-esteem and cynicism would have to be justified as well.

Figure 4. Example of a causal diagram in a mediation analysis focusing on interventional indirect effects, using the hypothetical social psychology experiment. There may be underlying causal effects between the mediators, or correlations among the mediators, notwithstanding the absence of arrows between the mediators. For visual simplicity, common causes C are represented by a single node, and arrows emanating from C are drawn in gray, although the effects of C on each variable are allowed to have different strengths.

Causal path	Justification	Details
Treatment --> Mediator		
Social exclusion --> Self-esteem	Temporal precedence	Self-esteem is measured at a later time after social exclusion.
	Possible confounders	None due to randomized assignment
Social exclusion --> Cynicism	Temporal precedence	Cynicism is measured at a later time after social exclusion.
	Possible confounders	None due to randomized assignment
Treatment --> Outcome		
Social exclusion --> Social withdrawal	Temporal precedence	Social withdrawal is measured at a later time after social exclusion.
	Possible confounders	None due to randomized assignment
Mediator --> Outcome		
Self-esteem --> Social withdrawal	Existing literature	Poor self-esteem implies one has low relational value to be included in relationships and groups (Leary & Baumeister, 2000). As a coping strategy, people may move away from social situations (van Kleef et al., 2010). These theoretical frameworks suggest that negative self esteem may give rise to social withdrawal.
	Temporal precedence	The mediator-outcome link can be strengthened by introducing a temporal ordering in the study design to ensure that the psychological processes represented by the mediators occur temporally before the outcome of social withdrawal. For example, the study may contain two parts. In part I, participants receive a randomized treatment of social exclusion, and then complete measures assessing state self-esteem and cynicism. In part II, participants are provided with an opportunity to complete a task alone or with other participants. Participants' preference for the subsequent solitary task (1=not at all, 5=extremely) will then be recorded as an indicator of their social withdrawal tendencies.
	Possible confounders	Pre-treatment measure of the mediator (trait level self-esteem), pre-treatment measure of the outcome (general social withdrawal tendencies), and other possible confounders should be measured. For example, past work has shown that lonely people tend to have low self esteem and withdraw socially (Ernst & Cacioppo, 1999), suggesting that chronic loneliness can be a common cause of low self-esteem and social withdrawal and thus should be measured and adjusted for.
Cynicism --> Social withdrawal	Existing literature	Cynical people tend to disrespect others (Stavrova et al., 2020), neglect partners' needs (Kaplan et al., 2004), and provide less social support (Hart, 1999). These findings show that cynicism contributes to a low interest in forming connections with others, providing support for the causal effect of cynicism on social withdrawal.
	Temporal precedence	see "Temporal precedence" above.
	Possible confounders	Pre-treatment measure of the mediator (trait cynicism), pre-treatment measure of the outcome (general social withdrawal tendencies), and other possible confounders should be measured. For example, past work has shown that experiencing disrespectful treatment leads to cynicism (Stavrova et al., 2020); and experiencing disrespectful treatment may motivate people to withdraw to avoid similar experience (Richman & Leary, 2009). Therefore, past experience with disrespectful treatment can be a common cause of cynicism and social withdrawal and thus should be measured and adjusted for.
Mediator --> Mediator		
Self-esteem and cynicism	Existing literature	No definitive causal effects between the mediators can be established. Interventional indirect effects will be estimated.

Note. The above justifications are used only to illustrate the guidelines for a mediation analysis in a hypothetical example. In actual practice, a more thorough literature review, and careful selection of confounders, are recommended.

Table 1

Example of justifications for, and possible confounders of, the posited causal effects in a mediation analysis, using the hypothetical social psychology experiment.

Recording pertinent pre-treatment confounders of the mediators and outcome

We encourage researchers to carefully consider relevant confounders of each mediator and the outcome. Researchers interested in mediated effects along compound paths should apply the same thoughtful consideration of confounders to each individual causal effect between the mediators. A rich selection of confounders can be put together by including relevant covariates based on existing theoretical knowledge and external empirical information, or in discussion with subject matter experts. Following Steyer et al. (2010), we recommend that researchers record pre-treatment measures of the mediators and outcome whenever possible, even in a randomized experiment without a pretest-posttest study design, for confounding adjustment. Confounders can be included in the causal diagram, along with their rationale for confounding adjustment, as part of study preregistration.

Continuing with our example above, possible pre-treatment confounders of each of self-esteem and cynicism (mediators), and social withdrawal (outcome), are included in the causal diagram of Figure 4. For each posited causal effect, the pre-treatment confounders to be recorded for confounding adjustment are stated in Table 1. Because interventional indirect effects are the focus of the mediation analysis in this example, confounders need only be considered for each mediator-outcome causal relation. Had we considered a serial mediation model instead, confounders of self-esteem and cynicism would have to be included as well.

We further suggest researchers measure possible confounders prior to treatment implementation when possible so that they may reasonably be considered to be causally unaffected by treatment. Confounders that are measured subsequent to treatment may be causally influenced by treatment, resulting in post-treatment confounding that complicates assessment of the indirect effects as described above. The complications can be further exacerbated if the putative confounders are only recognized during data analysis, and the causal structure between these post-treatment variables and the mediators cannot be rigorously justified. When post-treatment confounding is (a priori known to be) inevitable,

and interest is in investigating treatment effects transmitted via all possible intervening variables, we recommend adopting the perspective that all post-treatment confounders of the mediator(s)-outcome relation(s) be considered simply as possible mediators. Under such settings, researchers can simply leave the causal structure among the mediators unspecified, and adopt the interventional indirect effects framework.

Estimation and reporting of mediation analysis results

Estimation of the preregistered indirect (and direct) effects via the possible mediators can proceed after data has been collected for analysis. Prevailing estimation using path analysis within the SEM framework proceeds by fitting an assumed linear path analysis model (or set of recursive linear regression models) for the outcome and the mediator(s) to the observed data. These methods extend the Baron & Kenny (1986) approach for a single mediator to multiple mediators; see e.g., MacKinnon (2008) and Hayes (2018) for book-length presentations and the detailed references therein.

Researchers are advised to closely follow the preregistered mediation analysis. A scenario that may prompt deviation from the planned analysis is when the mediator residuals are empirically correlated. In extreme settings, the resulting multicollinearity in the outcome regression model can yield highly variable coefficient estimates of the so-called “*b*” paths. While a well-known remedy for multicollinearity is to drop one of the correlated predictors from the regression model, we advise caution when removing any mediators from the outcome model. Doing so may inadvertently admit hidden confounding of the remaining mediator(s) and outcome, especially when the dropped mediator is associated with both the remaining mediator(s) and outcome. The resulting statistical biases of the estimated indirect effect(s) via the remaining mediator(s) can potentially exceed the increased variability due to multicollinearity.

Confidence intervals (and not just p-values from null hypothesis tests) for each individual causal path, and each indirect effect, can be reported. Confidence intervals may

be constructed using either a Monte Carlo (or parametric bootstrap) approach (MacKinnon et al., 2002; Preacher & Selig, 2012) that resamples parameter values from the fitted models using their estimated sampling distribution from the observed data, or a nonparametric bootstrap (Efron & Tibshirani, 1994) that resamples observations with replacement, then refits the models to obtain parameter estimates. Readers interested in alternative inferential procedures are referred to the comparison studies conducted by Biesanz et al. (2010), Tofighi & Kelley (2020), Valente et al. (2016), and Yzerbyt et al. (2018).

When there are non-continuous mediators, or outcome, or both, the product-of-coefficients method may not result in a valid decomposition of the direct and indirect effects, due to misspecification of non-linear (e.g., logistic regression) models for the mediators or outcome; see MacKinnon et al. (2020) for the single mediator setting. Under such settings, we recommend researchers consider more general estimation methods that build on the counterfactual-based “causal mediation” framework developed using model-free definitions of *natural direct and indirect effects* (Pearl, 2001; Robins & Greenland, 1992). These methods readily permit flexible estimation procedures with linear and non-linear models, including treatment-mediator(-mediator) interactions; and allow for different requirements for models for the treatment and the mediators, or the mediators’ (joint) distribution, or mean models for the outcome, for valid inference. A review of the causal mediation analysis framework is beyond the scope of this article. Instead, we refer readers to Nguyen et al. (2021) for an accessible introduction and the references therein, and to VanderWeele (2015) for a book-length treatment. Moreover, available R (R Core Team, 2021) packages implementing methods built on this framework include **CMAverse** (Shi et al., 2021), **medflex** (Steen et al., 2017b), **mediation** (Tingley et al., 2014), and **paths** (Yamamoto & Zhou, 2021). Valente et al. (2020) compared software implementing causal mediation estimation procedures across various computing environments.

Sensitivity analyses to possible unmeasured confounding

Despite researchers' best efforts to adjust for confounding of the mediator(s)-outcome relation(s), there may plausibly remain hidden confounders, because they were impractical or unfeasible to measure, but can potentially invalidate causal inferences. It is therefore judicious to carry out a sensitivity analysis to hidden confounding of the mediator(s)-outcome relation(s). Methods for a single mediator include Cox et al. (2013), Fritz et al. (2016), Hong et al. (2018), Liu & Wang (2020), and Smith & VanderWeele (2019); methods for multiple mediators have recently been developed by Park & Esterling (2021) and Zhou & Yamamoto (2020). These methods essentially posit the impact of an additional hypothetical unmeasured (pre-treatment) confounder on a mediator and outcome, and systematically determine the (minimum) degree to which statistical inference would differ after adjusting for the hypothetical confounder.

Limitations

There are several issues we did not discuss in this article. First, we focused on randomized experiments and did not discuss mediation analysis using non-experimental data. Observational studies in which participants can select the treatment they are exposed to require (i) additional justifications for the posited treatment-mediator and treatment-outcome causal relations, and (ii) ensuring that confounders of these relations have been sufficiently adjusted for toward satisfying the respective assumptions of no unmeasured confounding. Second, we have assumed that all variables are measured without error. When the mediators (and outcome) are latent constructs, measurement error models for the latent mediators (and outcome) have to be incorporated into the mediation analysis (Gonzalez & MacKinnon, 2021). Such settings may require further scrutiny of whether measurement invariance holds between treatment groups and across time (Georgeson et al., 2021). Finally, we limited consideration to single level data (i.e., no nesting data structure) where each variable is measured at a single instance, although the

mediator(s) and outcome may be measured at different instances.

Conclusion

Mediation analysis is an important research tool for understanding the underlying causal mechanisms by which an intervention influences an outcome. But failure to scrutinize the scientific plausibility of the stringent causal assumptions can lead to misconstruing spurious associations as causal effects and misuse of mediation analysis. Our article offers a systematic presentation of the causal assumptions underpinning mediation analysis, and a step-by-step guide with concrete suggestions for researchers planning to conduct mediation analyses when designing randomized experiments. With this article, we hope to encourage explication, justification, and reflection of these causal assumptions to improve the research practice of mediation analysis in psychology research.

References

- Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 40(1), 37–47. doi: 10.2307/2094445
- Andrews, R. M., & Didelez, V. (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, 32(2).
- Avin, C., Shpitser, I., & Pearl, J. (2005, July). Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 357–363). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Bai, Y., Ocampo, J., Jin, G., Chen, S., Benet-Martinez, V., Monroy, M., . . . Keltner, D. (2021). Awe, daily stress, and elevated life satisfaction. *Journal of Personality and Social Psychology*, 120(4), 837–860. doi: 10.1037/pspa0000267
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173. doi: 10.1037/0022-3514.51.6.1173
- Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, 45(4), 661–701. Retrieved from <https://doi.org/10.1080/00273171.2010.498292> (PMID: 26735714) doi: 10.1080/00273171.2010.498292
- Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research* (pp. 301–328). Springer.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what’s the mechanism? (don’t expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558. doi: 10.1037/a0018933

- 590 Catapano, R., & Tormala, Z. L. (2021). Do i support that it's good or oppose that it's
591 bad? the effect of support-oppose framing on attitude sharing. *Journal of Personality
592 and Social Psychology*, 121(1), 23–42. doi: 10.1037/pspa0000266
- 593 Coffman, D. L., & Zhong, W. (2012). Assessing mediation using marginal structural
594 models in the presence of confounding and moderation. *Psychological Methods*, 17(4),
595 642. doi: 10.1037/a0029311
- 596 Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data:
597 Questions and tips in the use of structural equation modeling. *Journal of Abnormal
598 Psychology*, 112(4), 558–577. doi: 10.1037/0021-843X.112.4.558
- 599 Costello, C. K., & Srivastava, S. (2021). Perceiving personality through the grapevine: A
600 network approach to reputations. *Journal of Personality and Social Psychology*, 121(1),
601 151–167. doi: 10.1037/pspp0000362
- 602 Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013, Oct). Sensitivity
603 plots for confounder bias in the single mediator model. *Eval Rev*, 37(5), 405–431. doi:
604 10.1177/0193841X14524576
- 605 Cuartas, J., & McCoy, D. C. (2021). Causal mediation in developmental science: A primer.
606 *International Journal of Behavioral Development*, 45(3), 269–274. Retrieved from
607 <https://doi.org/10.1177/0165025420981640> doi:
608 10.1177/0165025420981640
- 609 Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal
610 mediation analysis with multiple mediators. *Biometrics*, 71(1), 1–14. doi:
611 10.1111/biom.12248
- 612 De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., & Micali, N. (2014). Mediation analysis
613 with intermediate confounding: Structural Equation Modeling viewed through the causal

inference lens. *American Journal of Epidemiology*, 181(1), 64–80. doi:

10.1093/aje/kwu239

Didelez, V., Dawid, A. P., & Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd conference on uncertainty in artificial intelligence* (pp. 138–146). Arlington, VA, USA: AUAI Press.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 72(1), 1–16. doi: 10.1086/224256

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC. doi: 10.1007/978-1-4899-4541-9

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-6094-3_13

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests—an analysis of articles published in 2015. *Journal of Experimental Social Psychology*, 75, 95–102. doi: 10.1016/j.jesp.2017.11.008

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, 47(6), 1231–1236. doi: 10.1016/j.jesp.2011.05.007

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

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 199–215. doi: 10.1111/j.1467-9868.2007.00584.x

- Georgeson, A. R., Valente, M. J., & Gonzalez, O. (2021). Evaluating response shift in statistical mediation analysis. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211012271. doi: 10.1177/25152459211012271
- Gonzalez, O., & MacKinnon, D. P. (2021). The measurement of the mediator and its influence on statistical mediation conclusions. *Psychological Methods*, 26(1), 1–17. doi: 10.1037/met0000263
- Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255. (PMID: 32727292) doi: 10.1177/1745691620921521
- Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., . . . Boadu, K. (2003). Pearl’s D-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(2), 289–311. doi: 10.1207/s15328007sem1002_8
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach* (2nd ed.). New York, NY, USA: Guilford press.
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if* (First ed.). Boca Raton, FL, USA: Chapman & Hall \CRC.
- Hong, G., Qin, X., & Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43(1), 32–56. doi: 10.3102/1076998617749561
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309. doi: 10.1037/a0020761

- Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2), 141–171. doi: 10.1093/pan/mps040
- Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, 118(1), 57–88. doi: <http://dx.doi.org/10.1037/pspi0000186>
- Kisbu-Sakarya, Y., MacKinnon, D. P., Valente, M. J., & Çetinkaya, E. (2020). Causal mediation analysis in the presence of post-treatment confounding variables: A Monte Carlo simulation study. *Frontiers in Psychology*, 11, 2067. doi: 10.3389/fpsyg.2020.02067
- Kline, R. B. (2015). The mediation myth. *Basic and Applied Social Psychology*, 37(4), 202–213. doi: 10.1080/01973533.2015.1049349
- Krems, J. A., Williams, K. E. G., Aktipis, A., & Kenrick, D. T. (2021). Friendship jealousy: One tool for maintaining friendships in the face of third-party threats? *Journal of Personality and Social Psychology*, 120(4), 977–1012. Retrieved from <https://www.proquest.com/scholarly-journals/friendship-jealousy-one-tool-maintaining/docview/2432063142/se-2?accountid=11077> doi: <http://dx.doi.org/10.1037/pspi0000311>
- Lemmer, G., & Gollwitzer, M. (2017). The “true” indirect effect won’t (always) stand up: When and why reverse mediation testing fails. *Journal of Experimental Social Psychology*, 69, 144–149. doi: 10.1016/j.jesp.2016.05.002
- Liu, X., & Wang, L. (2020, Jul). The impact of measurement error and omitting confounders on statistical inference of mediation effects and tools for sensitivity analysis. *Psychological Methods*. doi: 10.1037/met0000345

- Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G., & Vansteelandt, S. (2020). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomised studies. *Multivariate Behavioral Research*, 55(5), 763-785. doi: 10.1080/00273171.2019.1681251
- Loh, W. W., Moerkerke, B., Loeys, T., & Vansteelandt, S. (2021). Disentangling indirect effects through multiple mediators without assuming any causal structure among the mediators. *Psychological Methods*. Retrieved from psyarxiv.com/3q4hg doi: 10.1037/met0000314
- Lok, J. J., & Bosch, R. J. (2021). Causal organic indirect and direct effects: Closer to the original approach to mediation analysis, with a product method for binary mediators. *Epidemiology*, 32(3).
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114(1), 185–199. doi: 10.1037/0033-2909.114.1.185
- MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis* (1st ed.). New York, NY, USA: Routledge. doi: 10.4324/9780203809556
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1), 83–104. doi: 10.1037/1082-989X.7.1.83

- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, 19(1), 30–43. doi: 10.1177/1088868314542878
- MacKinnon, D. P., Valente, M. J., & Gonzalez, O. (2020). The correspondence between causal and traditional mediation analysis: the link is the mediator by treatment interaction. *Prevention Science*, 21(2), 147–157. doi: 10.1007/s11121-019-01076-4
- Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, 49(5), 425–442. doi: 10.1080/00273171.2014.931797
- Moerkerke, B., Loeys, T., & Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological Methods*, 20(2), 204. doi: 10.1037/a0036368
- Moreno-Betancur, M., & Carlin, J. B. (2018). Understanding interventional effects: A more natural approach to mediation analysis? *Epidemiology*, 29(5), 614–617. doi: 10.1097/EDE.0000000000000866
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
- Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*, 26(2), 255–271. doi: 10.1037/met0000299
- Park, S., & Esterling, K. M. (2021). Sensitivity analysis for pretreatment confounding with multiple mediators. *Journal of Educational and Behavioral Statistics*, 46(1), 85–108. doi: 10.3102/1076998620934500

- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th conference on uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York, NY, USA: Cambridge University Press. doi: 10.1017/CBO9780511803161
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York, NY, USA: Guilford Press. doi: 10.21236/ADA557445
- Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19(4), 459. doi: 10.1037/a0036434
- Pearl, J., & MacKenzie, D. (2018). *The book of why: The new science of cause and effect*. New York, NY, USA: Basic Books.
- Pek, J., & Hoyle, R. H. (2016). On the (in) validity of tests of simple mediation: Threats and solutions. *Social and personality psychology compass*, 10(3), 150–163. doi: 10.1111/spc3.12237
- Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66, 29–38. (Rigorous and Replicable Methods in Social Psychology) doi: 10.1016/j.jesp.2015.09.012
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879–891. doi: 10.3758/BRM.40.3.879
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2), 77–98.

- R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2), 143–155. doi: 10.1097/00001648-199203000-00013
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42. doi: 10.1177/2515245917745629
- Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2021, Feb). *That's a lot to process! pitfalls of popular path models*. PsyArXiv. Retrieved from psyarxiv.com/paeb7 doi: 10.31234/osf.io/paeb7
- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5(6), 359-371. doi: 10.1111/j.1751-9004.2011.00355.x
- Schumpe, B. M., Bélanger, J. J., Moyano, M., & Nisa, C. F. (2020). The role of sensation seeking in political violence: An extension of the significance quest theory. *Journal of Personality and Social Psychology*, 118(4), 743–761. Retrieved from <https://www.proquest.com/scholarly-journals/role-sensation-seeking-political-violence/docview/2129028175/se-2?accountid=11077> doi: <http://dx.doi.org/10.1037/pspp0000223>
- Shi, B., Choirat, C., Coull, B. A., VanderWeele, T. J., & Valeri, L. (2021). CMAverse: a suite of functions for reproducible causal mediation analyses. *Epidemiology, Publish Ahead of Print*. Retrieved from https://journals.lww.com/epidem/Fulltext/9000/CMAverse__a_suite_of_functions_for_reproducible.98264.aspx

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6), 1011–1035. doi: 10.1111/cogs.12058

Smith, L. H., & VanderWeele, T. J. (2019). Mediation e-values: approximate sensitivity analysis for unmeasured mediator–outcome confounding. *Epidemiology*, 30(6), 835–837.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89(6), 845–851. doi: 10.1037/0022-3514.89.6.845

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017a). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186(2), 184–193. doi: 10.1093/aje/kwx051

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017b). Medflex: an r package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*, 76(11).

Steyer, R., Fiege, C., & Rose, N. (2010). Analyzing total, direct and indirect causal effects in intervention studies. *ISSBD Bulletin*, 57, 10–13.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology*, 37(4), 235–246. doi: 10.1080/01973533.2015.1062380

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

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology*, 37(4), 226-234. doi: 10.1080/01973533.2015.1049351

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5), 1–38. Retrieved from <http://www.jstatsoft.org/v59/i05/> doi: 10.18637/jss.v059.i05

Tofighi, D., & Kelley, K. (2020). Improved inference in mediation analysis: Introducing the model-based constrained optimization procedure. *Psychological Methods*, 25(4), 496–515. doi: 10.1037/met0000259

Trafimow, D. (2015). Introduction to the special issue on mediation analyses: What if planetary scientists used mediation analysis to infer causation? *Basic and Applied Social Psychology*, 37(4), 197-201. Retrieved from <https://doi.org/10.1080/01973533.2015.1064290> doi: 10.1080/01973533.2015.1064290

Valente, M. J., Gonzalez, O., Miočević, M., & MacKinnon, D. P. (2016). A note on testing mediated effects in structural equation models: Reconciling past and current research on the performance of the test of joint significance. *Educational and Psychological Measurement*, 76(6), 889-911. (PMID: 27833175) doi: 10.1177/0013164415618992

Valente, M. J., & MacKinnon, D. P. (2017). Comparing models of change to estimate the mediated effect in the pretest–posttest control group design. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(3), 428-450. doi: 10.1080/10705511.2016.1274657

Valente, M. J., Rijnhart, J. J., Smyth, H. L., Muniz, F. B., & MacKinnon, D. P. (2020). Causal mediation programs in r, mplus, sas, spss, and stata. *Structural Equation*

Modeling: A Multidisciplinary Journal, 27(6), 975-984. (PMID: 33536726) doi:
10.1080/10705511.2020.1777133

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction* (1st ed.). Oxford University Press.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4), 457-468. doi:
10.4310/SII.2009.v2.n4.a7

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

Vansteelandt, S., & Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2), 258-265. doi: 10.1097/EDE.00000000000000596

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161-215. doi: 10.1214/aoms/1177732676

Yamamoto, T., & Zhou, X. (2021). paths: An imputation approach to estimating path-specific causal effects [Computer software manual]. Retrieved from
<https://CRAN.R-project.org/package=paths> (R package version 0.1.1)

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, 115(6), 929-943. doi:
10.1037/pspa0000132

847 Zhou, X., & Yamamoto, T. (2020, Jan). Tracing causal paths from experimental and
848 observational data. Retrieved from osf.io/preprints/socarxiv/2rx6p doi:
849 10.31235/osf.io/2rx6p

Appendix

Motivation of interventional (in)direct effects

In this section, we provide an intuitive motivation for interventional (in)direct effects using the simplest possible example with two mediators M_1 and M_2 . Interventional indirect effects are most appropriate for attributing the total effect of a treatment on an outcome to each of multiple possible mediators. Crucially, (possibly arbitrary or post-hoc) assumptions about the mediators' causal structure can be avoided in the statistical models fitted to the observed data. The development follows Loh et al. (2021). Intuitively, we will focus on only the “overall” effect of treatment on each mediator. For example, each mediator can be regressed on treatment X (and covariates C) only, so that the (partial) regression coefficient of X encodes the overall treatment effect on that mediator. In other words, consider the following set of linear regression models:

$$E(M_1|X, C) = a_1^*X + a_{1C}C \quad (1)$$

$$E(M_2|X, C) = a_2^*X + a_{2C}C \quad (2)$$

$$E(Y|M_2, M_1, X, C) = cX + b_1M_1 + b_2M_2 + b_C C, \quad (3)$$

We use asterisks in the coefficients' superscripts, e.g., a_1^* and a_2^* , to emphasize that they encode the overall effects of X on M_1 and M_2 , respectively, in the assumed model for each mediator. We reiterate that (1) and (2) do not imply that M_1 and M_2 are independent; rather, by excluding all other mediators from each model, we essentially sum over all underlying effects of treatment on the mediator in question via unknown intervening variables, and any unmeasured (pre-treatment) confounding of the mediators, to obtain the overall effect of X on each mediator.⁵ Then using the product-of-coefficients method, an estimator of the interventional indirect effect via a mediator is the product of the overall effect of treatment on the mediator in question, and the coefficient of the mediator in

⁵We elected not to use the label “total” effect of treatment X on each mediator to avoid potential confusion with the total effect of treatment X on outcome Y .

question in the outcome model, e.g., (3). For example, an estimator of the interventional indirect effect via M_1 is $a_1^*b_1$, and the interventional indirect effect via M_2 is $a_2^*b_2$.

We provide an example to illustrate why valid causal interpretation of the interventional indirect effect can be maintained regardless of the (possibly unknown) causal structure among the mediators from which the observed data is generated. Suppose that M_1 truly causally influences M_2 , such as in the causal diagram of Figure 1(b), and that both mediators share a hidden common cause U . In other words, suppose that the observed data was generated as follows:

$$E(M_1|X, U, C) = \alpha_1 X + \alpha_{1U}U + \alpha_{1C}C \quad (4)$$

$$E(M_2|M_1, X, U, C) = \alpha_2 X + \alpha_{21}M_1 + \alpha_{2U}U + \alpha_{2C}C \quad (5)$$

$$E(Y|M_2, M_1, X, U, C) = \gamma X + \beta_1 M_1 + \beta_2 M_2 + \beta_C C. \quad (6)$$

We use Greek alphabet letters (e.g., α, β, γ) to denote parameters in the true (possibly unknown) data-generating models, and Latin alphabet letters (e.g., a, b, c) to denote coefficients in assumed statistical models. Then a_1^* encodes the effect of treatment on M_1 along the path $X \rightarrow M_1$ alone, whereas a_2^* encodes the combined effect along the paths $X \rightarrow M_2$ and $X \rightarrow M_1 \rightarrow M_2$. Using the product-of-coefficients method, the former is α_1 and the latter is $\alpha_2 + \alpha_1\alpha_{21}$. Under the (correctly) assumed statistical models for the mediators and outcome in (1)–(3), the coefficients are thus unbiased estimators of the corresponding (combined) causal effects; i.e., $a_1^* = \alpha_1$, and $a_2^* = \alpha_2 + \alpha_1\alpha_{21}$. Note that even in a serial mediation model that correctly posits the arrow between M_1 and M_2 , the hidden confounder U precludes statistically unbiased estimation of the causal effect of M_1 on M_2 .

Now consider instead a different data-generating process where M_2 truly causally influences M_1 , as depicted in the causal diagram of Figure 1(d). It follows that a_1^* encodes the combined effect along the paths $X \rightarrow M_1$ and $X \rightarrow M_2 \rightarrow M_1$, whereas a_2^* encodes the effect along the path $X \rightarrow M_2$ alone. Similarly, if M_1 and M_2 do not causally influence

each other as depicted in the causal diagrams of Figures 1(a) and 1(c), then a_1^* and a_2^* encode the effect along the path $X \rightarrow M_1$ and $X \rightarrow M_2$, respectively. Therefore, while the constituent effects that a_1^* and a_2^* encode depend on the underlying causal structure of the mediators, the interpretations (and estimators) of a_1^* and a_2^* in (1) and (2) in terms of only the combined causal paths are invariant to that underlying causal structure.

We acknowledge that under assumed linear regression models with only main effects for treatment and mediators, the estimators of the interventional direct and indirect effects are functionally equivalent to those from parallel mediation models. However, the absence of posited causal effects between the mediators in a parallel mediation model (implicitly) imply a causal assumption that the mediators are independent of each other. In contrast, by interpreting the so-called “ a ” path for each mediator as the overall effect of treatment on that mediator instead (i.e., as “ a^* ”), prevalent product-of-coefficient estimators assuming parallel path models are thus endowed with the same causal interpretation as interventional indirect effects. In particular, the same interpretation is maintained even when there are unknown underlying causal dependencies among the mediators. Furthermore, the indirect effect estimates are robust against hidden confounding of the mediators, such as by U in Figure 1(c), that induces correlation among the mediators unexplained by the mediators’ dependence on treatment X and baseline confounders C . This is because the coefficient estimates of the “ a ” paths in a fitted parallel mediation model remain the same regardless of whether the mediator residuals covariances are constrained to zero (MacKinnon, 2000), or freely estimated (Preacher & Hayes, 2008). The former is implied when using a set of (separate) multiple linear regression models; the latter can be implemented when using SEM estimation methods by including the mediator residuals’ covariances as parameters to be (freely) estimated. Because these parameters do not enter the expressions of the indirect effect estimators (i.e., the product of the “ a ” and “ b ” path coefficients) under the assumed model, the resulting indirect effect estimates, therefore, are equivalent under both approaches.

918 However, the indirect effect estimates remain unchanged when the mediator residuals
919 are permitted to covary only if either (i) each mediator’s effect on the outcome is
920 unmoderated by another mediator, or (ii) the covariances do not depend on treatment.
921 When the observed data are generated under scenarios where both assumptions are
922 violated, merely assuming a parallel mediation model may lead to erroneous conclusions.
923 In particular, there can be an indirect effect that cannot be attributed to any specific
924 mediator due to the mediators’ *mutual dependence on one another*. Estimating this
925 indirect effect requires statistical models that allow for the effect of each mediator on the
926 outcome to be moderated by another mediator, and considering how the mediators’
927 covariances may be affected by treatment. The introduced interventional effects framework
928 therefore provides a clear set of rules for ascribing the model coefficients for such
929 (treatment-)mediator-mediator interaction effects to the corresponding expressions for the
930 estimators, notwithstanding theory and empirical data which may suggest such effects to
931 be negligible. Details on estimators of the interventional indirect effects are provided in
932 Loh et al. (2021).