Disentangling indirect effects through multiple mediators without assuming any causal structure among the mediators

Wen Wei Loh[1], Beatrijs Moerkerke[1], Tom Loeys[1], and Stijn Vansteelandt[2,3]

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

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium

[3] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, United Kingdom

Abstract

When multiple mediators exist on the causal pathway from treatment to outcome, path analysis prevails for disentangling indirect effects along paths linking possibly several mediators. However, separately evaluating each indirect effect along different posited paths demands stringent assumptions, such as correctly specifying the mediators' causal structure, and no unobserved confounding among the mediators. These assumptions may be unfalsifiable in practice and, when they fail to hold, can result in misleading conclusions about the mediators. Nevertheless, these assumptions are avoidable when substantive interest is in inference about the indirect effects specific to each distinct mediator. In this article, we introduce a new definition of indirect effects called interventional indirect effects from the causal inference and epidemiology literature. Interventional indirect effects can be unbiasedly estimated without the assumptions above while retaining scientifically meaningful interpretations. We show that under a typical class of linear and additive mean models, estimators of interventional indirect effects adopt the same analytical form as prevalent product-of-coefficient estimators assuming a parallel mediator model. Prevalent estimators are therefore unbiased when estimating interventional indirect effects - even when there are unknown causal effects among the mediators - but require a different causal interpretation. When other mediators moderate the effect of each mediator on the outcome, and the mediators' covariance is affected by treatment, such an indirect effect due to the mediators' mutual dependence (on one another) cannot be attributed to any mediator alone. We exploit the proposed definitions of interventional indirect effects to develop novel estimators under such settings.

Keywords: Direct and indirect effects; Interventional effects; Multiple mediation analysis; Path analysis

Disentangling indirect effects through multiple mediators without assuming any causal structure among the mediators

Introduction

Mediation analysis is widely used in the behavioral, psychological and social sciences to gain insight into the extent to which the causal effect of a treatment (A) on an outcome (Y) is transmitted through intermediate variables on the causal pathway from A to Y. Consider the following social psychology example by Voelkel et al. (2019), who investigated the causal effect of a political inclusion manipulation (A) on the level of prejudice toward a political outgroup (Y). Perceived worldview dissimilarity of the political outgroup (M_1) is considered a mediator if the manipulation affects how strongly an individual regards the political outgroup as holding political or social beliefs different from her/his own, which in turn causes a change in prejudice toward that outgroup. Similarly, perceived fairness of the political outgroup (M_2) is also considered a mediator if the manipulation affects how strongly an individual regards the outgroup as being open to different opinions, which in turn causes a change in prejudice toward that outgroup. Many realistic mediation analyses involve multiple mediators, either because interventions are designed to affect outcome by changing multiple (repeated measures of) mediators, or because scientific interest is in trying to understand the various causal pathways through (simultaneous) competing candidate mediators. *Path analysis* (Wright, 1934; Duncan, 1966) is therefore commonly used to disentangle the *indirect* or *mediated* effects of A on Y along the causal paths through the multiple mediators.

Building on our example, the *causal diagram* of Figure 1(a) depicts the causal relations between the variables when worldview dissimilarity (M_1) and fairness (M_2) are assumed not to affect each other. In this article, a causal diagram is a causal directed acyclic graph (DAG) (Hayduk et al., 2003; Pearl, 2012) that, similar to path diagrams in the structural equation modeling (SEM) framework, represents (assumed) causal relations among a set of variables. Vertices represent variables, and a directed edge e.g., from M_1 to Y, represents the causal effect M_1 may exert on Y. The absence of a

directed edge between two variables, e.g., between M_1 and M_2 in Figure 1(a), implies that neither variable causally affects the other, conditional on their common causes, e.g., treatment A and observed baseline covariate(s), such as political ideology, henceforth denoted by C. A summary of key concepts of causal DAGs can be found in e.g., Moerkerke et al. (2015, Figure 2). Here and throughout, the causal effects between the treatment and each mediator, between each mediator and the outcome, and between the treatment and the outcome, are assumed to be based on well-established scientific theoretical knowledge or empirical laws that satisfy logical and causal-temporal constraints (Fiedler et al., 2018). Unlike path diagrams, causal DAGs do not rely on (parametric) assumptions about the nature of the relationship between the variables; hence path coefficients and error terms are not displayed on causal diagrams in this article.

Mediation using path analysis within the SEM framework extends the Baron & Kenny (1986) approach for a single mediator by employing (multiple) linear regression models for the mediator(s) and the outcome; see e.g., MacKinnon (2008) and Hayes (2018) for book-length presentations and the detailed references therein. A linear path analysis model (or set of linear regression models) is first fitted to the outcome and the mediator(s) using SEM (or ordinary least squares; OLS). The effect of treatment transmitted along a particular path, as encoded by the (partial) regression coefficients of the variables on the path, is then calculated using the *product-of-coefficients* method (Alwin & Hauser, 1975; MacKinnon et al., 2002). This method prevails (as opposed to the "difference" method) when there are multiple mediators (MacKinnon, 2000; Preacher & Hayes, 2008). Continuing our example above in the causal diagram of Figure 1(a), when M_1 and M_2 are assumed not to affect each other, the indirect effect via M_1 in the corresponding (linear) path model is defined to be the product of the coefficient of A in the regression of M_1 on A and C, and the coefficient of M_1 in the regression of Y on A, M_1, M_2 and C.

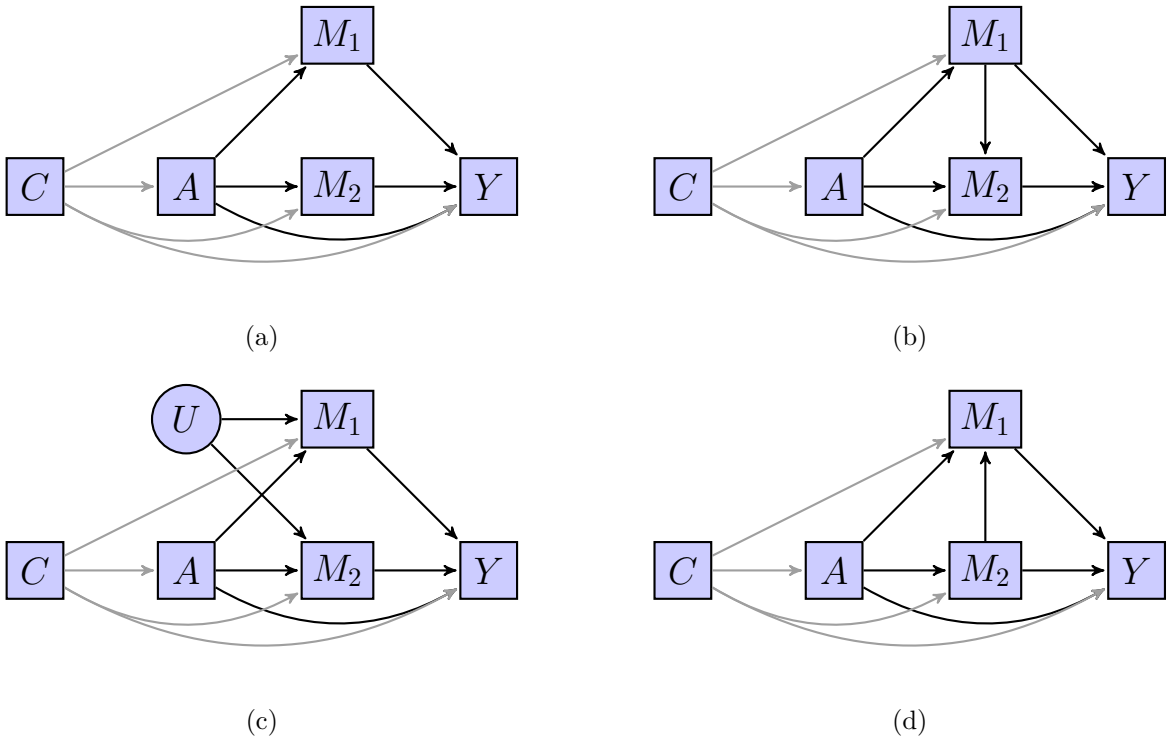

Figure 1. Causal diagrams with two mediators where either (a) M_1 and M_2 are independent conditional on A and C, or (b) M_1 causally precedes M_2, or (c) M_1 and M_2 do not affect each other but share an unobserved common cause U, or (d) M_2 causally precedes M_1. Rectangular nodes denote observed variables, while round nodes denote unobserved variables. For visual clarity, edges emanating from C are drawn in gray.

Existing path analysis approaches for multiple causally linked mediators

When the assumed causal structure of the mediators allows for "compound" paths from treatment to outcome that traverse several (causally) linked mediators, each indirect effect along any path that passes through at least one of the mediators can be separately assessed (Hayes, 2018). Using the motivating example, suppose that worldview dissimilarity (M_1) is assumed to causally affect fairness (M_2), as depicted in the causal diagram of Figure 1(b). The "three-path" mediated effect passing through both mediators along the path $A \to M_1 \to M_2 \to Y$ (Taylor et al., 2008) is defined to be the product of the coefficient of A in the regression of M_1 on A and C, the coefficient of M_1 in the regression of M_2 on A, M_1 and C, and the coefficient of M_2 in the regression of Y on A, M_1, M_2 and C. However, there are three potential pitfalls when

estimating the separate indirect effects along each path. First, the indirect effects are well-defined exclusively when the assumed (directions of the) causal effects among the mediators are correct. For example, suppose that the true causal relation between the mediators was that fairness (M_2) affected worldview dissimilarity (M_1) as shown in the causal diagram of Figure 1(d). The estimated mediated effect along the assumed path $A \to M_1 \to M_2 \to Y$ will not (generally) be unbiased for the true effect along the (different) path $A \to M_2 \to M_1 \to Y$. Second, even when the causal effects are correctly specified, estimates can be biased when there is hidden or unobserved confounding of the mediators. For example, suppose that worldview dissimilarity (M_1) and fairness (M_2) do not causally depend on each other, but instead share a hidden confounder U (such as prior adverse interactions with a political outgroup) as depicted in Figure 1(c). Because neither mediator exerts a causal effect on the other, there is no causal effect along any path where one mediator affects the other. But incorrectly assuming that M_1 affects M_2 will result in biased (non-zero) estimates of the mediated effect along the path $A \to M_1 \to M_2 \to Y$. Specifically, the bias is due to biased estimates of the (partial) regression coefficient of M_1 (in the regression of M_2 on A, M_1 and C) because M_1 and M_2 are correlated only due to hidden U, and not because M_1 influences M_2. Third, even when the causal effects are correctly specified, and there is no hidden confounding of the mediators, the causal effect transmitted along the path $A \to M_1 \to M_2 \to Y$ cannot be (non-parametrically) identified in general without (empirically untestable) assumptions about the mediators' joint distribution. We elaborate on this last point using the counterfactual-based mediation framework later in this article.

Notwithstanding the possibility of (unbiasedly) estimating separate indirect effects along different assumed paths, substantive interest may be in the indirect effects transmitted through each distinct mediator instead. We first consider settings where the causal structure among the mediators can be (correctly) assumed, and defer settings that do not require assuming any causal structure (among the mediators) to the next section. Separate indirect effects along a set of different paths may be combined based on a given definition of an indirect effect via a specific mediator (Bollen, 1987). For

example, Greene (1977) proposes a restrictive definition of the indirect effect via a mediator of interest to include only the path that intersects the particular mediator alone and no other mediators, whereas Brown (1997), following Fox (1980), proposes a less restrictive definition that includes all paths that intersect the particular mediator. The former approach fails to account for mediated effects along compound paths traversing several linked mediators, whereas the latter approach can potentially yield indirect effects whose sum is greater than the total (treatment) effect. Alwin & Hauser (1975) propose including all paths intersecting that mediator and any of its descendants, and excluding all paths via mediators that causally precede the mediator of interest. This approach ensures that the sum of the indirect effects via each mediator equals the joint indirect effect via all the mediators, but only when each indirect effect via a mediator is not moderated by any other mediator. Furthermore, we reiterate that it is necessary to correctly specify the (directions of the) causal effects among the mediators in the assumed path model for the above definitions of indirect effects via each distinct mediator to be valid.

In general when multiple mediators are correlated, there may be several plausible explanations of the associations, such as those depicted in the causal diagrams of Figures 1(b), 1(c), and 1(d) for M_1 and M_2. Researchers seeking to learn about the causal effects among the mediators may consider fitting different models assuming different (possibly conflicting) causal structures, then ideally select the model that best fits the observed data. But different models can be statistically indistinguishable, e.g., when they are saturated with zero degrees of freedom, or have identical goodness-of-fit measures. Merely assuming causal effects among the mediators can result in severely misleading conclusions about the mediated effects. Even when the assumed causal effects among the mediators are correct, it may be unrealistic to assume that the mediators do not share hidden or unobserved confounders, such as when the mediators are manifestations of an unknown latent variable or process.

A different definition of indirect effects via each mediator

In this article we propose a different definition of indirect effects when substantive interest is in the indirect effects that are specific to each mediator. The proposed definitions do not require (correctly) specifying how the mediators causally depend on one another. We first provide an intuitive motivation, and defer its formal derivation to the next section. Suppose that the outcome obeys the following linear and additive mean model:

$$\mathrm{E}(Y|A, M_1, M_2, C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_C C. \tag{1}$$

The effect of each mediator M_s on outcome is encoded by the (partial) regression coefficient $\beta_s, s = 1, 2$ in (1). To avoid specifying any causal dependence among the mediators, consider the *marginal* mean model for each mediator $M_s, s = 1, 2$, that depends only on treatment A and baseline covariate(s) C, and does not depend on the other mediator. In particular, suppose that the linear and additive (marginal) mean model for M_s is:

$$\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, \quad s = 1, 2. \tag{2}$$

It is important to note that (2) does not imply assuming that there are no causal effects between M_1 and M_2. Indeed, the intention is simply to leave the causal structure of the mediators unspecified, and to only consider the "overall" or "total" (i.e., marginal) effect of treatment A on each mediator $M_s, s = 1, 2$ (conditional on C). The overall effect of A on M_s captures all of the treatment effects that are transmitted through any intermediate variables on the causal pathway between A and M_s. This overall or total effect is simply encapsulated by the (partial) regression coefficient $\delta_s, s = 1, 2$ in (2). Using the product-of-coefficients method, the indirect effect via mediator M_s is therefore encoded by $\beta_s \delta_s$. Estimating this product requires no knowledge of the (possibly unknown) causal structure among the mediators from which the observed data is generated.

The mean models implied by (1) and (2) therefore adopt the same functional form as a (linear) path model corresponding to the causal diagrams of Figures 1(a) and 1(c),

where the mediators are a priori assumed not to causally affect each other. Such path models (without baseline covariates C) are termed "parallel" multiple mediator models (Hayes, 2018). In fact, a prevalent approach (MacKinnon, 2000; Preacher & Hayes, 2008) is to fit a parallel multiple mediator model (henceforth termed a *parallel path model* for simplicity) to the observed data. Recent examples across different areas in psychology implementing this prevailing approach for multiple mediation analysis are provided in the Discussion section. In a fitted parallel path model, the formulae for the indirect effect is $\beta_s\delta_s$, because there is only one path from A to Y that intersects each mediator M_s. Hence, existing indirect effects using parallel path models possess the same interpretation as the proposed indirect effects when considering only the marginal effect of treatment on each mediator, regardless of the unknown causal structure among the mediators. Furthermore, the indirect effect estimates are robust against unobserved 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 A and baseline covariates C. This is because the (point) estimates of the path coefficients β_s and δ_s in a fitted parallel path model remain the same regardless of whether the mediator residuals covariances are constrained to zero (MacKinnon, 2000), or freely estimated (Preacher & Hayes, 2008).

The indirect effect estimates remain unchanged when the mediator residuals are permitted to covary only if either (i) there are no mediator-mediator interaction terms in the outcome mean model (1) or (ii) the covariances do not depend on treatment. When other mediators moderate the effect of each mediator on the outcome, and the mediators' covariances are affected by treatment, existing indirect effect estimates assuming a parallel path model may be severely biased. Furthermore, there can be an indirect effect that cannot be attributed to any specific mediator due to the mediators' *mutual dependence on one another*. We therefore exploit the proposed definitions in this article to develop indirect effects under such settings. We propose novel estimators of the indirect effects that allow for the effect of each mediator on the outcome to be moderated by treatment, or another mediator, or both, without having to specify a

causal structure among the mediators. Unlike Hayes & Rockwood (2020) who consider settings where covariates (unaffected by treatment) moderate the indirect effects of each mediator, in this article we will focus on settings where mediators (possibly affected by treatment) moderate each other's indirect effects, and the mediators' covariances are permitted to change with treatment. The rest of this article is organized as follows. A brief introduction to causal mediation using the counterfactual-based framework is given. We describe conceptual definitions of the proposed *interventional* (in)direct effects using the minimal example above with two mediators. Interpretations using the causal diagrams of Figure 1 are provided to give readers intuition into the proposed (in)direct effects. We formally demonstrate that under assumed linear and additive models for the mediators and outcome, estimators of the interventional (in)direct effects have the same analytical form as existing product-of-coefficient estimators in a fitted parallel path model. Next, we propose novel estimators of the indirect effect due to the mediators' mutual dependence (on one another). We exploit the mediators' covariances under the assumed linear mean models, which simplifies closed form solutions for settings with multiple mediators. Simulation studies based on a substantive mediation analysis are used to illustrate estimating the proposed interventional (in)direct effects. We empirically demonstrate how a misspecified path model, by either incorrectly assuming causal effects among the mediators, or omitting mediator-mediator interaction terms in the outcome model, can result in misleading conclusions about the indirect effects. The proposed methods are utilized to assess the extent that the effect of political inclusion on political prejudice is mediated by six distinct possible mediators in a social psychology experiment. All scripts used to carry out the simulation studies, and to estimate the interventional direct and indirect effects in the applied example, are implemented in the open source `R` (R Core Team, 2019) statistical software environment. The scripts are freely available online[1], with more user-friendly functions for applied researchers under development. We conclude with recommendations and practical considerations for applied researchers using multiple mediation analyses to

[1] `https://github.com/wwloh/disentangle-multiple-mediators`

answer substantive questions.

Interventional direct and indirect effects

Mediation analysis using a counterfactual-based framework

Notwithstanding the widespread use of parametric approaches to mediation analysis, a counterfactual-based framework for mediation analysis has been developed using model-free definitions of *natural direct and indirect effects* (Robins & Greenland, 1992; Pearl, 2001). This development enables extensions to non-additive and non-linear models, and formalizes the "ignorability" assumptions needed to identify the natural (in)direct effects, without relying on a specific statistical model; see e.g., Imai, Keele, & Tingley (2010) and Pearl (2014) for the single mediator setting. Under these assumptions, the total effect can be decomposed into a direct and an indirect effect. Using linear and additive (i.e., without interactions) mean models for the mediator and outcome in the mediation formula yields the same estimators as the path analysis approach using the product-of-coefficients method (Imai, Keele, & Yamamoto, 2010).

When there are multiple mediators, identifying the natural indirect effects via each mediator is challenging because one mediator that is affected by treatment can concurrently be a confounder of the mediator-outcome association for another mediator, also known as *post-treatment* or *treatment-induced confounding*; see e.g., VanderWeele et al. (2014) and Moerkerke et al. (2015). Here and throughout we will consider all post-treatment confounders to be competing possible mediators. Continuing our example above, suppose that worldview dissimilarity (M_1) affects fairness (M_2), as depicted in Figure 1(b), so that M_1 is a post-treatment confounder of the $M_2 - Y$ relation. Then the natural indirect effect via M_2 cannot be (non-parametrically) identified because M_1 is a *recanting witness* that is set to different counterfactual values along the different paths $A \to M_1 \to Y$ and $A \to M_1 \to M_2 \to Y$ (Avin et al., 2005). Strong (empirically untestable) parametric assumptions about the joint distribution of the (counterfactual) mediators are required to identify the natural effects (Shpitser, 2013). Recent proposals of counterfactual-based mediation analysis for multiple

(repeated measures of) mediators have thus relied on stringent ("sequential ignorability") assumptions to carefully identify natural indirect effects either along certain causal pathways (Daniel et al., 2015; Steen et al., 2017; Albert et al., 2019), or assuming no causal effects among the mediators (Lange et al., 2013; Taguri et al., 2018). But (fine-grained) decompositions of indirect effects using these existing methods require correctly specifying the (absence of) causal effects among the mediators, and assuming that the mediators share no hidden confounders. In most realistic scenarios, the directions of the causal effects between the various mediators are unknown, thus either violating the assumptions needed to identify the indirect effects, or demanding additional assumptions about the correct specification of the causal structure.

In contrast, *interventional (in)direct effects*, first introduced by Didelez et al. (2006) and VanderWeele et al. (2014) for a single mediator, then generalized by Vansteelandt & Daniel (2017) to the multiple mediator setting, can be identified under much weaker conditions than natural effects, and still achieve an exact decomposition of the total effect. Unlike natural effects that are defined in terms of individual-level (deterministic) interventions on the mediator, interventional effects consider population-level (stochastic) interventions that set the value of the mediator to a random draw from its counterfactual distribution. Continuing with the motivating example above, the natural indirect effect via worldview dissimilarity (M_1) is the average change in prejudice (Y) when each individual's (counterfactual) value of worldview dissimilarity is manipulated from being political included ($A = 1$) to that under control ($A = 0$). In contrast, the interventional indirect effect is the average change in prejudice when the (counterfactual) distribution of worldview dissimilarity under the political inclusion manipulation is shifted to that under control. Interventional effects can therefore be scientifically meaningful even when the treatment cannot be realistically manipulated at the individual level. For example, Jackson & VanderWeele (2018) describe interventional (in)direct effects using race as the treatment and socioeconomic status as the mediator, without having to define nested potential outcomes (for each individual) where race is set to one group but socioeconomic status

is simultaneously set to its potential value under a different group, depending on the treatment effect of race. Quynh Nguyen et al. (2019) compare different definitions of direct and indirect effects used in causal mediation analysis that may be motivated by different research questions. We refer interested readers to Lin & VanderWeele (2017), Moreno-Betancur & Carlin (2018), and Lok (2019) among many others for discussions of interventional effects in the causal inference and epidemiology literature.

Definition of potential outcomes

To facilitate the conceptual development of interventional (in)direct effects, we present definitions under a setting with two mediators M_1 and M_2, and defer results for more than two mediators to the Online Supplemental Materials. In this article, uppercase letters denote (observed) random variables and (possibly unobserved) potential outcomes, and lowercase letters denote specific values, for each individual. For $s = 1, 2$, let $M_{s,a^{(s)}}$ denote the potential outcome for M_s if, possibly counter to fact, treatment A is set to $a^{(s)}$. Let $Y_{am_1m_2}$ denote the (individual) potential outcome for Y if, possibly counter to fact, A is set to a, and when each mediator M_s is set to the value $m_s, s = 1, 2$.

Definition of interventional indirect and direct effects

In this section, we formally define the interventional indirect and direct effects and describe the exact decomposition of the total effect for a binary treatment A. We provide interpretations of the interventional (in)direct effects in terms of the underlying causal path(s) in the causal diagrams of Figure 1.

Define the interventional indirect effect of treatment on outcome via M_1, henceforth denoted by IE_1, as:

$$\text{E}\left[\sum_{m_1,m_2} \text{E}(Y_{1m_1m_2}|C)\left\{\Pr(M_{1,1} = m_1|C) - \Pr(M_{1,0} = m_1|C)\right\}\Pr(M_{2,0} = m_2|C)\right]. \quad (3)$$

The interventional indirect effect via mediator M_1 is the treatment effect of changing M_1 from its marginal (counterfactual) distribution under treatment level $a^{(1)} = 1$ to its distribution under level $a^{(1)} = 0$, while fixing the mediator M_2 at its distribution under

treatment level $a^{(2)} = 0$, and the individual values of treatment at $a^{(0)} = 1$. Continuing the example above, the interventional indirect effect via worldview dissimilarity is the average difference in political prejudice (Y) when the distribution of worldview dissimilarity is shifted from political inclusion to the control condition, while holding the distribution of perceived fairness (M_2) fixed under the control condition, among individuals in the political inclusion group. In other words, the indirect effect describes how political prejudice is potentially affected (on average) when worldview dissimilarity is randomly drawn from its (counterfactual) distribution under treatment, as compared to a specific other distribution under control. The distributions under treatment and under control need not differ only in terms of the location (mean) parameter, or the scale (variance) parameter, or both, and may generally adopt different (parametric) forms.

The interventional indirect effect via M_1 in (3) is defined to be a function of the difference in (marginal) probabilities $\Pr(M_{1,1} = m_1|C) - \Pr(M_{1,0} = m_1|C)$. When the underlying causal structure among the mediators is as depicted in the causal diagrams of Figures 1(a) – 1(c), the indirect effect via M_1 therefore corresponds to the causal effect transmitted along the path $A \to M_1 \to Y$; whereas in the causal diagram of Figure 1(d), the indirect effect combines the effects along the paths $A \to M_1 \to Y$ and $A \to M_2 \to M_1 \to Y$. The interventional indirect effect via M_1 thus captures all of the treatment effect that is mediated by M_1, and any other mediators causally preceding M_1, in the underlying causal diagram.

Similarly, define the interventional indirect effect of treatment on outcome via M_2, henceforth denoted by IE_2, as:

$$\text{E}\left[\sum_{m_1,m_2} \text{E}(Y_{1m_1m_2}|C)\Pr(M_{1,1} = m_1|C)\left\{\Pr(M_{2,1} = m_2|C) - \Pr(M_{2,0} = m_2|C)\right\}\right]. \quad (4)$$

The interventional indirect effect via mediator M_2 can be analogously interpreted as the indirect effect via M_1. In particular, when the underlying causal structure among the mediators is as depicted in the causal diagrams of Figures 1(a), 1(c), and 1(d), the interventional indirect effect via M_2 (4) corresponds to the causal effect transmitted along the path $A \to M_2 \to Y$; in the causal diagram of Figure 1(b), the indirect effect

combines the effects along the paths $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$. As before, the interventional indirect effect via M_2 is interpreted as the effect of treatment that is mediated by M_2, and any other mediators causally preceding M_2, in the underlying causal diagram. In general, when there are $t > 2$ distinct mediators, the interventional indirect effect via each mediator M_s, henceforth denoted by IE_s, for $s = 1, \ldots, t$, is defined in the Online Supplemental Materials.

In this article, subscripts in the notation for mediators are merely used to distinguish the different mediators, and not to indicate an a priori assumed causal ordering of the mediators; e.g., M_1 is not necessarily assumed to causally precede M_2. Nonetheless, the definitions of the indirect effects via each mediator will generally, but not necessarily, differ by changing the indices of the mediators, due to fixing the other mediator at its distribution under a different hypothetical treatment level. For example, had worldview dissimilarity been merely indexed as M_2, the indirect effect via worldview dissimilarity would hold the distribution of perceived fairness (now M_1) fixed under the political inclusion condition instead. The hypothetical treatment levels are fixed at different values merely to ensure that the separate indirect effects via each mediator add up to the same quantity in (5) defined below, regardless of the (typically arbitrary) indices used solely to label the mediators for statistical analysis. Notwithstanding such differences, we emphasize that the conceptual interpretation of the interventional indirect effect via each mediator - in terms of the causal pathways in the underlying causal structure - is invariant to the different mediator indices. In later sections, we describe the estimators when the effect of each mediator on the outcome is moderated by treatment, or the other mediator, or both, and describe a sensitivity analysis.

It follows that the sum of the separate interventional indirect effects via each mediator, i.e., $\text{IE}_1 + \text{IE}_2$, is:

$$\text{E}\Big[\sum_{m_1,m_2} \text{E}(Y_{1m_1m_2}|C)\,\{\Pr(M_{1,1}=m_1|C)\Pr(M_{2,1}=m_2|C) - \Pr(M_{1,0}=m_1|C)\Pr(M_{2,0}=m_2|C)\}\Big]. \quad (5)$$

This indirect effect describes the average difference in the outcome when both marginal

(counterfactual) distributions of the mediators M_1 and M_2 are simultaneously shifted from the treated group ($a^{(1)} = a^{(2)} = 1$) to the control group ($a^{(1)} = a^{(2)} = 0$). Because the product of the marginal distributions does not equal the joint distribution of the mediators in general, define the *joint* indirect effect via the mediators, henceforth denoted by IE_{jo}, as:

$$\text{E}\Big[\sum_{m_1,m_2} \text{E}(Y_{1m_1m_2}|C)\,\{\Pr(M_{1,1}=m_1, M_{2,1}=m_2|C) - \Pr(M_{1,0}=m_1, M_{2,0}=m_2|C)\}\Big]. \quad (6)$$

We emphasize that in (6) the joint (counterfactual) distribution of the mediators M_1 and M_2 is shifted, instead of the marginal distributions as defined in (5). The difference between IE_{jo} in (6) and $\text{IE}_1 + \text{IE}_2$ in (5), henceforth denoted by IE_{mu}, is therefore:

$$\text{E}\left[\sum_{m_1,m_2} \text{E}(Y_{1m_1m_2}|C)\left\{\Pr(M_{1,1}=m_1, M_{2,1}=m_2|C) - \prod_{s=1}^{2}\Pr(M_{s,1}=m_s|C) - \Pr(M_{1,0}=m_1, M_{2,0}=m_2|C) + \prod_{s=1}^{2}\Pr(M_{s,0}=m_s|C)\right\}\right]. \quad (7)$$

We refer to (7) as the *indirect effect due to the mediators' mutual dependence on each other* (Vansteelandt & Daniel, 2017). We will demonstrate in the next section that under assumed linear models for the means of the mediators and the outcome, this indirect effect is non-zero only if (i) the effect of each mediator on the outcome is moderated by the other mediator, and (ii) the covariance of the mediators is affected by treatment. Because this effect cannot be attributed to any mediator alone, it should be considered separately from the indirect effects via each mediator.

The interventional direct effect of treatment on outcome that avoids both mediators, henceforth denoted by DE, is correspondingly defined as:

$$\text{E}\left[\sum_{m_1,m_2} \{\text{E}(Y_{1m_1m_2}|C) - \text{E}(Y_{0m_1m_2}|C)\}\Pr(M_{1,0}=m_1, M_{2,0}=m_2|C)\right]. \quad (8)$$

The direct effect (8) is the treatment effect when controlling the joint (counterfactual) distribution of the mediators M_1 and M_2 to be under control, i.e., $a^{(1)} = a^{(2)} = 0$. The direct effect (8) corresponds to the causal effect along the path $A \to Y$ that avoids all the mediators in the causal diagrams of Figure 1. Define the sum of the joint indirect

effect (6) and the direct effect (8) to be the *total* effect of treatment on outcome, henceforth denoted by TE, as:

$$\mathrm{E}\Bigg[\sum_{m_1,m_2}\Big\{\mathrm{E}(Y_{1m_1m_2}|C)\Pr(M_{1,1}=m_1,M_{2,1}=m_2|C) - \mathrm{E}(Y_{0m_1m_2}|C)\Pr(M_{1,0}=m_1,M_{2,0}=m_2|C)\Big\}. \quad (9)$$

In other words, $\mathrm{TE}=\mathrm{DE}+\mathrm{IE}_{jo}=\mathrm{DE}+\mathrm{IE}_1+\mathrm{IE}_2+\mathrm{IE}_{mu}$. Definitions for settings with more than two distinct mediators are provided in the Online Supplemental Materials.

Identification of interventional effects

Identification of the interventional effects defined above requires the following assumptions (Vansteelandt & Daniel, 2017):

(A1) The effect of treatment A on outcome Y is unconfounded conditional on C.

(A2) The effect of both mediators M_1, M_2 on outcome Y is unconfounded conditional on A and C.

(A3) The effect of treatment A on both mediators is unconfounded conditional on C.

Assumption (A1) states that there are no unobserved confounders between A and Y, or equivalently, that the observed covariate(s) C are sufficient to adjust for confounding of the effect of A on Y. This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of A and Y.

Because the potential outcome $Y_{am_1m_2}$ is unknown for each value of (a, m_1, m_2) except for the observed realization (A, M_1, M_2), assumption (A2) states that there is available sufficient covariate information observed in C so that the association between any of (M_1, M_2) and Y is unconfounded within levels of the covariate(s) C. This assumption requires that there is no confounder of any mediator-outcome association that is affected by treatment; such potential post-treatment confounders can merely be included in the set of possible mediators under the multiple mediator setting considered in this article. This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of any of (M_1, M_2) and Y. Hence one possible

scenario under which assumption (A2) is violated, as illustrated by Mayer et al. (2014) in their opening example (for a single mediator), is when baseline measurements of a mediator and the outcome are correlated and unadjusted for, even in a randomized experiment. For this reason, baseline covariates were adjusted for in models (1) and (2).

Because $M_{1,a^{(1)}}$ and $M_{2,a^{(2)}}$ are unknown for each value of $\{a^{(1)}, a^{(2)}\}$ except when $a^{(1)} = a^{(2)} = A$, assumption (A3) states that there are no unobserved confounders between A and any of (M_1, M_2), or equivalently, that the observed covariate(s) C are sufficient to adjust for confounding of the effects of A on (M_1, M_2). This assumption is implied in the causal diagrams of Figure 1 by the absence of any hidden common causes of A and any of (M_1, M_2). Note that assumptions (A1) and (A3) are satisfied in randomized trials when A is randomly assigned. When treatment is not randomly assigned, observed (baseline) confounders of the treatment-mediator(s) and treatment-outcome should be included in C and adjusted for in the mediator and outcome models.

When the assumptions (A1)–(A3) hold, the interventional direct and indirect effects defined above can be inferred from the observed data. In particular, the average potential outcomes, and joint distribution of the counterfactual mediators, can be (non-)parametrically identified by the following observable quantities:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y_{a^{(0)}m_1m_2}|C) \prod_{s=1}^{2} \Pr(M_{s,a^{(s)}} = m_s|C)\right] \\
&= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y|A = a^{(0)}, M_1 = m_1, M_2 = m_2, C) \prod_{s=1}^{2} \Pr(M_s = m_s|A = a^{(s)}, C)\right]; \qquad (10) \\
&\mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y_{a^{(0)}m_1m_2}|C) \Pr(M_{1,a^{(1)}} = m_1, M_{2,a^{(1)}} = m_2|C)\right] \\
&= \mathrm{E}\left[\sum_{m_1,m_2} \mathrm{E}(Y|A = a^{(0)}, M_1 = m_1, M_2 = m_2, C) \Pr(M_1 = m_1, M_2 = m_2|A = a^{(1)}, C)\right]. \qquad (11)
\end{aligned}$$

In practice, unbiased estimation of the interventional direct and indirect effects therefore depends on correctly modeling the outcome conditional on treatment, mediators, and covariates, e.g., $\mathrm{E}(Y|A = a, M_1 = m_1, M_2 = m_2, C)$, that is unbiased for $\mathrm{E}(Y_{am_1m_2}|C)$, and a joint distribution of the observed mediators conditional on

treatment and covariates, e.g., $\Pr(M_1 = m_1, M_2 = m_2|A = a, C)$, that is unbiased for $\Pr(M_{1,a} = m_1, M_{2,a} = m_2|C)$.

Estimators of interventional indirect and direct effects

In this section we describe estimators of the interventional (in)direct effects defined in the previous section. We first assume a parallel path model with no interaction terms in the outcome model, then relax this assumption to allow for treatment-mediator, mediator-mediator, and treatment-mediator-mediator interactions.

Outcome models without interaction terms

Suppose that the outcome obeys the linear and additive mean model in (1), i.e., $\mathrm{E}(Y|A, M_1, M_2, C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_C C$. Further suppose that the marginal treatment effect on each mediator, given baseline covariate(s) C, is parametrized by the (partial) regression coefficient of treatment A in the linear and additive (marginal) mean models in (2), i.e., $\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{Cs} C, s = 1, 2$. The interventional indirect effect via each mediator $M_s, s = 1, 2$, is identified upon plugging the assumed outcome model (1) and mediator models (2) into (10); i.e.,

$$\mathrm{IE}_s = \beta_s \left\{\mathrm{E}(M_s|A = 1) - \mathrm{E}(M_s|A = 0)\right\} = \beta_s \delta_s.$$

The joint indirect effect (IE_{jo}) and direct effect (DE) are similarly identified under the assumed outcome model (1) and mediator models (2) using (11); i.e.,

$$\mathrm{IE}_{jo} = \sum_{s=1}^{2} \beta_s \left\{\mathrm{E}(M_s|A = 1) - \mathrm{E}(M_s|A = 0)\right\} = \sum_{s=1}^{2} \beta_s \delta_s, \ \mathrm{DE} = \beta_A.$$

Unbiased estimation of the interventional (in)direct effects thus requires correctly specifying the outcome mean model (1) and mediator (marginal) mean models (2) under assumptions (A1)–(A3).

As previously noted in the introduction, the derivation of the interventional effects imply mean models (1) and (2) that adopt the same form for the expected values of the outcome Y and mediators $M_s, s = 1, 2$, as a parallel path model where the mediators are assumed not to causally affect each other. The interventional indirect effect via each

mediator M_s thus equals the indirect effect using the product-of-coefficients method $\beta_s\delta_s$ for the path $A \to M_s \to Y$ in the parallel path model. Similarly, the interventional direct effect equals β_A for the path $A \to Y$ that avoids both mediators in the parallel path model. Estimators of the interventional effects can therefore be straightforwardly obtained by fitting the parallel path model to the observed data using linear SEM or OLS, then plugging in estimates of the (partial) regression coefficients for the respective effects using the product-of-coefficients method. Standard errors can be estimated using a nonparametric percentile bootstrap procedure (Efron & Tibshirani, 1994) that randomly resamples observations with replacement. In general when there are $t > 2$ distinct mediators, the estimators of the interventional indirect effects via each mediator $M_s, s = 1, \ldots, t$, are described in the Online Supplemental Materials.

Again we emphasize that δ_s in (2) encodes the overall or total effect of A on M_s and captures all of the underlying treatment effects that are transmitted from A to M_s through any causal ancestors of M_s. To see why fitting mean models (1) and (2) is sufficient to obtain unbiased estimators of the interventional (in)direct effects, consider the continuing example from the introduction corresponding to the causal diagram of Figure 1(b). Suppose that the observed data is generated from a true (but unknown) path model where M_1 has a causal effect on M_2, with the mediator and outcome models:

$$\begin{aligned}
\mathrm{E}(M_1|A, C) &= \alpha^*_{01} + \alpha^*_1 A + \alpha^*_{C1} C, \\
\mathrm{E}(M_2|A, M_1, C) &= \alpha^*_{02} + \alpha^*_2 A + \eta^*_{12} M_1 + \alpha^*_{C2} C, \\
\mathrm{E}(Y|A, M_1, M_2, C) &= \beta^*_0 + \beta^*_A A + \beta^*_1 M_1 + \beta^*_2 M_2 + \beta^*_C C.
\end{aligned}$$

(Asterisks denote parameters in the true but unknown data-generating model.) Note that the α_s parameter encodes the conditional association between treatment A and mediator $M_s, s = 1, 2$, possibly given other mediators such as M_1 in the model for M_2, and will therefore differ from δ_s in general. The mean of the implied marginal

distribution of M_2, obtained by averaging over the distribution of M_1, is then:

$$\begin{aligned}
\mathrm{E}(M_2|A,C) &= \sum_{m_1} \mathrm{E}(M_2|A,M_1=m_1,C)\Pr(M_1=m_1|A,C) \\
&= \alpha^*_{02} + \alpha^*_2 A + \eta^*_{12}\,\mathrm{E}(M_1|A,C) + \alpha^*_{C2}C \\
&= \alpha^*_{02} + \alpha^*_2 A + \eta^*_{12}(\alpha^*_{01} + \alpha^*_1 A + \alpha^*_{C1}C) + \alpha^*_{C2}C \\
&= (\alpha^*_{02} + \eta^*_{12}\alpha^*_{01}) + (\alpha^*_2 + \eta^*_{12}\alpha^*_1)A + (\alpha^*_{C2} + \eta^*_{12}\alpha^*_{C1})C.
\end{aligned}$$

The interventional indirect effect via M_2 in the true but unknown model is thus identified by $\beta^*_2(\alpha^*_2 + \eta^*_{12}\alpha^*_1)$. By fitting to the observed data a parallel path model with outcome mean model (1), so that $\beta_2 = \beta^*_2$, and mediator mean model (2), so that $\delta_2 = \alpha^*_2 + \eta^*_{12}\alpha^*_1$, it follows that the interventional indirect effect can be unbiasedly estimated using the product-of-coefficients method because $\beta_2\delta_2 = \beta^*_2(\alpha^*_2 + \eta^*_{12}\alpha^*_1)$ (assuming (A1)–(A3) hold). Hence the parallel path model is used merely to obtain estimators of the interventional indirect and direct effects using ubiquitous linear SEM or OLS estimation methods. Unbiased estimation does not require the mediators to be causally independent, as implied in the parallel path model; in fact, the (marginal) mean model (2) is used precisely so that the interventional indirect effects are agnostic to the underlying causal dependence among the mediators.

Outcome models with treatment-mediator, mediator-mediator, and treatment-mediator-mediator interaction terms

Under the assumed outcome model (1), the joint indirect effect via both mediators (IE_{jo}) equalled the sum of both separate indirect effects via each mediator ($\mathrm{IE}_1 + \mathrm{IE}_2$). As we will demonstrate next, this was a consequence of excluding mediator-mediator interaction terms from the outcome model, which precluded a non-zero estimate of the indirect effect due to the mediators' mutual dependence on each other (IE_{mu}). The effect of a mediator on the outcome is moderated by a third variable if the effect depends on, or is a function of, the third variable. In this section, we develop estimators for the interventional (in)direct effects when the effect of each mediator on the outcome is moderated by treatment, or the other mediator, or both. First, allow for the following

treatment-mediator, mediator-mediator, and treatment-mediator-mediator interaction terms in the assumed linear mean model (1); i.e.,

$$\begin{aligned}\mathrm{E}(Y|A, M_1, M_2, C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_{A1} A M_1 + \beta_{A2} A M_2 \\ + \beta_{12} M_1 M_2 + \beta_{A12} A M_1 M_2 + \beta_C C. \end{aligned} \tag{12}$$

Next, allow the mediators' covariance to depend on treatment A, which we denote by $\mathrm{cov}(M_1, M_2|A) = \Sigma(A)$ for notational simplicity. Then under the outcome mean model (12) and mediator mean models (2), the indirect effect due to the mediators' mutual dependence (7) is identified by:

$$\begin{aligned}&\mathrm{E}\Bigg[\sum_{m_1,m_2} \mathrm{E}(Y|A=1, m_1, m_2, C) \\ &\quad\times \Big\{ \Pr(M_1 = m_1, M_2 = m_2|A=1, C) - \Pr(M_1 = m_1|A=1, C)\Pr(M_2 = m_2|A=1, C) \\ &\qquad - \Pr(M_1 = m_1, M_2 = m_2|A=0, C) + \Pr(M_1 = m_1|A=0, C)\Pr(M_2 = m_2|A=0, C)\Big\}\Bigg] \\ &= (\beta_{12} + \beta_{A12})\,\mathrm{E}\{\mathrm{cov}(M_1, M_2|A=1, C) - \mathrm{cov}(M_1, M_2|A=0, C)\} \\ &= (\beta_{12} + \beta_{A12})\{\Sigma(1) - \Sigma(0)\}. \end{aligned} \tag{13}$$

We make the perhaps obvious point that the indirect effect (13) is non-zero only if (i) there is a non-zero (treatment-)mediator-mediator interaction in the outcome model (12), i.e., $\beta_{12} + \beta_{A12} \neq 0$; and (ii) the covariance of the mediators is affected by treatment, i.e., $\Sigma(1) - \Sigma(0) \neq 0$. Continuing the motivating example from the introduction, perceived worldview dissimilarity (M_1) and perceived fairness (M_2) may become more weakly correlated among those who were politically included ($A = 1$) than those in the control condition ($A = 0$). The treatment group-specific covariances of the mediators can be straightforwardly estimated from the observed data using the empirical covariances within each observed treatment group.

The interventional indirect effect via mediator M_1 (3) under the outcome mean

model (12) is identified by:

$$\begin{aligned}
&\mathrm{E}\Bigg[\sum_{m_1,m_2}\mathrm{E}(Y|A=1,m_1,m_2,C)\\
&\quad\times\Big\{\Pr(M_1=m_1|A=1,C)-\Pr(M_1=m_1|A=0,C)\Big\}\Pr(M_2=m_2|A=0,C)\Bigg]\\
&=\{(\beta_1+\beta_{A1})+(\beta_{12}+\beta_{A12})\,\mathrm{E}(M_2|A=0)\}\delta_1\\
&=\{(\beta_1+\beta_{A1})+(\beta_{12}+\beta_{A12})(\delta_{02}+\delta_{C2}\mu_C)\}\delta_1,
\end{aligned}$$

where we denote $\mu_C = \mathrm{E}(C)$ for simplicity. The interventional indirect effect via mediator M_2 (4) is similarly identified by:

$$\{(\beta_2+\beta_{A2})+(\beta_{12}+\beta_{A12})\,\mathrm{E}(M_1|A=1)\}\delta_2=\{(\beta_2+\beta_{A2})+(\beta_{12}+\beta_{A12})(\delta_{01}+\delta_1+\delta_{C1}\mu_C)\}\delta_2.$$

Estimators of the interventional indirect effects when there are more than two mediators are provided in the Online Supplemental Materials. We emphasize that unbiased estimation under the above (correctly-assumed) linear mean models for the mediators and outcome therefore requires no distributional assumptions on the random errors for the variables.

When a mean outcome model without interactions, such as (1), is assumed, the resulting indirect effect estimators are invariant to the mediator indices. In other words, merely switching the indices, e.g., by denoting worldview dissimilarity and fairness by M_2 and M_1 respectively, yields the same estimators. In contrast, under the assumed outcome model (12), the estimator of the indirect effect via one mediator is a (linear) function of the mean value of the other mediator under a given treatment level. Hence different indices lead to different estimators of the interventional indirect effects. For example, the indirect effect via M_1 would be $\{(\beta_1 + \beta_{A1}) + (\beta_{12} + \beta_{A12})\,\mathrm{E}(M_2|A = 1)\}\delta_1$ instead, where M_2 is now fixed at its mean value under treatment; similarly, the indirect effect via M_2 would be $\{(\beta_2 + \beta_{A2}) + (\beta_{12} + \beta_{A12})\,\mathrm{E}(M_1|A = 0)\}\delta_2$ instead, with M_1 fixed at its mean value under control. Because the choice of mediator indices can lead to different estimators, we describe in the applied example how to carry out a sensitivity analysis where the mediator indices are permuted, and the indirect effects estimated under each permutation. When the (true) effect of each mediator on the outcome is

moderated by treatment, or the other mediator, or both, (incorrectly) assuming no interactions in the outcome model can lead to biased estimates of the indirect effects. We demonstrate this empirically using a simulation study in the next section.

Simulation studies

Three simulation studies were conducted to illustrate estimating the proposed interventional indirect effects. In each study, all mediators were correlated due to an unobserved confounder (of the mediators). In Study 1, a setting with two mediators was used simply to demonstrate the estimators presented in the preceding section. In Studies 2 and 3, we considered more complex and realistic settings with four mediators. To provide an overview, in Study 2, a posited path model that incorrectly assumed (the presence of) causal effects among merely correlated mediators was used to estimate separate mediated effects along different paths. For comparison, existing product-of-coefficient estimators under a parallel path model, that adopted the same analytical form as interventional (in)direct effect estimators assuming linear and additive models for the mediators and outcome, were calculated. In Study 3, the effect of each (true) mediator on the outcome was moderated by another mediator, and the mediators' covariance (possibly) depended on treatment. The interventional indirect effects, including the indirect effect due to the mediators' mutual dependence on one another, were estimated using an outcome model that included all mediator-mediator interaction terms. For comparison, indirect effects using a misspecified parallel path model (with only main effects for the mediators in the outcome model) were estimated.

Simulation Study 1

Each observed dataset was generated with the following linear models corresponding to the causal diagram of Figure 2:

$$
\begin{aligned}
A &\sim \text{Bernoulli}(0.5) \\
C, U &\sim \mathcal{N}(1, 1) \\
M_1 &= \alpha_1 A + \alpha_{C1} C + \alpha_{U1} U + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2) \\
M_2 &= \alpha_2 A + \eta_{12} M_1 + \alpha_{C2} C + \alpha_{U2} U + \epsilon_2, \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2) \\
Y &= \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_C C + \epsilon_Y,\ \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2).
\end{aligned}
$$

Both mediators shared an unobserved (baseline) common cause U that precluded unbiased estimation of the separate (three-path) mediated effect along $A \to M_1 \to M_2 \to Y$ in general. The observed covariate C was a (baseline) confounder of both mediators and the outcome. The variables A, C, U, and residuals $\epsilon_1, \epsilon_2, \epsilon_Y$, were mutually independent of each other. The values of the (partial) regression coefficients in the data-generating model were set as $\alpha_1 = -0.09, \beta_2 = -0.66, \alpha_2 = \beta_1 = \beta_a = 0$, where the non-zero values were the path coefficient estimates in Figure 5a of Voelkel et al. (2019) that the motivating example was based on. The values of the remaining coefficients and residual variances were set to one for simplicity. The interventional indirect effect via M_1 corresponded to the causal effect along the path $A \to M_1 \to Y$, and was identified by $\beta_1 \alpha_1$. In this study, this indirect effect was zero because M_1 did not affect outcome ($\beta_1 = 0$). The interventional indirect effect via M_2 corresponded to the combined causal effect along the paths for $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$, and was identified by $\beta_2(\alpha_2 + \eta_{12}\alpha_1)$. In this study, even though M_2 was unaffected by treatment directly ($\alpha_2 = 0$), this indirect effect was non-zero because $\eta_{12}\alpha_1$ was non-zero. For simplicity, the direct effect β_A was zero.

The interventional (in)direct effects were estimated by simply fitting to each generated observed data: the (linear and additive) outcome mean model shown in (1), and a (marginal) linear and additive mean model for each mediator, where each mediator depended only on treatment A and observed (baseline) covariate C as shown

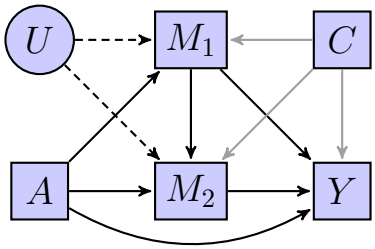

Figure 2. Causal diagram used to generate each simulated dataset with two possible mediators in simulation study 1. Rectangular nodes denote observed variables, while round nodes denote unobserved variables. For visual clarity, edges emanating from the baseline covariate C are drawn in gray, while edges emanating from the hidden confounder U are drawn as broken lines.

in (2). Under the assumed models, the interventional (in)direct effect estimators adopted the same analytical form as existing product-of-coefficient estimators under a parallel path model. We reiterate that the mediators need not be causally independent, as implied by the fitted parallel path model. The (marginal) mean models (2) are used solely to estimate the interventional indirect effects that are agnostic to the underlying causal structure among the mediators. 10000 observed datasets with sample size of either 50, 200, or 1000 were generated. Average estimates and empirical standard errors of the mediated effects from fitting the parallel path model to the generated data are displayed in Table 1. As expected, all the interventional (in)direct effects were unbiasedly estimated.

Simulation Study 2

This simulation study was motivated by a substantive mediation analysis on climate change beliefs and attitudes from the psychology literature (van der Linden et al., 2015). Each observed dataset was generated with the following linear and additive

Table 1

Average estimates ("Est.") and empirical standard errors ("Ese.") of the mediated effects in simulation study 1. The sample size (n) was either 50, 200, or 1000. All results were rounded to two decimal places.

		$n = 50$		$n = 200$		$n = 1000$	
Effect	True value	Est.	Ese.	Est.	Ese.	Est.	Ese.
IE_1	0.00	0.00	0.09	-0.00	0.02	0.00	0.01
IE_2	0.06	0.06	0.48	0.06	0.23	0.06	0.10
DE	0.00	-0.00	0.30	0.00	0.14	-0.00	0.06

models corresponding to the causal diagram of Figure 3:

$$
\begin{aligned}
A &\sim \text{Bernoulli}(0.5) \\
C, U &\sim \mathcal{N}(0, 1) \\
M_s &= \alpha_s A + \alpha_{Cs} C + \alpha_{Us} U + \epsilon_s, \quad s = 1, 2, 3, \\
M_4 &= \alpha_{04} + \alpha_{C4} C + \alpha_{U4} U + \epsilon_4, \\
\epsilon_s &\sim \mathcal{N}(0, \sigma_s^2), \quad s = 1, 2, 3, 4, \\
Y &= \beta_2 M_2 + \beta_3 M_3 + \beta_4 M_4 + \beta_C C + \epsilon_Y, \\
\epsilon_Y &\sim \mathcal{N}(0, \sigma_Y^2).
\end{aligned}
$$

Under the (true) data-generating model, the unobserved variable U was a (baseline) confounder of all four mediators, and the observed variable C was a (baseline) confounder of all four mediators and the outcome. The variables A, C, U, and residuals $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_Y$ were mutually independent of each other. Following the path coefficient estimates in Figure 2 of van der Linden et al. (2015), the values of the (partial) regression coefficients in the data-generating model were set as $\alpha_1 = 12.80, \alpha_2 = 1.50, \alpha_3 = 1.92, \beta_2 = \beta_3 = 0.08, \beta_4 = 0.19$. The values of the remaining coefficients and residual variances were set to one for simplicity. There were no indirect effects via M_1 or via M_4 because M_1 did not affect Y, and M_4 did not depend on any of the other mediators or treatment and hence unaffected by treatment. The

interventional indirect effects via M_2 and M_3 corresponded to the causal effects along the paths for $A \to M_2 \to Y$ and $A \to M_3 \to Y$ respectively, and were identified by $\beta_2\alpha_2$ and $\beta_3\alpha_3$. For simplicity, the direct effect was zero.

The interventional (in)direct effects were estimated by fitting to each generated observed data: a (marginal) linear and additive mean model for each mediator, where each mediator depended only on treatment A and observed (baseline) covariate C as shown in (2), and the following (linear and additive) outcome mean model:

$$\mathrm{E}(Y|M_1, M_2, M_3, M_4, C) = \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_3 M_3 + \beta_4 M_4 + \beta_C C.$$

Under the assumed models, the interventional (in)direct effect estimators adopted the same analytical form as existing product-of-coefficient estimators under a parallel path model.

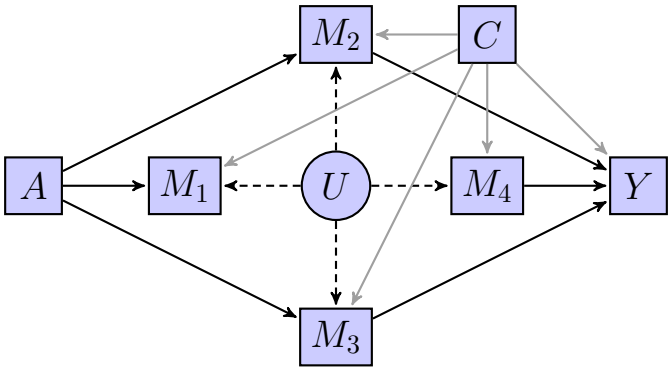

Figure 3. Causal diagram used to generate each simulated dataset with four possible mediators. Rectangular nodes denote observed variables, while round nodes denote unobserved variables. For visual clarity, edges emanating from the baseline covariate C are drawn in gray, while edges emanating from the hidden confounder U are drawn as broken lines.

Now suppose that substantive interest was in assessing indirect effects through the mediators by positing certain causal effects between the mediators, such as in the causal diagram of Figure 4. This particular path model was posited by van der Linden et al. (2015) as a "gateway belief model" representing "causal associations" between perceptions of scientific consensus, key beliefs in climate change, and support for

climate action. Treatment A was a randomly assigned consensus-message intervention, mediators M_1, M_2, M_3, and M_4 were perceived level of scientific consensus, belief that climate change is happening, belief in human causation (of climate change), and worry about climate change respectively, and outcome Y was support for public action. In particular, a causal structure among the variables was posited that assumed (i) the consensus-message intervention (A) affected only the level of perceived consensus (M_1), and no other variables, (ii) the level of perceived consensus (M_1) affected the key beliefs in climate change (M_2, M_3, M_4), (iii) belief that climate change is happening (M_2), and belief in human causation (M_3) subsequently affected worry about climate change (M_4), and (iv) support for public action (Y) was causally affected by the key beliefs in climate change (M_2, M_3, M_4), and neither level of perceived consensus (M_1) or the consensus-message intervention (A) directly.

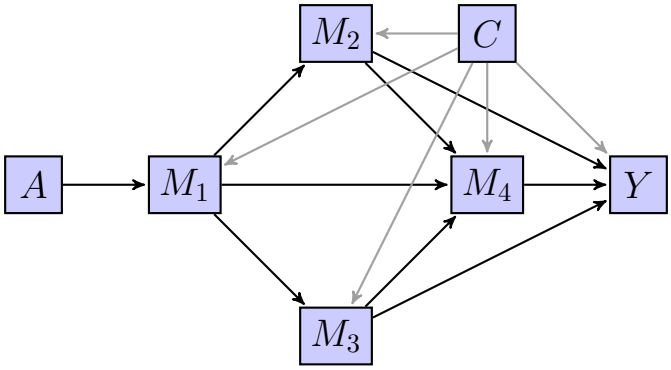

Figure 4. Causal diagram for positing indirect effects assuming the causal structure of a "gateway belief model" (van der Linden et al., 2015). For visual clarity, edges emanating from the baseline covariate C are drawn in gray.

A discussion of how to use causal diagrams to carefully represent causal mechanisms in theoretical models based on established scientific knowledge and prior careful experimentation is beyond the scope of this paper; we refer readers to Grosz et al. (2020). Instead, we will only consider whether estimates of the separate mediated effects along different assumed paths in Figure 4 can be unbiasedly estimated when the observed data was generated using the model in Figure 3. Denote the three-path mediated effect for the path $A \to M_1 \to M_s \to Y$ by PE_{1s} for $s = 2, 3, 4$, and denote the

"four-path" mediated effect for the path $A \to M_1 \to M_s \to M_4 \to Y$ by PE_{1s4} for $s = 2, 3$. The true values of all these mediated effects were zero under the data-generating model because M_1 did not exert any causal effects on the other mediators.

10000 observed datasets with sample size of either 50, 200, or 1000 were generated. Average estimates and empirical standard errors of the mediated effects from fitting the posited path model in Figure 4 to the generated data are displayed in Table 2. As expected, all the estimated mediated effects for the separate paths were empirically biased due to unobserved confounding of the mediators. The estimates could be either positively or negatively biased, and remained biased even at large sample sizes. In contrast, all the interventional indirect effects were unbiasedly estimated.

Table 2

Average estimates ("Est.") and empirical standard errors ("Ese.") of the mediated effects under the posited "gateway belief" path model in the simulation study. The sample size (n) was either 50, 200, or 1000. All results were rounded to two decimal places.

		$n = 50$		$n = 200$		$n = 1000$	
Effect	True value	Est.	Ese.	Est.	Ese.	Est.	Ese.
PE_{12}	0.00	0.14	0.22	0.14	0.10	0.14	0.05
PE_{13}	0.00	0.17	0.25	0.17	0.12	0.17	0.05
PE_{14}	0.00	-0.19	0.16	-0.19	0.07	-0.19	0.03
PE_{124}	0.00	0.11	0.10	0.11	0.04	0.11	0.02
PE_{134}	0.00	0.13	0.11	0.13	0.05	0.13	0.02
IE_1	0.00	0.03	1.79	0.02	0.82	-0.00	0.37
IE_2	0.12	0.12	0.22	0.12	0.10	0.12	0.04
IE_3	0.15	0.15	0.27	0.15	0.13	0.15	0.05
IE_4	0.00	-0.00	0.10	0.00	0.04	-0.00	0.02
DE	0.00	-0.03	1.71	-0.01	0.79	0.00	0.35

Simulation Study 3

This simulation study was used to demonstrate how to obtain estimators of the interventional indirect effects that allowed for mediator-mediator interaction terms in the outcome model. Each observed dataset was generated with the following linear models corresponding to the causal diagram of Figure 3:

$$
\begin{aligned}
A &\sim \text{Bernoulli}(0.5) \\
C, U &\sim \mathcal{N}(0, 1) \\
M_s &= \alpha_s A + \alpha_{Cs} C + \alpha_{Us} U + \alpha_{UAs}(U \times A) + \epsilon_s, \quad s = 1, 2, 3, \\
M_4 &= \alpha_{04} + \alpha_{C4} C + \alpha_{U4} U + \alpha_{UA4}(U \times A) + \epsilon_4, \\
\epsilon_s &\sim \mathcal{N}(0, \sigma_s^2), \quad s = 1, 2, 3, 4, \\
Y &= \beta_2 M_2 + \beta_3 M_3 + \beta_4 M_4 + \beta_{23} M_2 M_3 + \beta_C C + \epsilon_Y, \\
\epsilon_Y &\sim \mathcal{N}(0, \sigma_Y^2).
\end{aligned}
$$

The (true) data-generating model in this study differed from the previous study in two respects. The effect of the unobserved confounder U on each mediator was moderated by treatment A due to the $U - A$ interaction term in each mediator mean model. The effect of each (true) mediator on the outcome was moderated by the other (true) mediator due to the $M_2 - M_3$ interaction term in the outcome mean model. The true value of the coefficient for this interaction was set to e.g., $\beta_{23} = -(\beta_2 + \beta_3)$, simply to amplify possible biases that may arise when omitting this term in the fitted outcome model. For simplicity, there were no other interaction effects in the outcome model, and the direct effect was zero. The variables A, C, U, and residuals $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_Y$ were mutually independent of each other. The covariance between the mediators, induced by the unobserved confounder U, was $\Sigma_{sp}(A) = (\alpha_{Us} + \alpha_{UAs}A)(\alpha_{Up} + \alpha_{UAp}A)$ for $s \neq p$, and therefore depended on treatment when $\alpha_{UAs} \neq 0$ or $\alpha_{UAp} \neq 0$. Following the Online Supplemental Materials, the indirect effect due to the mediators' mutual dependence was identified by $\beta_{23}(\alpha_{U2}\alpha_{UA3} + \alpha_{U3}\alpha_{UA2} + \alpha_{UA2}\alpha_{UA3})$. The indirect effects via M_2 and via M_3 were respectively identified by $\{\beta_2 + \beta_{23}\,\mathrm{E}(M_3|A = 0)\}\alpha_2$ and $\{\beta_3 + \beta_{23}\,\mathrm{E}(M_2|A = 1)\}\alpha_3$.

We generated data under each of two different scenarios in turn. (I) In the first setting, the coefficients for the confounder-treatment interactions in the mediator models were set to zero, i.e., $\alpha_{UAs} = 0, s = 1, \ldots, 4$, so that the mediators' covariance did not differ by treatment. The indirect effect due to the mutual dependence of the mediators was therefore zero. However, the indirect effect estimates assuming a parallel path model would be biased due to omitting the mediator-mediator interaction terms in the fitted outcome model. (II) In the second setting, set the coefficients $\alpha_{UAs} = 0.1, s = 1, \ldots, 4$, so that the indirect effect due to the mediators' mutual dependence was non-zero. The remaining coefficients and residual variances for both scenarios were set to the same values as in Simulation Study 1.

The interventional indirect effects were estimated by fitting to each generated data a (marginal) mean model for each mediator, where each mediator depended only on treatment A and observed (baseline) covariate C as shown in (2), and the following outcome model that included all mediator-mediator interaction terms:

$$\begin{aligned}
&\mathrm{E}(Y|M_1, M_2, M_3, M_4, C) \\
&= \beta_0 + \beta_A A + \beta_1 M_1 + \beta_2 M_2 + \beta_3 M_3 + \beta_4 M_4 + \beta_C C \\
&\quad + \beta_{12} M_1 M_2 + \beta_{13} M_1 M_3 + \beta_{14} M_1 M_4 + \beta_{23} M_2 M_3 + \beta_{24} M_2 M_4 + \beta_{34} M_3 M_4.
\end{aligned}$$

For simplicity, no treatment-mediator or treatment-mediator-mediator interaction terms were included. The indirect effect estimators, as shown in the Online Supplemental Materials, were respectively:

$$\begin{aligned}
\mathrm{IE}_1 &= \{\beta_1 + \beta_{12}\,\mathrm{E}(M_2|A=0) + \beta_{13}\,\mathrm{E}(M_3|A=0) + \beta_{14}\,\mathrm{E}(M_4|A=0)\}\delta_1, \\
\mathrm{IE}_2 &= \{\beta_2 + \beta_{12}\,\mathrm{E}(M_1|A=1) + \beta_{23}\,\mathrm{E}(M_3|A=0) + \beta_{24}\,\mathrm{E}(M_4|A=0)\}\delta_2, \\
\mathrm{IE}_3 &= \{\beta_3 + \beta_{13}\,\mathrm{E}(M_1|A=1) + \beta_{23}\,\mathrm{E}(M_2|A=1) + \beta_{34}\,\mathrm{E}(M_4|A=0)\}\delta_3, \\
\mathrm{IE}_4 &= \{\beta_4 + \beta_{14}\,\mathrm{E}(M_1|A=1) + \beta_{24}\,\mathrm{E}(M_2|A=1) + \beta_{34}\,\mathrm{E}(M_3|A=1)\}\delta_4, \\
\mathrm{IE}_{mu} &= \sum_{\substack{k,l=1,\\ k<l}}^{4} \beta_{kl}\{\Sigma_{kl}(1) - \Sigma_{kl}(0)\}.
\end{aligned}$$

Under the assumed outcome model, different indices for the mediators lead to different (definitions and) estimators of the indirect effects. In particular, swapping the labels for

M_2 and M_3 would lead to an estimator for the indirect effect via M_2 where M_3 is fixed at its mean value under treatment instead; i.e., $\{\beta_2 + \beta_{12}\,\mathrm{E}(M_1|A=1) + \beta_{23}\,\mathrm{E}(M_3|A=1) + \beta_{24}\,\mathrm{E}(M_4|A=0)\}\delta_2$. Similarly, the estimator for the indirect effect via M_3 would fix M_2 at its mean value under control instead; i.e., $\{\beta_3 + \beta_{13}\,\mathrm{E}(M_1|A=1) + \beta_{23}\,\mathrm{E}(M_2|A=0) + \beta_{34}\,\mathrm{E}(M_4|A=0)\}\delta_3$. A sensitivity analysis where the mediator indices are permuted, and the indirect effects estimated under each permutation, is carried out and discussed in the applied example, and hence not considered here.

10000 observed datasets with sample size of either 50, 200, or 1000 were generated, and estimators of the indirect effects from a fitted outcome model that either included all mediator-mediator interactions, or excluded such interactions in a parallel path model, were obtained. The model syntax in `R` describing the fitted models and the indirect effect estimators are provided online as part of the `R` scripts for this article[2]. Average estimates and empirical standard errors of the (in)direct effects are displayed in Table 3. As expected, the (in)direct effects assuming a parallel path model were empirically biased under both scenarios even in large samples. Furthermore, the indirect effect via a mediator may be of the opposite sign as the true effect. In contrast, estimates of the interventional indirect effects allowing for mediator-mediator interaction terms in the outcome model were empirically unbiased under both scenarios.

Summary of simulation studies

The results of Simulation Study 1 empirically demonstrated that the parallel path model can be used to unbiasedly estimate the interventional indirect and direct effects proposed in this article, even when the mediators can be causally ordered, and there is hidden confounding among the mediators. However, this is predicated on the outcome model being correctly specified in that there are no mediator-mediator interaction terms. Unbiased estimation does not require the mediators to be causally independent, as implied in the parallel path model; in fact, the (marginal) mean model (2) is used

[2] `https://github.com/wwloh/disentangle-multiple-mediators`

Table 3

Average estimates ("Est.") and empirical standard errors ("Ese.") of the interventional indirect and direct effects under each fitted model in the simulation study. An interaction between the mediators' unmeasured confounder U and the treatment A was either absent ("$U - A$ int."=False) or present ("$U - A$ int."=True) when generating the data. The assumed outcome model either included all mediator-mediator interactions ("$M - M$ int."), or main effects only ("Parallel"). The sample size was either 50, 200, or 1000. All results were rounded to two decimal places.

				$n = 50$		$n = 200$		$n = 1000$	
Fitted model	Effect	$U - A$ int.	True value	Est.	Ese.	Est.	Ese.	Est.	Ese.
$M - M$ int.	IE_1	False	0.00	-0.01	1.93	0.01	0.86	-0.00	0.37
		True	0.00	0.00	1.94	-0.01	0.85	0.00	0.36
	IE_2	False	0.12	0.13	0.48	0.12	0.19	0.12	0.08
		True	0.12	0.13	0.46	0.12	0.18	0.12	0.08
	IE_3	False	-0.31	-0.32	0.44	-0.31	0.19	-0.31	0.08
		True	-0.31	-0.34	0.47	-0.31	0.20	-0.31	0.09
	IE_4	False	0.00	0.00	0.12	0.00	0.04	0.00	0.02
		True	0.00	-0.00	0.15	0.00	0.06	-0.00	0.02
	IE_{mu}	False	0.00	0.00	0.19	-0.00	0.09	-0.00	0.04
		True	-0.48	-0.48	0.31	-0.48	0.15	-0.48	0.06
	DE	False	0.00	0.01	1.79	-0.01	0.81	0.00	0.35
		True	0.00	-0.01	1.80	0.01	0.79	-0.00	0.34
Parallel	IE_1	False	0.00	0.01	2.14	0.02	1.02	-0.01	0.45
		True	0.00	-0.45	2.52	-0.48	1.20	-0.48	0.53
	IE_2	False	0.12	-0.11	0.28	-0.11	0.13	-0.11	0.06
		True	0.12	-0.18	0.35	-0.17	0.16	-0.17	0.07
	IE_3	False	-0.31	-0.08	0.36	-0.08	0.17	-0.08	0.07
		True	-0.31	-0.16	0.43	-0.15	0.20	-0.15	0.09
	IE_4	False	0.00	0.00	0.10	0.00	0.04	0.00	0.02
		True	0.00	-0.01	0.13	0.00	0.05	-0.00	0.02
	DE	False	0.00	-0.01	2.04	-0.02	0.98	0.00	0.43
		True	0.00	0.12	2.40	0.14	1.15	0.13	0.51

precisely so that the interventional indirect effects are agnostic to the underlying causal dependence among the mediators. The results of Simulation Study 2 showed that the estimated mediated effects along separate paths were empirically biased when the

statistical associations between the mediators were (incorrectly) assumed to be causal effects. In contrast, the existing product-of-coefficient indirect effect estimates under a parallel path model were unbiased. When the (true) mediators moderated each other's effects on the outcome in Simulation Study 3, the indirect effect estimates under a fitted parallel path model were empirically biased even in large samples. The biases were due to omitting mediator-mediator interaction terms in the (misspecified) outcome model fitted to the data. Hence allowing for (all) mediator-mediator interaction terms in the (fitted) outcome model yielded unbiased estimates of the interventional indirect effects.

We have focused on unbiased estimation of the proposed interventional indirect effects in the simulation studies. Inference using a non-parametric bootstrap as suggested in this paper is straightforward and builds on established bootstrap theory and procedures. For this reason, we have chosen not to evaluate the statistical efficiency of the resulting confidence intervals or hypothesis tests. Comparisons of the (non-parametric) bootstrap with other (parametric) methods, such as the Monte Carlo method (MacKinnon et al., 2004; Preacher & Selig, 2012), are thus deferred to future work.

Application

The proposed estimation procedure was illustrated using a publicly-available data set from a randomized study assessing the effect of (non-)political inclusion on political prejudice that was possibly mediated by six different mediators (Voelkel et al., 2019). The data set is available as part of a preregistered study via the Open Science Framework [3]. The goal of the study was to assess the causal effect of either political inclusion or non-political inclusion versus control on momentary prejudice toward the political outgroup. The sample consisted of college freshmen from a large university in the Netherlands who received course credit in a psychology course for their participation. Participants were randomly assigned to one of three conditions: political inclusion, non-political inclusion, or control. For the purposes of illustration, we

[3] https://osf.io/jcmmp/?view_only=3af8cb6b1f2845b1ba3fd69cb0b89585

considered only the 183 participants assigned to either political inclusion ($A = 1$) or control ($A = 0$). In the treatment group, participants' political inclusion experiences were manipulated using an online political discussion. In the control group, participants experienced a neutral scenario where no discussion (political or non-political) occurred, and they were only asked to fill in a questionnaire. The outcome Y (prejudice) was an average of three items: dislike of, social distance from, and perceived immorality of, the participant's political outgroup. Larger values indicated higher levels of prejudice.

To understand how political inclusion affected prejudice, the authors of the study considered six possible mediators of the causal relationship between political inclusion and prejudice: satisfaction of the need to belong (M_1), satisfaction of the need for self-esteem (M_2), satisfaction of the need for control (M_3), satisfaction of the need for meaningful existence (M_4), perceived worldview dissimilarity of the political outgroup (M_5), and perceived fairness of the political outgroup (M_6). In addition, we considered political ideology ("Ideology"), age in years ("Age"), and gender ("Gender") as (baseline) confounders of the mediator-outcome relation for all the mediators. Summaries of the variables for each treatment group are provided in Table 4. We adjusted for the baseline covariates toward satisfying the assumptions (A1)–(A3) needed to identify the interventional direct and indirect effects.

The total effect of treatment was estimated by regressing prejudice on treatment, political ideology, gender and age (without any mediators). The estimated total effect of the political inclusion manipulation (versus control) was an average change in prejudice by -0.076 (95% confidence interval (CI) $= (-0.135, -0.017)$). All standard errors and 95% (percentile) confidence intervals were constructed using 1000 (non-parametric) bootstrap samples. To estimate the interventional (in)direct effects, the following mediator and outcome models were fitted to the observed data:

$$\mathrm{E}(M_s|A, C) = \delta_{0s} + \delta_s A + \delta_{C1s}\mathrm{Ideology} + \delta_{\mathrm{C2s}}\mathrm{Age} + \delta_{\mathrm{C3s}}\mathrm{Gender}, \quad \mathrm{s} = 1, \ldots, 6;$$

$$\mathrm{E}(Y|A, M_1, \ldots, M_6, C) = \beta_0 + \beta_A A + \sum_{s=1}^{6}(\beta_s + \beta_{As}A)M_s + \sum_{\substack{k,l=1,\\ k<l}}^{6}(\beta_{kl} + \beta_{Akl}A)M_k M_l$$

$$+ \beta_{C1}\mathrm{Ideology} + \beta_{\mathrm{C2}}\mathrm{Age} + \beta_{\mathrm{C3}}\mathrm{Gender}.$$

Table 4

Sample means and standard deviations (in brackets) for the baseline confounders, mediators and outcome for each treatment group in the applied example.

Treatment group	$A = 0$	$A = 1$
Number of participants	95	88
Ideology	-0.51 (0.9)	-0.43 (0.7)
Gender	0.73 (0.4)	0.76 (0.4)
Age	20.0 (2.3)	20.1 (2.5)
M_1 (belong)	0.81 (0.2)	0.69 (0.2)
M_2 (self-esteem)	0.50 (0.2)	0.50 (0.2)
M_3 (control)	0.37 (0.2)	0.30 (0.2)
M_4 (meaningful existence)	0.83 (0.2)	0.81 (0.2)
M_5 (worldview dissimilarity)	0.65 (0.2)	0.56 (0.2)
M_6 (fairness)	0.48 (0.2)	0.62 (0.2)
Y	0.46 (0.2)	0.38 (0.2)

Each mediator depended only on treatment, political ideology, age, and gender. The outcome model included all treatment-mediator, mediator-mediator, and treatment-mediator-mediator interaction terms. Closed form expressions for the interventional direct and indirect effect estimators as functions of the outcome and mediator model parameters are provided in the Online Supplemental Materials. The estimated effects using the observed data are shown in Table 5.

The estimated interventional direct effect was $-0.013\,(95\%\,\text{CI} = (-0.088, 0.085))$, suggesting that politically included individuals had lower prejudice (than if assigned to the control condition), when holding the (counterfactual) distributions of all mediators (given ideology, age, and gender) fixed under those of the control condition. There was only one mediator with a statistically significant indirect effect. The estimated indirect effect via fairness was $-0.077\,(95\%\,\text{CI} = (-0.139, -0.021))$, so that shifting the

Table 5

Interventional (in)direct effect estimates, bootstrap standard errors ("SE") and 95% bootstrap (percentile) confidence intervals ("CI") for the applied example. All results were rounded to three decimal places.

Interventional effect	Estimate	Bootstrap SE	95% CI
Indirect effect via M_1 (belong)	0.016	0.017	(-0.016, 0.053)
Indirect effect via M_2 (self-esteem)	0.000	0.007	(-0.015, 0.015)
Indirect effect via M_3 (control)	-0.002	0.012	(-0.029, 0.020)
Indirect effect via M_4 (meaningful existence)	-0.003	0.008	(-0.024, 0.008)
Indirect effect via M_5 (worldview dissimilarity)	-0.006	0.011	(-0.029, 0.015)
Indirect effect via M_6 (fairness)	-0.077	0.030	(-0.139, -0.021)
Indirect effect due to mutual dependence	0.009	0.025	(-0.048, 0.051)
Direct effect	-0.013	0.044	(-0.088, 0.085)
Total effect	-0.076	0.031	(-0.135, -0.017)

(counterfactual) distribution of fairness from the political inclusion manipulation to that under control resulted in lower prejudice on average, while holding treatment and the distributions of all other mediators fixed. The estimated indirect effects via the remaining mediators were not statistically significant at 5%. For example, the indirect effect of changing the (counterfactual) distribution of the need to belong from the politically included group to the control group (holding treatment and distributions of all other mediators fixed) was an increase in prejudice by 0.016 on average, but the 95% CI of $(-0.016, 0.053)$ included zero. The indirect effect due to the mediators' mutual dependence on one another was $0.009\,(95\%\,\text{CI} = (-0.048, 0.051))$, suggesting a positive (but not statistically significant) average effect on prejudice when changing the mediators' covariance from the political inclusion manipulation to that under control.

The mediator indices were arbitrarily labelled and used merely to distinguish the mediators for statistical analysis when considering simultaneous mediators, and did not

represent or imply any causal ordering of the mediators. Under the assumed outcome model where the effect of each mediator on the outcome was moderated by treatment, or another mediator, or both, a different permutation of the (arbitrary) mediator labels corresponded to a different decomposition of the joint indirect effect, whereby the indirect effects via each mediator held the (counterfactual) mediator distributions fixed at different hypothetical treatment levels. We carried out a sensitivity analysis, by considering each of the $6! = 720$ possible permutations of the six mediators in turn, and calculated the indirect effects (and the 95% CIs) under each permutation. Each permutation of the indices thus implies a different decomposition of the joint indirect effect, and subsequently, may result in different estimates of the separate indirect effects. Details on enumerating all possible permutations, especially with a large number of mediators, when implementing the sensitivity analysis are provided online as part of the R scripts for this article[4]. The minimum and maximum estimates (and bounds of the 95% CIs) across all the permutations are shown in Table 6. Inference for the indirect effects was unchanged across the different decompositions resulting from different permutations of the mediator indices. We again emphasize that the conceptual interpretations of the interventional indirect effects via each mediator using the causal paths in the underlying causal structure remain the same regardless of the chosen decomposition. For example, the estimated indirect effect via fairness was significantly different from zero (statistically at the 5% level) regardless of the chosen decomposition. Conversely, the 95% CIs for the indirect effect via the need to belong always included zero. In general, when changing the mediators' labels leads to conflicting inferences about the indirect effects, theoretical knowledge may be used to determine the most scientifically relevant decomposition (of the joint indirect effect) and the implied definitions of the indirect effects. These results suggested that the total diminishing effect of political inclusion on prejudice was primarily explained by the mediating effect through perceived fairness of the political outgroup.

[4] https://github.com/wwloh/disentangle-multiple-mediators

Table 6

Interventional (in)direct effect estimates and 95% bootstrap (percentile) confidence intervals ("CI") for the applied example. The minimum ("min.") and maximum ("max.") estimates, and 95% CI lower and upper bounds, across all $6! = 720$ *possible permutations of the mediator indices are presented. All results were rounded to three decimal places.*

	Estimate		95% CI	
Interventional indirect effect	Min.	Max.	Min. (lower)	Max. (upper)
Belong	0.011	0.025	-0.045	0.073
Self-esteem	0.000	0.000	-0.019	0.021
Control	-0.012	0.012	-0.044	0.053
Meaningful existence	-0.004	-0.001	-0.034	0.014
Worldview dissimilarity	-0.014	0.006	-0.049	0.039
Fairness	-0.094	-0.074	-0.156	-0.013

Discussion

Recommendations for multiple mediation analysis

When there are multiple or competing mediators on the causal pathway from treatment to outcome, path analysis is commonly used to disentangle the indirect effects transmitted along causal path(s) through each mediator. But indirect or mediated effects along separate paths traversing several linked mediators are valid only when (i) the causal dependence among the mediators is correctly specified, and (ii) there is no unobserved confounding of the mediators. When these assumptions are violated, estimates of mediated effects along separate paths can be severely biased; the biases were demonstrated empirically using a simulation study in this article.

When scientific interest is in inferring indirect effects transmitted through each distinct mediator, the aforementioned assumptions are avoidable by using the interventional indirect effects proposed in this article. Intuitively, the causal dependence

among the mediators is left unspecified, by focusing on only the marginal mean model for each mediator that captures the overall treatment effect (on that mediator). The interventional indirect effect via a mediator of interest is interpreted as the combined effect along all (unknown) causal pathways from treatment to outcome that intersect that mediator and any others that causally precede the mediator in question. Interventional indirect effects are therefore agnostic to the (unknown) causal structure of the mediators. Estimators assuming linear and additive mean models for the mediators and the outcome, such as (1) and (2), imply the same analytical form as prevailing indirect effects using parallel path models. But when treatment affects the mediators' covariance, and the effect of each mediator on the outcome is moderated by the other mediator(s), such an indirect effect that is due to the mediators' mutual dependence on one another cannot be attributed to any mediator alone. We proposed new estimators of interventional indirect effects under such settings that exploit the mediators' covariance under the assumed linear mean models, thus simplifying closed form solutions when there are more than two mediators. Unbiased estimators of the interventional effects can be straightforwardly obtained using conventional OLS estimation methods, and are robust against an incorrectly specified causal structure of the mediators, and unobserved confounding among the mediators.

Practical considerations for applied researchers

It is important to note that we are not advocating researchers avoid multiple mediation analyses using "serial" mediation models (Hayes, 2018) whose causal structure among the mediators represent theoretical models. On the contrary, we encourage mediation analysis using causal structures that are grounded in established scientific knowledge or prior thoughtful experimentation; see e.g., Pek & Hoyle (2016) and Fiedler et al. (2018) for the single mediator setting. For example, in the single mediator setting, experimentally manipulating the mediator to examine its effect on the outcome can lend empirical support for the posited causal effects (Spencer et al., 2005). In principle, such methods may be extended to multiple mediator settings toward

establishing the mediators' causal structure, by experimentally manipulating each mediator in turn to examine its causal effect(s) on other mediator(s). Valid inference of causal effects using causal diagrams that carefully represent theoretical models is invaluable toward understanding and reasoning of underlying causal mechanisms (Grosz et al., 2020). But often in practice the plausibility of an assumed (path) model is evaluated based solely on statistical associations or observed goodness-of-fit criteria. Practitioners of multiple mediation analysis should therefore be cognizant of the implied causal assumptions when inferring causal effects linking different mediators and the implications when the assumptions are violated.

Applied researchers across different areas in psychology (often) seek to explore attributing the total effect of a treatment on an outcome to each of multiple possible mediators, without having to specify (arbitrary) causal effects among the mediators. Recent examples include Bergfeld & Chiu (2017), Brooks et al. (2019), Irwin et al. (2019), Schroeder et al. (2019), and Ren et al. (2020), among many others. Possible reasons may be that the mediators were contemporaneously measured, or there was simply insufficient theoretical or experimental justification to warrant positing a causal structure among the mediators. The framework proposed in this article is particularly well-suited for such (common) research settings, because interventional indirect effects have the benefit of being well-defined and possessing the same interpretation, regardless of the mediators' underlying causal structure. Existing indirect effects using prevailing parallel path models are unbiased estimators of the interventional indirect effects - without necessarily requiring the absence of causal effects among the mediators as implied by the fitted model - under certain assumptions. When these assumptions fail to hold, such as when the (true) effect of each mediator on the outcome is moderated by another mediator, incorrectly fitting a parallel path model can lead to biased estimates of the interventional indirect effects. We therefore recommend applied researchers conducting multiple mediation analysis to consider outcome models with mediator-mediator interactions terms when feasible, and investigate the indirect effect due to the mediators' mutual dependence, which may reveal part of the treatment effect

that simply cannot be attributable to any mediator alone.

One of the assumptions (A2) required to identify the interventional (in)direct effects presented in this paper is that there be no hidden common causes of the mediators and outcome. Future research could include extending sensitivity analyses to unobserved confounding of the mediator-outcome relations for a single mediator (Cox et al., 2013; Fritz et al., 2016; Hong et al., 2018; Liu & Wang, 2020) to the multiple mediator setting. The path analysis approach can be extended to accommodate latent mediators or outcome, or both, by including latent variable models; see e.g., Loeys et al. (2014) and Loh, Moerkerke, Loeys, Poppe, et al. (2020). VanderWeele & Tchetgen Tchetgen (2017) proposed interventional indirect effects for mediation analysis with longitudinal data under a formal causal framework, and described estimators using sets of linear structural equation models under the so-called "Autoregressive Model III" of MacKinnon (2008). The interventional indirect effects defined in this paper have focused on a binary treatment, continuous mediators, and a continuous outcome. Continuous treatments may be accommodated by extending the interventional effect models proposal (Loh, Moerkerke, Loeys, & Vansteelandt, 2020a) to parameterize a linear treatment effect in future work. 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. Assuming non-linear models for the means of the mediators, or the outcome, or both, will generally lead to different estimators of the interventional indirect effects defined in this paper. For example, when the outcome is binary and rare, and a logistic outcome model is assumed, the indirect effect due to the mediators' mutual dependence (7) is non-zero only if (i) the main effect of each mediator on the outcome is non-zero, and (ii) the covariance of the (continuous) mediators is affected by treatment (Loh, Moerkerke, Loeys, & Vansteelandt, 2020b). Unlike the linear setting assumed in this paper, estimating this indirect effect may not require a mediator-mediator interaction term in the (non-linear) outcome model. More

general estimation strategies for non-continuous mediators, or outcome, or both, are described in Vansteelandt & Daniel (2017) and Loh, Moerkerke, Loeys, & Vansteelandt (2020a). Estimation requires (correctly) specifying a model for the joint distribution of the mediators and a (mean) model for the outcome, and proceeds via Monte Carlo integration.

References

Albert, J. M., Cho, J. I., Liu, Y., & Nelson, S. (2019). Generalized causal mediation and path analysis: Extensions and practical considerations. *Statistical Methods in Medical Research*, *28*(6), 1793–1807. doi: 10.1177/0962280218776483

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

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.

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

Bergfeld, J. R., & Chiu, E. Y. (2017). Mediators in the relationship between minority stress and depression among young same-sex attracted women. *Professional Psychology: Research and Practice*, *48*(5), 294–300. doi: 10.1037/pro0000155

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 37–69. doi: 10.2307/271028

Brooks, M., Graham-Kevan, N., Robinson, S., & Lowe, M. (2019). Trauma characteristics and posttraumatic growth: The mediating role of avoidance coping, intrusive thoughts, and social support. *Psychological Trauma: Theory, Research, Practice, and Policy*, *11*(2), 232–238. doi: 10.1037/tra0000372

Brown, R. L. (1997). Assessing specific mediational effects in complex theoretical models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 142–156. doi: 10.1080/10705519709540067

Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013, Oct). Sensitivity plots for confounder bias in the single mediator model. *Eval Rev*, *37*(5), 405–431. doi: 10.1177/0193841X14524576

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

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

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

Fox, J. (1980). Effect analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Methods & Research*, *9*(1), 3-28. doi: 10.1177/004912418000900101

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

Greene, V. L. (1977). An algorithm for total and indirect causal effects. *Political Methodology*, *4*(4), 369–381. Retrieved from `https://www.jstor.org/stable/25791510`

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. Retrieved from https://doi.org/10.1177/1745691620921521 (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.

Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist*, *64*(1), 19–54.

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. Retrieved from https://doi.org/10.3102/1076998617749561 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., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71. doi: 10.2307/41058997

Irwin, A., Li, J., Craig, W., & Hollenstein, T. (2019). The role of shame in chronic peer victimization. *School Psychology*, *34*(2), 178–186. doi: 10.1037/spq0000280

Jackson, J. W., & VanderWeele, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, *29*(6), 825–835. doi: 10.1097/EDE.0000000000000901

Lange, T., Rasmussen, M., & Thygesen, L. C. (2013). Assessing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, *179*(4), 513–518. doi: 10.1093/aje/kwt270

Lin, S.-H., & VanderWeele, T. (2017). Interventional approach for path-specific effects. *Journal of Causal Inference*, *5*(1). doi: 10.1515/jci-2015-0027

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. *Psychol Methods*. doi: 10.1037/met0000345

Loeys, T., Moerkerke, B., Raes, A., Rosseel, Y., & Vansteelandt, S. (2014). Estimation of controlled direct effects in the presence of exposure-induced confounding and latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 396–407. doi: 10.1080/10705511.2014.915372

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. (2020a). Heterogeneous indirect effects for multiple mediators using interventional effect models. *arXiv e-prints*, arXiv:1907.08415.

Loh, W. W., Moerkerke, B., Loeys, T., & Vansteelandt, S. (2020b). Non-linear mediation analysis with high-dimensional mediators whose causal structure is unknown. *arXiv preprint arXiv:2001.07147*.

Lok, J. J. (2019, Mar). Causal organic direct and indirect effects: closer to Baron and Kenny. *arXiv preprint arXiv:1903.04697*.

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., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128.

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

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. (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

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

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.

Quynh Nguyen, T., Schmid, I., & Stuart, E. A. (2019, Apr). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *arXiv preprint arXiv:1904.08515*.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ren, D., Wesselmann, E. D., & van Beest, I. (2020). Seeking solitude after being ostracized: A replication and beyond. *Personality and Social Psychology Bulletin*. Retrieved from `https://doi.org/10.1177/0146167220928238` (PMID: 32515281) doi: 10.1177/0146167220928238

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

Schroeder, J., Risen, J. L., Gino, F., & Norton, M. I. (2019). Handshaking promotes deal-making by signaling cooperative intent. *Journal of Personality and Social Psychology*, *116*(5), 743–768. doi: 10.1037/pspi0000157

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

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. (2017). Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, *186*(2), 184–193. doi: 10.1093/aje/kwx051

Taguri, M., Featherstone, J., & Cheng, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical Methods in Medical Research*, *27*(1), 3-19. doi: 10.1177/0962280215615899

Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, *11*(2), 241–269. doi: 10.1177/1094428107300344

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015, 02). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLOS ONE*, *10*(2), 1-8. doi: 10.1371/journal.pone.0118489

VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 917–938. doi: 10.1111/rssb.12194

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.0000000000000596

Voelkel, J. G., Ren, D., & Brandt, M. J. (2019). Political inclusion reduces political prejudice. *PsyArXiv*. doi: 10.31234/osf.io/dxwpu

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