



## A Regression Framework for Causal Mediation Analysis with Applications to Behavioral Science

Christina T. Saunders & Jeffrey D. Blume

To cite this article: Christina T. Saunders & Jeffrey D. Blume (2019): A Regression Framework for Causal Mediation Analysis with Applications to Behavioral Science, Multivariate Behavioral Research, DOI: [10.1080/00273171.2018.1552109](https://doi.org/10.1080/00273171.2018.1552109)

To link to this article: <https://doi.org/10.1080/00273171.2018.1552109>



Published online: 01 Apr 2019.



Submit your article to this journal [↗](#)



View Crossmark data [↗](#)



# A Regression Framework for Causal Mediation Analysis with Applications to Behavioral Science

Christina T. Saunders  and Jeffrey D. Blume

Department of Biostatistics, Vanderbilt University

## ABSTRACT

We introduce and extend the classical regression framework for conducting mediation analysis from the fit of only one model. Using the essential mediation components (EMCs) allows us to estimate causal mediation effects and their analytical variance. This single-equation approach reduces computation time and permits the use of a rich suite of regression tools that are not easily implemented on a system of three equations. Additionally, we extend this framework to non-nested mediation systems, provide a joint measure of mediation for complex mediation hypotheses, propose new visualizations for mediation effects, and explain why estimates of the total effect may differ depending on the approach used. Using data from social science studies, we also provide extensive illustrations of the usefulness of this framework and its advantages over traditional approaches to mediation analysis. The example data are freely available for download online and we include the R code necessary to reproduce our results.

## KEYWORDS

Causal modeling; mediation; direct and indirect effects

## Introduction


Psychologists and social scientists concerned with dynamic relations and the mechanisms by which an exposure affects an outcome have been studying mediation processes for decades (Alwin & Hauser, 1975; Baron & Kenny, 1986; Woodworth, 1928). A large body of literature and a variety of methods exist for conducting mediation analyses, e.g., the Baron-Kenny causal steps approach (Baron & Kenny, 1986; Zhao, Lynch, & Chen, 2010), the structural equation modeling (SEM) approach (Gunzler, Chen, Wu, & Zhang, 2013), the potential outcomes (PO) approach (Imai, Keele, & Tingley, 2010; Pearl, 2001; Robins & Greenland, 1992; VanderWeele, 2015) and others (Gelfand, Mensinger, & Tenhave, 2009; MacKinnon, 2008; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Preacher, 2015). This article presents key extensions to the single-model mediation framework recently proposed by Saunders and Blume (2017). These extensions are essential for applying the proposed framework to complex mediation models with multiple mediators, nonlinear mediation effects, and exposure-mediator interactions.

For behavioral research scientists, epidemiologists, and statisticians implementing mediation analyses, the

gap between existing frameworks can be hard to navigate. To help bridge this gap, we provide a crosswalk in Table 1 and important background information in Appendix A.1. Whereas Saunders and Blume (2017) is a technically focused paper, the intent of this manuscript is to introduce the new method to an audience accustomed to mediation analysis in socialscience research and to provide essential extensions that accommodate complex models. Instead of fitting separate regression equations and aggregating coefficients to estimate mediation effects, the proposed single-model approach uses a simple formula to estimate the portion eliminated (PE) from the fit of one model. The PE is a generalization of the difference of coefficients approach for estimating the indirect effect that measures the reduction in the total effect when indirect paths are blocked. This measure is important because it pertains to effects of interventions that could actually be carried out and requires fewer identification assumptions than the natural indirect effect (NIE) (Naimi, Kaufman, & MacLehose, 2014; Pearl, 2012a; VanderWeele, 2013; VanderWeele, 2015). For example, when studying how an intervention (such as seeing a psychologist for help with depression) can

**CONTACT** Christina T. Saunders  [christina.t.saunders@vumc.org](mailto:christina.t.saunders@vumc.org)  Department of Biostatistics, Vanderbilt University, 2525 West End Ste. 11000, Nashville, TN 37203, USA.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hmbr](http://www.tandfonline.com/hmbr).

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2019 Taylor & Francis Group, LLC

**Table 1.** Nomenclature and definitions of causal mediation effects for exposure  $X$ , mediator  $M$ , covariates  $C$ , and outcome  $Y$ . Effects compare the value  $X = x$  to  $X = x_0$  (the referent exposure).

SEM Name	PO Name(s)	PO Definition	Regression-based estimand for continuous $X$	for (0,1) $X$
Total effect <sup>b</sup> of $X$		$Y(x) - Y(x_0)$	$(\beta_X + \beta_X^*)(x - x_0)$	$\beta_X^*$
Direct effect <sup>b</sup> of $X$		$Y(x, m) - Y(x_0, m)$	$(\beta_X + \beta_{XM})(x - x_0)$	$(\beta_X + \beta_{XM})$
a) Set $M = m$				
b) Set $M = E[M x_0, c]$		$Y(x, M(x_0)) - Y(x_0, M(x_0))$	$(\beta_X + \beta_{XM}E[M x_0, c])(x - x_0) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x_0 + \alpha_X c))(x - x_0)$	$\beta_X + \beta_{XM}(\alpha_0 + \alpha_X c)$
		Pure direct effect <sup>d</sup>		
		Average direct effect (control) <sup>e</sup>		
c) Set $M = E[M x, c]$		$Y(x, M(x)) - Y(x_0, M(x))$	$(\beta_X + \beta_{XM}E[M x, c])(x - x_0) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x + \alpha_X c))(x - x_0)$	$\beta_X + \beta_{XM}(\alpha_0 + \alpha_X c)$
		Total direct effect <sup>d</sup>		
		Average direct effect (treatment) <sup>e</sup>		
Indirect effect <sup>b</sup> of $X$		$Y(x_0, M(x)) - Y(x_0, M(x_0))$	$(\beta_M + \beta_{MX})(E[M x, c] - E[M x_0, c]) = (\beta_M + \beta_{MX}\alpha_X)(x - x_0)$	$\beta_M \alpha_X$
a) Set $X = x_0$				
b) Set $X = x$		$Y(x, M(x)) - Y(x_0, M(x_0))$	$(\beta_M + \beta_{MX})(E[M x, c] - E[M x_0, c]) = (\beta_M + \beta_{MX}\alpha_X)(x - x_0)$	$(\beta_M + \beta_{MX})\alpha_X$
		Total indirect effect <sup>d</sup>		
		Average causal mediation effect (treat) <sup>e</sup>		
Portion eliminated		$Y(x) - Y(x_0) - (Y(x, m) - Y(x_0, m))$	$(\beta_X^* + \beta_X^*)(x - x_0)$	$\beta_X^* - (\beta_X + \beta_{XM})$

<sup>a</sup>The marginal model  $E[Y|X, C]$  is obtained by  $E_{M|X, C}[Y|X, M, C]$ . Notice that the marginal model has an  $X^2$  term and the full model does not.

<sup>b</sup>As discussed in the Interactions and moderated mediation section, the maximum likelihood fit of the total effect  $(\beta_X^* + \beta_X^*)(x - x_0)$  does not always equal the sum of the estimated natural direct and indirect effects. If  $X$  is a binary variable, then fitting the model  $E[Y|X, C]$  will drop the  $X^2$  term and the total effect estimate  $\hat{\beta}_X^*$  will equal the sum of the natural direct and indirect effects.

<sup>c</sup>T. J. VanderWeele (2015).

<sup>d</sup>Robins and Greenland (1992).

<sup>e</sup>Imai et al. (2010).

prevent adverse mental health outcomes (such as suicide), the PE measures the maximum preventive effect of said therapy on the mediating pathways.

There are several practical benefits to using the single-model framework for mediation analysis. First, this framework yields a model-based formula for the variance of the PE, eliminating the need to rely on approximations or resampling methods. The efficiency gains can be quite large. Example: the simple mediation model section discusses an example where the model-based standard error is more than five times smaller than the approximations from Sobel's formula or the case-based bootstrap. Second, specifying a system of separate equations can lead to conflicting estimates of mediation effects when the system is not internally consistent, as discussed in the Empiric versus implied total effect estimates section. The single-model framework avoids this problem, which is not uncommon in complex models. A third benefit of the single-model framework is faster computation in high-dimensional settings. With the onslaught of big data and data science applications, this is a nontrivial advance. Fourth, the single-model approach imparts substantial convenience in applying modeling tools such as imputation of missing data, penalization, and cross-validation. These increasingly common techniques are cumbersome to apply to a system of equations, and it is often unclear how they can be applied at all. We illustrate the new single-model approach and compare it to existing methods in a series of detailed, reproducible examples in the Examples using data from social psychology research section. The R code is in the Supplement and the data we used are freely available online.

## Background and notation

### What is a mediator?

Several types of variables may be present when analyzing the relationship between an exposure and some outcome of interest. A confounder is a variable related to two factors of interest that falsely obscures or accentuates the relationship between them (Meinert, 1986). We want to adjust for appropriate confounders to obtain an unbiased estimate of the relationship between the exposure and the outcome. By contrast, a moderator (also known as "effect-modifier" in the epidemiologic literature) is a variable that affects the direction or strength of the relationship between the exposure and outcome (Baron & Kenny, 1986). In regression analysis, we typically account for moderators by including interaction terms in the model.

Lastly, a mediator represents “the generative mechanism through which the focal independent variable is able to influence the dependent variable” (Baron & Kenny, 1986) or “a variable that occurs in a causal pathway from an independent variable to a dependent variable” (Last, 1988). We will return to the importance of distinguishing between these types of variables when we discuss the assumptions of mediation models. Next, we introduce the simple mediation model.

### The simple mediation model

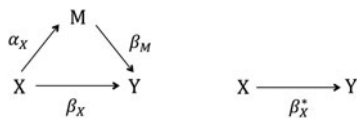
By partitioning the *total effect* of an exposure into its *direct* and *indirect* components, mediation analysis seeks to understand how much of an exposure’s effect on an outcome is transmitted through intermediate pathways. The total effect of the exposure variable  $X$  on the outcome  $Y$  represents how much a change in  $X$  results in a change  $Y$ , irrespective of the mechanisms by which the change occurs; the part of the total effect that is not transmitted through intervening variables is called the direct effect (Alwin & Hauser, 1975); the indirect effect is the part of a variable’s total effect that is transmitted to the outcome via mediating variable(s)  $M$ . Suppose we have one exposure  $X$ , one continuous mediator  $M$ , and one continuous outcome  $Y$ . The Baron-Kenny *simple mediation model* is illustrated in Figure 1 and represented by the following three regression equations.

$$E[Y|X, M] = \beta_0 + \beta_X X + \beta_M M \quad (1)$$

$$E[M|X] = \alpha_0 + \alpha_X X \quad (2)$$

$$E[Y|X] = \beta_0^* + \beta_X^* X \quad (3)$$

This model assumes linear relationships, no interaction among the variables, and normally distributed errors. Although the original Baron-Kenny model did not include covariates, it is important to adjust for confounders of the type listed in Figure 2 so that mediation effects are identifiable (simply add the set of relevant confounders to each model).

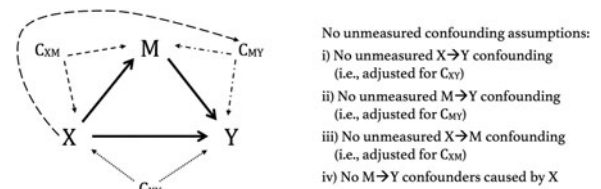


**Figure 1.** Simple mediation model for exposure  $X$ , continuous mediator  $M$ , and continuous outcome  $Y$ . The coefficients  $\alpha_X$ ,  $\beta_X$ ,  $\beta_M$ , and  $\beta_X^*$  are estimated from the system of three regression equations.

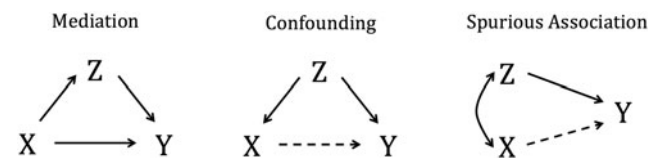
### Assumptions for causal inference

Critical assumptions concerning the relationships in a proposed mediation model rely on theory and empirical support. To infer causality from a mediation analysis, one must assume the confounders enumerated in Figure 2 have been accounted for, T. J. VanderWeele (2015). In addition, mediation analysis assumes the temporal order of the variables was correctly specified (Judd & Kenny, 1981; Stone & Sobel, 1990). A mediator must truly be a dependent variable relative to the exposure and an independent variable relative to the outcome (it should be noted that the timing of a variable’s *measurement* may be different than the timing of the variable itself).

Mediation analysis also relies on correctly specified causal directions (McDonald, 1997). Consider the diagrams in Figure 3 where  $X$  is the exposure,  $Y$  is the outcome, and  $Z$  is a third variable influencing the effect of  $X$  on  $Y$ . The dashed lines represent a spurious  $X \rightarrow Y$  effect. The first panel displays  $Z$  acting as a mediator; the second panel shows  $Z$  influencing both  $X$  and  $Y$ , leading to a spurious  $X \rightarrow Y$  effect; the



**Figure 2.** Assumptions required for estimating causal mediation effects from a model with exposure  $X$ , mediator  $M$ , and outcome  $Y$ .  $C_{XY}$  represents confounding variables of the  $X \rightarrow Y$  relationship,  $C_{MY}$  represents confounders of the  $M \rightarrow Y$  relationship, and  $C_{XM}$  represents confounders of the  $X \rightarrow M$  relationship. To estimate causal mediation effects, the researcher assumes that he has controlled for  $C_{XY}$ ,  $C_{MY}$ ,  $C_{XM}$  and that there are no  $M \rightarrow Y$  confounders caused by  $X$ . Identifying controlled direct effects requires that assumptions i) and ii) be met. Identifying natural direct and indirect effects requires that all four assumptions be met. Figure adapted from T. J. VanderWeele (2015) book *Explanation in Causal Inference: Methods for Mediation and Interaction*.



**Figure 3.** Statistically indistinguishable three-variable systems: The arrows represent the causal direction of the effects between variables and dashed lines represent a spurious effect between  $X$  and  $Y$ . The left panel represents the simple mediation model where  $Z$  mediates the effect of  $X$  on  $Y$ . The middle panel shows  $Z$  confounding the relationship between  $X$  and  $Y$ . The right panel shows  $X$  and  $Z$  as two covariates having a reciprocal relationship.

third panel shows a reciprocal relationship between  $X$  and  $Z$ , with  $Z$  affecting  $Y$  without being causally intermediate. Although these models are conceptually distinct, they are mathematically equivalent and cannot be empirically distinguished from one another with cross-sectional data (Cole & Maxwell, 2003).

### Intersection of existing frameworks

Many articles on mediation analysis tout the benefits of one framework over another for making causal inferences. Social and behavioral scientists tend to adopt the SEM language, while statisticians more often use the PO approach (and may even be unfamiliar with the basic tenets of SEM). Although one framework may lend itself to a specific research question, the SEM and PO frameworks are “logically equivalent,” a result proven formally by Galles and Pearl (1998).

The essential difference between the SEM and PO frameworks is that the former encodes causal knowledge in the form of functional relationships among ordinary variables, observable as well as latent, while the latter encodes such knowledge in the form of statistical relationships among hypothetical (or counterfactual) variables, whose value is determined only after a treatment is enacted... A systematic analysis of the syntax and semantics of the two notational systems reveals that they are logically equivalent; a theorem in one is a theorem in the other, and an assumption in one has a parallel interpretation in the other (Bollen & Pearl, 2013).

Providing definitions of causal mediation effects and their regression-based estimands, Table 1 is a crosswalk between the SEM and PO nomenclature. A brief introduction to the terminology and notation of SEM is provided in Appendix A.1.4.

### Beyond the simple mediation model

This section provides a brief introduction to the recently proposed single-model regression framework by Saunders and Blume (2017). We then show how we have extended this framework to accommodate non-nested mediation systems. We present a novel method for visualizing mediation effects that leverages the single-model approach. A joint measure of mediation is also proposed.

#### The recently proposed single-model regression framework

Recall that the simple mediation model (1–3) assumes  $X$  is linearly related to  $Y$ . A more general formulation allows the effect of the exposure to

be nonlinear and includes additional covariates  $C$ :  $E[Y|X, M, C] = \beta_0 + \beta_X h(X) + \beta_M M + \beta_C C$ , where  $h(X)$  is a flexible function of  $X$  (e.g.,  $\log X$ ). Consider  $p$  exposures  $X$ ,  $j$  mediators  $M$ , and  $l$  covariates  $C$  such that the full model and reduced model are

$$E[Y|X, M, C] = \beta_0 + \beta_X h(X) + \beta_M M + \beta_C C \quad (4)$$

$$E[Y|X, C] = \beta_0^* + \beta_X^* h(X) + \beta_C^* C \quad (5)$$

The vector  $h(X)$ , such as  $[X, X^2]$  or  $[X, XC]$ , captures the nonlinear trends in  $X$ . In a recent article, we named the difference in exposure pathway coefficients  $\Delta = \beta_X^* - \beta_X$  the essential mediation components (EMCs) of  $X$  (Saunders & Blume, 2017). We derived analytical estimates of the EMCs and their model-based variance from the fit of a single regression model. Because the fit of only one model is required, inference for causal mediation effects (which are functions of the EMCs) follows naturally.

Our method uses the “full” outcome model (4) and the model for the total effect of the exposure (5). The general idea of our approach is to use the sweep operator on the full model to obtain coefficients from any nested reduced model, without having to actually fit said reduced model (Goodnight, 1979). This allows us to obtain the EMCs from the fit of the full model alone. Saunders and Blume (2017) discuss the advantages that result from having to fit only one model (e.g., simplified application of regression tools and reduced computation time).

### The essential mediation components

A general formula for estimating the EMCs from the fit of the full regression model (4) is

$$\begin{aligned} \hat{\Delta} = \hat{\beta}_X^* - \hat{\beta}_X &\equiv -\hat{V}_{XM} \hat{V}_M^{-1} \hat{\beta}_M \\ &= -\widehat{\text{cov}}(\hat{\beta}_X, \hat{\beta}_M) \widehat{\text{var}}(\hat{\beta}_M)^{-1} \hat{\beta}_M \end{aligned} \quad (6)$$

where  $\hat{\beta}_X$  and  $\hat{\beta}_M$  are the vectors of estimated exposure and mediator coefficients from the full model,  $\hat{V}_{XM}$  is the covariance between  $\hat{\beta}_X$  and  $\hat{\beta}_M$ , and  $\hat{V}_M^{-1}$  is the inverse variance of  $\hat{\beta}_M$ . For the simple mediation model,  $-\widehat{\text{cov}}(\hat{\beta}_X, \hat{\beta}_M) \widehat{\text{var}}(\hat{\beta}_M)^{-1} \hat{\beta}_M$  equals the Baron-Kenny product of coefficients  $\alpha_X \beta_M$ , where  $-\widehat{\text{cov}}(\hat{\beta}_X, \hat{\beta}_M) \widehat{\text{var}}(\hat{\beta}_M)^{-1}$  equals  $\alpha_X$ .

The distinction between changes in the EMCs and causal mediation estimands (e.g., PE, NIE) is critical. For the simple mediation model (1–3), the EMC  $\hat{\Delta} = \hat{\beta}_X^* - \hat{\beta}_X$  is mathematically equivalent to the indirect effect for a unit change in the exposure; more generally, causal mediation estimands are functions of the EMCs. The PE is the difference between the total effect and the controlled direct effect (CDE). To



**Table 2.** Results from the simple mediation model example in Example: the simple mediation model section.

	Simple mediation model		Simple mediation model adjusting for confounders	
	Portion eliminated	SE	Portion eliminated	SE
Proposed approach <sup>a</sup>	−0.0103	0.0016	−0.0107	0.0016
Simulation-based <sup>b</sup>	−0.0103	0.0089	−0.0107	0.0094
Difference of coefficients	−0.0103	–	−0.0107	–
Product of coefficients	−0.0103	–	−0.0107	–
Residual bootstrap	–	0.0017	–	0.0016
Case-based bootstrap	–	0.0088	–	0.0093
Sobel	–	0.0087	–	0.0093

<sup>a</sup>Saunders and Blume (2017).<sup>b</sup>Imai et al. (2010).

In this example, the portion eliminated equals the natural indirect effect. The proposed approach by Saunders and Blume (2017) provides an estimate of the portion eliminated and its model-based variance from the fit of one model. The simulation-based approach by Imai et al. (2010) simulates estimates of the portion eliminated and its variance. Point estimates of the portion eliminated can be obtained using the difference of coefficients and product of coefficients approaches. Standard error estimates can be obtained using the residual bootstrap, the case-based bootstrap, and Sobel's approximation.

estimate the PE comparing some referent level of the exposure  $x_o$  to  $x$ , use

$$PE(x, x_o) = [\mathbf{h}(x) - \mathbf{h}(x_o)] \Delta \quad (7)$$

When the exposure or mediator effects are non-scalar, formulas (6) and (7) allow for estimation of multidimensional mediation effects from the fit of a single fitted regression model (4), rather than fitting separate models and combining effect estimates. A list of commonly encountered mediation models for which the controlled and natural direct effects are equivalent (and as a result the PE and the NIE are the same) can be found in the Supplemental Table 2 of Saunders and Blume (2017). For these models, since the PE and the NIE are equal to each other, the NIE can also be estimated as a function of the EMCs:  $NIE(x, x_o) = [\mathbf{h}(x) - \mathbf{h}(x_o)] \Delta$ .

### The model-based variance

Saunders and Blume (2017) provide an analytical solution for the variance of the EMCs, making the variance of mediation effects readily accessible. The expression for the fully conditional variance of the EMCs is given by  $\text{Var}(\hat{\Delta} | \mathbf{X}, \mathbf{M}) = \mathbf{V}_{XM} \mathbf{V}_M^{-1} \mathbf{V}_{MX}$ . The model-based variance of the PE is simply  $[\mathbf{h}(x) - \mathbf{h}(x_o)] \text{Var}(\hat{\Delta} | \mathbf{X}, \mathbf{M}) [\mathbf{h}(x) - \mathbf{h}(x_o)]^T$ , which requires fitting only model (4). The standard conditional variance is used for inference in classical regression settings and one could argue in favor of treating both the exposure and the mediator as fixed since the mediator is a theoretic consequent of the exposure (Pearl, 2012a). Alternately, one can marginalize over  $M$  as discussed in Saunders and Blume (2017) Section 3.3. A parametric (residual-based) bootstrap approximates the conditional model-based variance, while the nonparametric (case-based) bootstrap approximates

the fully unconditional variance (marginalized over both exposure and mediator). Information on commonly used approximations to the variance is provided in Appendix A.2. From properties of a regression model, for a scalar PE and a unit change in the exposure we have  $\frac{\hat{PE} - PE}{\sqrt{\text{Var}(\hat{PE})}} \sim t(df = n - k, \text{scale} = -1)$  and the 95% confidence interval is  $\hat{PE} \pm t_{.975, n-k} \times -\hat{V}_{XM} \hat{V}_M^{-1} \sqrt{\hat{\text{Var}}(\hat{\beta}_M)}$ .

### Extensions to non-nested mediation systems

A direct application of the EMC formula is not applicable to mediation models in which the marginal model is not nested within the full model. Here we show how a recursive sweep algorithm solves this problem. Simply specify a global model under which the full model (4) and the reduced model (5) are nested. Use formula (6) to “sweep” from the global model  $G$  to the reduced model  $R$  (to obtain  $\Delta_{RG}$ ) and also from the global model  $G$  to the full model  $F$  (to obtain  $\Delta_{FG}$ ). After this “double sweep,” subtract the corresponding  $\Delta$ s to get the proper EMCs:  $\text{EMCs} = \Delta_{RG} - \Delta_{FG}$ . This expands the applicability of the classical regression framework to non-nested mediation systems.

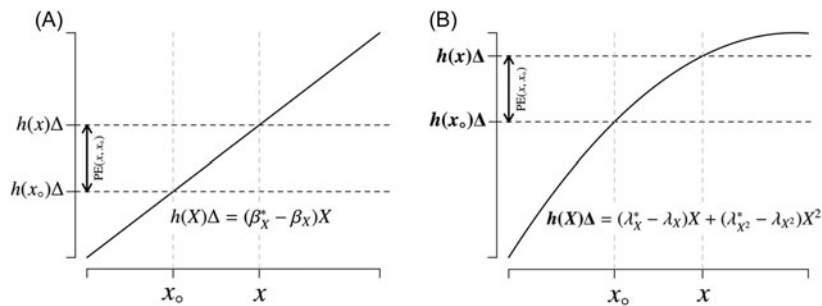
Suppose the outcome model, mediator model, and the implied marginal model are  $E[Y | X, M] = \delta_0 + \delta_X X + \delta_M M$ ,  $E[M | X, C] = \alpha_0 + \alpha_X X + \alpha_C C$ , and  $E[Y | X, C] = \kappa_0 + \kappa_X X + \kappa_C C$ , respectively. Estimating the EMCs  $\kappa_X - \delta_X$  using formula (6) requires the reduced model to be a defacto submodel of the full model. However, by fitting a global model  $E[Y | X, M, C] = \lambda_0 + \lambda_X X + \lambda_M M + \lambda_C C$  and using (6) to estimate  $\Delta_{RG} = \kappa_X - \lambda_X$  and  $\Delta_{FG} = \delta_X - \lambda_X$ , we can estimate the EMCs using functionals of only the global model:  $\Delta_{RG} - \Delta_{FG} = (\kappa_X - \lambda_X) - (\delta_X - \lambda_X) = \kappa_X - \delta_X$ .

Now consider the model  $E[Y|X, M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$ . If  $E[M|X] = \alpha_0 + \alpha_X X$ , the marginal model for the total effect is  $E[Y|X] = \beta_0^* + \beta_X^* X + \beta_{X^2} X^2$ . To estimate the EMCs  $\Delta^T = [\beta_X^* - \beta_X, \beta_{X^2}^* - 0]$ , fit a global full model  $E[Y|X, M] = \gamma_0 + \gamma_X X + \gamma_M M + \gamma_{XM} XM + \gamma_{X^2} X^2$  and use (6) to estimate  $[\beta_X^* - \gamma_X, \beta_{X^2}^* - \gamma_{X^2}] - [\beta_X - \gamma_X, 0 - \gamma_{X^2}] = [\beta_X^* - \beta_X, \beta_{X^2}^* - 0]$ . Thus, one can use the difference of coefficients approach and fit only one model to estimate mediation effects from systems where the reduced model is not a submodel of the full model.

### Visualizing mediation effects

Mediation effects can be visualized as functions of the exposure and mediator values. Recall that for the simple mediation model specified in (1–3), the EMC  $\Delta = \beta_X^* - \beta_X$  and the PE  $= [h(x) - h(x_0)]\Delta = (\beta_X^* - \beta_X)(x - x_0)$ . Figure 4 panel A shows the PE (which equals the indirect effect in this simple case) as the vertical distance between the line  $h(X)\Delta$  evaluated at  $x$  and  $x_0$ .

Now consider the full model with a quadratic exposure effect so that  $\mathbf{h}(\mathbf{X}) = [X, X^2]$ ,  $E[Y|X, M] = \lambda_0 + \lambda_X X + \lambda_{X^2} X^2 + \lambda_M M$ , and the reduced model  $E[Y|X] = \lambda_0^* + \lambda_X^* X + \lambda_{X^2}^* X^2$ . The EMCs  $[\Delta_1, \Delta_2]^T = [\lambda_X^* - \lambda_X, \lambda_{X^2}^* - \lambda_{X^2}]$  and the PE is  $[\mathbf{h}(x) - \mathbf{h}(x_0)]\Delta = [x - x_0, x^2 - x_0^2]\Delta = (\lambda_X^* - \lambda_X)(x - x_0) + (\lambda_{X^2}^* - \lambda_{X^2})(x^2 - x_0^2)$ . The PE is simply the distance between the parabola  $\mathbf{h}(\mathbf{X})\Delta = \Delta_1 X + \Delta_2 X^2$  evaluated at  $x$  and  $x_0$ , as shown in Figure 4 panel B. This is a helpful way to illustrate complex mediation behavior and accommodates multidimensional exposures. Alternatively, depictions of the relationships among the exposure, mediator, and outcome variables can be used to illustrate the indirect effect and its components (MacKinnon, 2008); however, this approach can be cumbersome when the number of exposures or mediators is large.



**Figure 4.** Visualizing the portion eliminated on the vertical axis as a function of the exposure values  $(x, x_0)$  on the horizontal axis. Panel A shows the portion eliminated as the distance (arrow  $\leftrightarrow$ ) between the line  $h(X)\Delta$  evaluated at  $x$  and  $x_0$ . Panel B shows the portion eliminated as the distance between the parabola  $\mathbf{h}(\mathbf{X})\Delta = \Delta_1 X + \Delta_2 X^2$  evaluated at  $x$  and  $x_0$ .

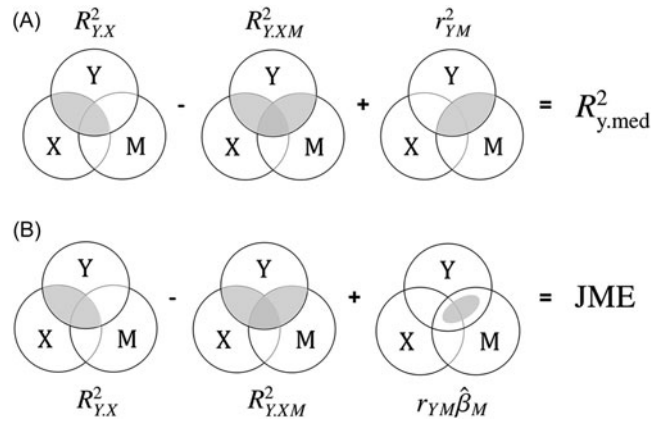
### Proposed measure of joint mediation

Until now, we have considered mediation of a single exposure through one mediator. Suppose the mediation model is more complex, such that we are interested in the mediation of a set of  $p$  exposures  $\mathbf{X} = (X_1, \dots, X_p)$  by a set of  $j$  mediators  $\mathbf{M} = (M_1, \dots, M_j)$ . One can fit the full model and obtain separate estimates of each exposure's indirect effect using formula (6). However, summing separate indirect effect estimates to measure total mediation may fail to account for overlapping mediation effects when exposures are mediated jointly.

To address this problem, we propose a general measure of the *joint mediation effect* (JME). This measure aims to capture the amount of mediation that a group of exposure variables is responsible for as a whole, which is not necessarily the sum of the indirect effects. Let  $R_{Y.XMC}^2$  and  $R_{Y.XC}^2$  be the coefficients of determination from the full model  $E[Y|\mathbf{X}, \mathbf{M}, \mathbf{C}] = \beta_0 + \beta_X \mathbf{X} + \beta_M \mathbf{M} + \beta_C \mathbf{C}$  and the reduced model  $E[Y|\mathbf{X}, \mathbf{C}] = \beta_0^* + \beta_X^* \mathbf{X} + \beta_C^* \mathbf{C}$ , respectively. To measure the JME of multiple exposures through multiple mediators, it is helpful to scale the variables to have unit variance. We define the JME as a linear combination of  $p$  individual EMCs (which, when standardized, are on the same scale), where the EMC  $\hat{\Delta}_i$  of exposure  $X_i$  is multiplied by  $X_i$ 's correlation with the outcome  $Y$ :

$$r_{YX}^T \hat{\Delta} = \sum_{i=1}^p r_{YX_i} (\hat{\beta}_{X_i}^* - \hat{\beta}_{X_i}) = \sum_{k=1}^j r_{YM_k} \hat{\beta}_{M_k} - (R_{Y.XMC}^2 - R_{Y.XC}^2) - \sum_{h=1}^l r_{YC_h} (\hat{\beta}_{C_h}^* - \hat{\beta}_{C_h}). \quad (8)$$

The JME is unitless and will give the same numerical value whether the data are standardized or unstandardized. Note that when the data are unstandardized, the JME is  $\sum_{i=1}^p \frac{\text{cov}(Y, X_i)}{\text{var}(Y)} \hat{\Delta}_i = \sum_{k=1}^j \frac{\text{cov}(Y, M_k)}{\text{var}(Y)} \hat{\beta}_{M_k} - (R_{Y.XMC}^2 - R_{Y.XC}^2) - \sum_{h=1}^l \frac{\text{cov}(Y, C_h)}{\text{var}(Y)} (\hat{\beta}_{C_h}^* - \hat{\beta}_{C_h})$ , which



**Figure 5.** A comparison of MacKinnon's R-squared measure of mediation,  $R^2_{y.med}$  (MacKinnon, 2008) in panel A to the proposed joint mediation effect (JME) in panel B for the simple mediation model with exposure  $X$ , mediator  $M$ , and outcome  $Y$  (see Proposed measure of joint mediation section for discussion).  $R^2_{Y,XM}$  and  $R^2_{Y,X}$  are the coefficients of determination from the full and marginal outcome models, respectively;  $r^2_{YM}$  is the squared correlation between  $Y$  and  $M$ , and  $\hat{\beta}_M$  is the regression coefficient for the mediator from the full outcome model. The gray shaded areas of the Venn diagrams represent the amount of variation in the outcome variable  $Y$  accounted for by  $R^2_{y.med}$  and the JME.

may appear incongruous with the unscaled  $\hat{\Delta}_i$ s because of the difference in units among exposures. We can show this measure is bounded between  $(-2,2)$ , although there may be tighter achievable bounds.

MacKinnon provides an  $R^2$  measure “designed to localize the amount of variance in  $Y$  that is explained by  $M$  specific to the mediated effect... by identifying the variance in  $Y$  explained by both  $M$  and  $X$  but not by  $X$  alone or  $M$  alone:”  $R^2_{y.med} = r^2_{YM} - (R^2_{Y,XM} - R^2_{Y,X})$  (MacKinnon, 2008). de Heus (2012) argues that  $R^2_{y.med}$  assigns all overlap between the direct and indirect effects to the indirect effect, which is problematic because they are “heavily interdependent.” For the simple mediation model, the JME replaces the correlation  $r_{YM}$  with the semipartial correlation  $\hat{\beta}_M = r_{YM.X}$ , giving  $r_{YX}\hat{\Delta} = r_{YM}\hat{\beta}_M - (R^2_{Y,XM} - R^2_{Y,X})$ . The first term  $r_{YM}\hat{\beta}_M$  will be less than  $r^2_{YM}$  if  $r_{YM.X} < r_{YM}$ . Our  $R^2$  measure matches MacKinnon's in two rather extreme cases:  $\hat{\beta}_M = r_{YM}$  if either  $X$  and  $M$  are uncorrelated or the effect of  $X$  on  $Y$  adjusted for  $M$  in the full model is zero (in the Baron-Kenny framework, this is the definition of “complete” mediation). The Venn diagrams in Figure 5 help intuit the similarity between our measure of the JME and MacKinnon's measure for the simple mediation model.

### Empiric versus implied total effect estimates

The equations in a mediation model form an interdependent system and need to be specified so that they remain congruous. For example, suppose one believes  $M$  depends on  $X$  quadratically and writes the following model:

$$\begin{aligned} E[Y|X, M] &= \beta_0 + \beta_X X + \beta_M M \\ E[M|X] &= \alpha_0 + \alpha_X X + \alpha_{X^2} X^2 \\ E[Y|X] &= \beta_0^* + \beta_X^* X \end{aligned}$$

This system is not congruous because marginalizing over the full outcome model  $E_{M|X}[Y|X, M]$  does not yield the specified marginal model  $E[Y|X]$ . A properly specified system of equations would include an  $X^2$  term in the third equation. System congruency is important because the difference of coefficients approach relies on the full model and the marginal model, while the product of coefficients approach relies on the full model and the mediator model. When the specified system is not congruent, the difference and product approaches are no longer comparable because they are actually using *different systems* to estimate mediation effects. Thus, discrepancies between the two approaches may be due to differences in their underlying mediation systems. Importantly, our proposed approach avoids this issue because it fits only the full model.

At its core, mediation analysis aims to decompose the total effect of an exposure into the direct and mediated components. Our approach, a generalization of the difference of coefficients approach, uses the empiric total effect. This “empiric” effect is the maximum likelihood estimate of the exposure's total effect on the outcome from the marginal model. In contrast, the product of coefficients approach estimates the direct and indirect effects from fitting the full outcome model and the mediator model, and then sum these to obtain the implied total effect. This estimate of the total effect is “implied” because the marginal model



$E[Y|X]$  is not actually fit to the data. The sum of the estimated natural direct and indirect effects is routinely used as the total effect estimate in the PO approach (for instance, this is how the R mediation package by Imai et al. (2010) estimates the total effect).

Identifying the PE as the difference between the total and CDEs relies on only the first two assumptions in Figure 2, whereas estimating the natural direct and indirect effects (and subsequently summing them to estimate the total effect) requires all four assumptions be met (Robins & Greenland, 1992). As a result, assumptions about potential mediation pathways play an outsized role in determining the implied total effect.

The difference and product approaches often yield the same conclusion. For the simple mediation model (1–3), it is well known that  $E[Y|X] = E_{M|X}[E[Y|X, M]] = \beta_0 + \beta_X X + \beta_M(\alpha_0 + \alpha_X X) = \beta_0^* + \beta_X^* X$ , which proves  $\beta_X^* = \beta_X + \alpha_X \beta_M$  (an alternative proof is given by MacKinnon, Warsi, and Dwyer (1995)). That is, the estimated total effect  $\hat{\beta}_X^*$  obtained from fitting the marginal model  $E[Y|X]$  equals the sum of the estimated natural direct and indirect effects. However, this is not always the case. In our experience, the discrepancies present themselves in settings with exposure–mediator interactions (exposure–covariate interactions do not present this problem). When fitting the full model  $E[Y|X, M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$ , the mediator model  $E[M|X] = \alpha_0 + \alpha_X X$ , and the implied marginal model  $E[Y|X] = \gamma_0 + \gamma_X X + \gamma_X^2 X^2$ , the empiric estimate of the total effect  $\hat{\gamma}_X(x - x_0) + \hat{\gamma}_X^2(x^2 - x_0^2)$  does not equal the sum of the estimated natural direct and indirect effects. Thus, even when the marginal models are theoretically equivalent, the estimates of the total effect can differ. This can lead to conflicting results since the total effect is used to gauge the overall decomposition of the exposure effect.

Why does this happen? Once the full outcome model and the mediator model are specified, there is one theoretical marginal model. However, there can be several mediator models that imply the same form for the marginal model. For example, consider the full model  $E[Y|X, M, W] = \beta_0 + \beta_X X + \beta_M M + \beta_W W + \beta_{XW} XW$  and two mediator models: (a)  $E[M|X, W] = \alpha_0 + \alpha_X X + \alpha_W W$  and (b)  $E[M|X, W] = \delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW$ . Notice that model (b) includes an  $XW$  interaction and model (a) does not. Marginalizing the full model over  $M$  using (a) gives  $E[Y|X, W] = E_{M|X, W}[Y|X, M, W] = (\beta_0 + \alpha_0 \beta_M) + (\beta_X + \alpha_X \beta_M)X + (\beta_W + \alpha_W \beta_M)W + \beta_{XW} XW$ . Marginalizing over  $M$

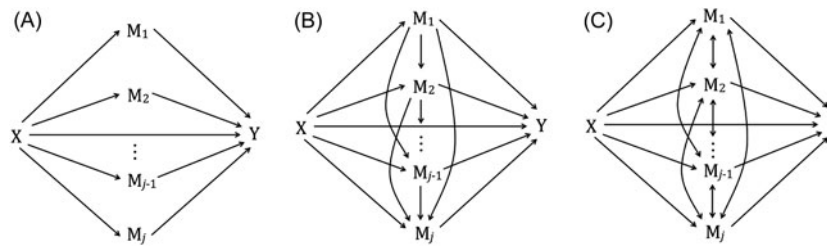
using (b) gives  $E[Y|X, W] = (\beta_0 + \delta_0 \beta_M) + (\beta_X + \delta_X \beta_M)X + (\beta_W + \delta_W \beta_M)W + (\beta_{XW} + \delta_{XW} \beta_M)XW$ . Both marginal models have the form  $E[Y|X, W] = \gamma_0 + \gamma_X X + \gamma_W W + \gamma_{XW} XW$ . As a result, and importantly, the fitted estimates of the marginal model effects  $\hat{\beta}$  will be the same (even though the implied coefficients from the two marginal models are different).

In a conditional process model meant to represent “moderation of only the direct effect,” (Hayes, 2013, p. 335) specifies the same full model as above and a mediator model that omits  $W$  and  $XW$ :  $E[M|X] = \alpha_0 + \alpha_X X$ . If  $E[M|X, W] = \delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW$ , then  $E[M|X] = E_{W|X}[E[M|X, W]] = E_{W|X}[\delta_0 + \delta_X X + \delta_W W + \delta_{XW} XW] = \delta_0 + \delta_X X + (\delta_W + \delta_{XW} X)E[W|X]$ . So, writing  $E[M|X] = \alpha_0 + \alpha_X X$  assumes  $\delta_W = \delta_{XW} = 0$ . From this example, we see that the “implied” total effect (obtained from summing the estimated direct and indirect effects) may rely on hidden assumptions about the mediation mechanism. Estimating the marginal model directly can be used to assess the degree to which these assumptions are supported by the data.

## Multiple mediators

Multiple mediator models are useful when researchers hypothesize that the exposure affects the outcome through several intermediate pathways. Consider the full model that contains  $j$  mediators  $E[Y|X, M, C] = \beta_0 + \beta_X X + \sum_{i=1}^j \beta_i M_i + \beta_C C$  and the corresponding reduced model  $E[Y|X, C] = \beta_0^* + \beta_X^* X + \beta_C^* C$ . The total and direct effects of  $X$  on  $Y$  are given by  $\beta_X^*(x - x_0)$  and  $\beta_X(x - x_0)$ , respectively. To identify mediation effects from multiple mediator models, all four no unmeasured confounding assumptions outlined in Figure 2 must hold with respect to the set of mediators  $M$ .

Estimating the *total indirect effect* aims to determine if the set of  $j$  mediators transmits the effect of  $X$  to  $Y$ . This is analogous to conducting a regression analysis with several exposures, with the aim of determining if an overall effect exists. The *mediator-specific indirect effect* represents the amount of the exposure’s effect on the outcome that is mediated by  $M_i$  above and beyond the other  $j - 1$  mediators and adjusted for the covariates in the model. The mediator-specific effects are often attenuated due to collinearity among the mediators (Preacher & Hayes, 2008a); that is, if two or more mediators share a role in transmitting the effect of  $X$  to  $Y$ , then the effect attributed uniquely to mediator  $M_i$  may exclude this overlapping effect. Additionally, specific indirect effects might have



**Figure 6.** Comparison of multiple mediator models. Panel A depicts the single-step or parallel multiple mediator model. Panel B depicts the serial multiple mediator model. Panel C depicts our proposed single-model framework for assessing mediation with multiple mediators. The directions of arrows indicate the assumed causal pathways. The parallel model assumes the mediators do not causally affect each other; the serial model assumes some temporal ordering of the mediators (represented by the unidirectional arrows between mediators); the single-model framework allows the mediators to covary and does not assume the temporal ordering is known (represented by the bidirectional arrows between mediators).

different signs, leading to inconsistent mediation. As a result, mediator-specific indirect effects do not necessarily sum to the total indirect effect.

A clear advantage of our approach in the presence of multiple mediators is that it requires fitting only one model to obtain an estimate of the total indirect effect through  $M$  and estimates of mediator-specific indirect effects. It also yields model-based variance estimates that do not require the computation time of resampling methods. Using formula (7), the total indirect effect through  $M$  and the mediator-specific indirect effect of  $X$  through  $M_i$  are easily estimated using  $-\hat{V}_{XM}\hat{V}_M^{-1}\hat{\beta}_M(x-x_o)$  and  $-\hat{V}_{XM_i}\hat{V}_{M_iM_i}^{-1}\hat{\beta}_{M_i}(x-x_o)$ , respectively. The corresponding variances are estimated by  $\text{Var}([h(x)-h(x_o)]\hat{\Delta}) = (x-x_o)^2\hat{V}_{XM}\hat{V}_M^{-1}\hat{V}_{MX}$  and  $(x-x_o)^2\hat{V}_{XM_i}\hat{V}_{M_iM_i}^{-1}\hat{V}_{M_iX}$ . This approach does not assume the mediators act independently nor does it assume a particular order of effects (see panel C of Figure 6).

### Comparison of multiple mediator models

In the context of multiple mediators, existing approaches include the *single-step multiple mediator model* (MacKinnon, 2008), also termed the *parallel multiple mediator model* (Hayes, 2013), and the *serial multiple mediator model* (Hayes, 2013). The parallel approach specifies a separate model for each mediator and assumes that each mediator independently affects the outcome. In other words, the parallel model assumes the mediators do not covary (see panel A of Figure 6). The serial model assumes the temporality of the mediators is known and allows the mediators to covary according to the specified order in which mediators affect each other (see panel B of Figure 6). Our single-model approach allows each mediator to depend on the others but does not assume an ordering of the mediators (see panel C of Figure 6).

Our approach gives the same estimated total effect, direct effect, and PE (which equals the total indirect effect if there is no exposure-mediator interaction) as the parallel and serial models. However, our approach does not give the same *mediator-specific* indirect effect estimates as these models. The difference between the mediator-specific estimates is due to the way each approach specifies the mediator models (which reflect the assumptions of each approach). The parallel model assumes the mediators independently affect the outcome, so each mediator is modeled separately and not conditional on the others. The serial approach assumes an ordering of the mediators, so each mediator is modeled separately and is conditional on those that precede it in the causal chain. Our approach implicitly incorporates the covariance between all of the mediators. As a result, our approach does not give the result of the parallel approach (unless the mediators are conditionally independent) or the serial approach (since we don't specify the order of the mediators). Appendix A.3 provides a detailed description of these and other approaches to multiple mediator models. Table 3 shows the mediator equations that coincide with the parallel model, serial model, and our approach for the two-mediator setting. While our approach does not actually fit separate mediator models, we present the underlying models to show how each mediator-specific effect is conditional on all the other mediators.

### Advantages of estimating the total indirect effect through $M$

We recommend the researcher's primary interest lie in the total indirect effect rather than the amount mediated by a specific mediator. Importantly, both our framework and existing approaches to mediation with multiple mediators specify the *same full model* for the outcome  $Y$ , and as a result yield the same

**Table 3.** Regression equations from the parallel multiple mediator model, the serial multiple mediator model, and the proposed single-model approach for the two-mediator setting.

Parallel	Serial	Single-model
$E[Y X, M_1, M_2] = \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2$	$E[Y X, M_1, M_2] = \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2$	$E[Y X, M_1, M_2] = \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2$
$E[Y X] = \beta_0^* + \beta_X^* X$	$E[Y X] = \beta_0^* + \beta_X^* X$	$E[Y X] = \beta_0^* + \beta_X^* X$
$E[M_1 X] = \alpha_{01} + \alpha_1 X$	$E[M_1 X] = \alpha_{01} + \alpha_1 X$	$E[M_1 X, M_2] = \kappa_{01} + \kappa_1 X + \delta_{12} M_2$
$E[M_2 X] = \alpha_{02} + \alpha_2 X$	$E[M_2 X, M_1] = \kappa_{02} + \kappa_2 X + \delta_{21} M_1$	$E[M_2 X, M_1] = \kappa_{02} + \kappa_2 X + \delta_{21} M_1$

For space considerations, we omit confounders  $C$  from the equations. Notice that all three approaches specify the same full outcome model  $E[Y|X, M_1, M_2]$  and reduced outcome model  $E[Y|X]$ , but the models for the mediators  $M_1$  and  $M_2$  differ. The parallel model assumes  $M_1$  and  $M_2$  independently affect the outcome. The serial model assumes the ordering of the mediators is known (in this case,  $M_2$  depends on  $M_1$  but  $M_1$  does not depend on  $M_2$ ). The proposed approach allows all mediators to be interdependent, and we present the single-model framework's implied mediator models in gray to help intuit this ( $M_1$  and  $M_2$  can both depend on each other). Note that the models in gray do not need to be fit; the single-model framework requires fitting only the full outcome model  $E[Y|X, M_1, M_2]$ .

estimate of the direct effect of  $X$  and of the total indirect effect of  $X$  through  $M$ . The estimated total indirect effect through  $M$  comes from the full model containing all of the mediators and does not suffer bias from the misspecification of inter-mediator relationships, whereas estimates of mediator-specific indirect effects rely heavily on assumed inter-mediator relationships (such as the order of causal effects in serial multiple mediator models, which can be unverifiable with cross-sectional data). We present relevant examples of this in Example: multiple mediators section.

### Interactions and moderated mediation

To use the proposed framework with exposure–exposure, mediator–mediator interactions, and exposure–confounder interactions, simply include the interaction terms in the full model and use formulas (6) and (7) to estimate the EMCs and causal mediation effects. We provide examples in Example: interactions section.

Exposure–mediator interactions, on the other hand, lead to mediation effects that are less clearly defined. Judd and Kenny (1981) use the term *moderated mediation* to describe when  $X$  moderates its own indirect effect on  $Y$  by moderating the effect of  $M$  on  $Y$ . They discuss the importance of including exposure–mediator interactions in the model, and they recommend doing significance testing on the interaction terms. If the interaction effect is statistically significant, one can conclude the treatment modifies the mediation process; however, they do not explain how to subsequently estimate the direct and indirect effects in the presence of an exposure–mediator interaction. Preacher, Rucker, and Hayes (2007) provide an approach to estimating mediation effects by considering *conditional indirect effects*; Hayes calls models in which the mediated effects are conditional on moderator variable(s) *conditional process models* (2013). Additional background information on estimating mediated and moderated effects can be found in

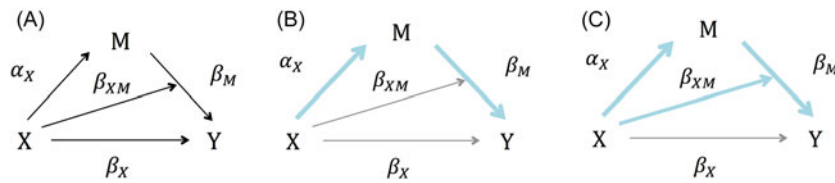
Total Effect			
Controlled Direct Effect	Reference Interaction	Mediated Interaction	Pure Indirect Effect
Controlled Direct Effect	Portion Attributed to Interaction		Pure Indirect Effect
Pure Direct Effect		Mediated Interaction	Pure Indirect Effect
Total Direct Effect			Pure Indirect Effect
Pure Direct Effect		Total Indirect Effect	
Controlled Direct Effect	Reference Interaction	Total Indirect Effect	
Controlled Direct Effect	Portion Eliminated		

**Figure 7.** The two, three, and four-way decompositions of the total effect (TE) into the controlled direct effect (CDE), reference interaction ( $INT_{ref}$ ), mediated interaction ( $INT_{med}$ ), portion attributable to interaction (PAI), pure indirect effect (PIE), pure direct effect (PDE), total direct effect (TDE), total indirect effect (TIE), and the portion eliminated (PE). Formulas for these decompositions are provided and proven formally in T. J. VanderWeele (2015).

Kraemer, Stice, Kazdin, Offord, and Kupfer (2001); Morgan-Lopez and MacKinnon (2006); MacKinnon (2008). The PO approach provides decompositions of the total effect into mediated and moderated components. Figure 7 depicts the various two, three, and four-way decompositions of the total effect (VanderWeele, 2015).

Consider the full model  $E[Y|X, M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$ . With binary  $X$ , the reduced model is  $E[Y|X] = \beta_0^* + \beta_X^* X$ . The CDE ( $\beta_X + \beta_{XM} m$ ) and its variance  $\text{var}(\beta_X) + m^2 \text{var}(\beta_{XM}) + 2m \text{cov}(\beta_X, \beta_{XM})$  are functions of  $m$ . The PE and its variance are functions of the mediator as well. The  $PE = TE - CDE(m) = [\beta_X^* - (\beta_X + \beta_{XM} m)](x - x_o)$  can be estimated using  $[\hat{\Delta} - \hat{\beta}_{XM} m](x - x_o)$ . The variance follows by direct calculation:  $(x - x_o)^2 [V_{XM} V_M^{-1} V_{MX} + m^2 \text{Var}(\hat{\beta}_{XM}) + 2m V_{XM} V_M^{-1} \text{Cov}(\hat{\beta}_M, \hat{\beta}_{XM})]$ . We suggest reporting mediation effects for meaningful values of the moderator, such as the sample mean or quartiles (Aiken, West, & Reno, 1991).

If the exposure is continuous, then the exposure–mediator interaction model above implies a marginal model that includes an  $X^2$  term:  $E[Y|X] = \lambda_0 + \lambda_X X + \lambda_{X^2} X^2$ . To estimate the EMCs  $\Delta^T = [\lambda_X - \beta_X, \lambda_{X^2} - 0]$ , use the



**Figure 8.** Panel A provides a conceptual diagram of the mediation model with an exposure-mediator interaction:  $E[Y|X, M] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM$ . The interaction term  $\beta_{XM}$  can be attributed to either the *moderated effect* or the *mediated effect*. In order to obtain the so-called *pure indirect effect*, one attributes the interaction term to the direct effect of the exposure  $X$ . The pure indirect effect is depicted by the bolded lines in panel B. The *portion eliminated*, which is the portion of the total effect attributed to both interaction and mediation, is depicted by the bolded lines in panel C.

“double sweep” approach described in the Extensions to non-nested mediation systems section. Alternatively, one can use the mediation formula (provided in the [Appendix A.1.3](#)) to proceed with estimation in this setting. Recall that the estimate of the total effect from fitting the marginal model may not equal the implied total effect from summing the natural direct and indirect effects.

Notice that because  $M$  acts simultaneously as a moderator and a mediator, both the direct and indirect effects are affected by the interaction term  $\beta_{XM}$ . As a result, there is more than one way to decompose the total effect (Kraemer et al., 2001). If one attributes  $\beta_{XM}$  to the indirect effect, then the total effect decomposes into the natural (or pure) direct and total indirect effects; if one attributes  $\beta_{XM}$  to the direct effect, then the total effect decomposes into the total direct and pure indirect effects (Pearl, 2001; Robins & Greenland, 1992). The PE does not depend on the choice of decomposition because it is the portion of the total effect attributed to *both* interaction and mediation. Figure 8 shows how the pure indirect effect and the PE measures account for the interaction effect  $\beta_{XM}$  in the presence of an exposure-mediator interaction.

### Examples using data from social psychology research

We provide several examples of mediation models to illustrate the efficiency and coherence of the proposed framework. We compare variance estimates obtained from the model-based formula and percentiles of 5000 bootstrap replications. We also include results from the mediation software by (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014); in models with interactions, we found that 5000 simulations were required to yield results matching the analytical solution (the default is 1000). These examples do not cover all aspects of the analysis process and the results are not meant to be interpreted scientifically. Rather, they are intended to demonstrate how to use the methods discussed throughout the article and to aid

researchers who want to implement the newly proposed approach to mediation analysis. All models assume errors  $\varepsilon \sim N(0, \sigma^2)$ . Unless otherwise specified, we consider unit changes in continuous exposures so that  $(x - x_o) = 1$ .

### Data accessibility

Example: the simple mediation model section uses a subset of data from the Jobs Search Intervention Study (JOBS II) (Vinokur & Schul, 1997) that can be downloaded using the R package mediation <http://www.jstatsoft.org/v59/i05/> (Tingley et al., 2014). The remaining examples use data from several studies described in Hayes (2017). The data is available for download at <http://afhayes.com/introduction-to-mediation-moderation-and-conditional-process-analysis.html>. We provide R code in the Supplement so that anyone can reproduce our results.

### Example: the simple mediation model

Consider  $N = 899$  subjects from the JOBS II study, an experiment that randomly assigned unemployed workers to either treatment (job skills workshops) or control (a booklet with job-search tips). Researchers hypothesized that the workshops would lead to reduced depression score by enhancing unemployed workers' confidence in their ability to find a job. Let  $X = \text{treat}$  be an indicator of whether the patient was randomized to receive treatment or control,  $M = \text{jobseek}$  be a continuous measure of job search self-efficacy, and  $Y = \text{depress2}$  be a continuous measure of depressive symptoms. Baseline covariates include pretreatment depression, age, gender, race, education level, and level of economic hardship.

Does confidence in job-finding ( $\text{jobseek}$ ) mediate the effect of job-skills workshops ( $\text{treat}$ ) on depression levels ( $\text{depress2}$ )? Consider the simple mediation model that includes baseline depression ( $\text{depress1}$ ). The full model is  $\text{depress2} =$



$\beta_0 + \beta_X \text{treat} + \beta_M \text{jobseek} + \beta_C \text{depress1} + \varepsilon$  and the estimated EMC  $\hat{\Delta} = -\hat{V}_{XM} \hat{V}_M^{-1} \hat{\beta}_M = -\text{cov}(\hat{\beta}_X, \hat{\beta}_M) \text{var}(\hat{\beta}_M)^{-1} \hat{\beta}_M$ . The mediated effect of treatment on depression is  $[h(x) - h(x_o)] \hat{\Delta} = (x - x_o) \hat{\Delta} = -0.0103$  (SE = 0.0016) with 95% CI  $-0.0135$  to  $-0.0071$ . The residual bootstrap SE = 0.0017, Sobel's SE = 0.0087, and the case bootstrap SE = 0.0088. Importantly, the model-based estimate of the SE is over five times smaller than the estimated SEs from Sobel's or the case-based bootstrap approximation. Notice that the fully conditional standard error estimate aligns with the residual bootstrap, whereas Sobel's estimate aligns with the case-based bootstrap.

To strengthen the validity of the causal mediation analysis assumptions (outlined in Figure 2), we include additional pretreatment covariates: age, gender, race, education level, and economic hardship. The full model is now specified as  $\text{depress2} = \gamma_0 + \gamma_X \text{treat} + \gamma_M \text{jobseek} + \gamma_C \text{depress1} + \gamma_{C2} \text{age} + \gamma_{C3} \text{gender} + \gamma_{C4} \text{race} + \gamma_{C5} \text{educ} + \gamma_{C6} \text{econhard} + \varepsilon$ . The new estimated indirect effect is  $(x - x_o) \hat{\Delta} = -0.0107$  (SE = 0.0016). The 95% CI is  $(-0.0138, -0.0075)$ , compared to the residual bootstrap  $(-0.0138, -0.0075)$ , the case-based bootstrap  $(-0.0304, 0.0067)$ , and the mediation function  $(-0.0297, 0.0074)$ . The point estimates and standard errors of the mediation effects given by the various approaches are shown in Table 2. For this simple mediation model with no exposure-mediator interaction, the difference of coefficients equals the product of coefficients (and in the PO terminology, the PE equals the NIE). As a result, while the standard error estimates differ, the point estimates for the various measures of mediation are all identical.

### Example: multiple mediators

To demonstrate how the approach proposed by Saunders and Blume (2017) works in the multiple mediator setting, we now consider data from the Media Influence Study, which analyzed subjects' reactions to a newspaper article about a likely sugar shortage (Tal-Or, Cohen, Tsafati, & Gunther, 2010). Half of the subjects were told the article would be published on the front page and the other half were told it would be published in an internal supplement. After reading the article, researchers measured the subjects' beliefs about the article's influence and importance. The presumed media influence (PMI) and the perceived issue importance (import) were two beliefs hypothesized to mediate the relationship between the article's (location) and intentions to buy sugar

(reaction). We fit a multiple mediator model that adjusts for covariates  $C = \{\text{gender}, \text{age}\}$ .

The conceptual diagrams in Figure 9 depict the single-model approach for multiple mediators, the serial multiple mediator model, and the parallel multiple mediator model. For all three methods, the full model is  $\text{reaction} = \beta_0 + \beta_X \text{location} + \beta_{M1} \text{import} + \beta_{M2} \text{PMI} + \beta_C \text{gender} + \beta_{C2} \text{age} + \varepsilon$  and the reduced model for the total effect of article location is  $\text{reaction} = \beta_0^* + \beta_X^* \text{location} + \beta_C^* \text{gender} + \beta_{C2}^* \text{age} + \varepsilon$ . Thus, the direct effect of location is  $\beta_X(x - x_o)$ , the total effect is  $\beta_X^*(x - x_o)$ , and the total indirect effect of location mediated through perceived importance and PMI is  $(\beta_X^* - \beta_X)(x - x_o)$ , regardless of whether the analyst uses the single-model, parallel, or serial approach. Importantly, the total indirect effect does not depend on the order or directionality of the mediators, whereas the amount of mediation attributed *specifically* to perceived importance or PMI will differ across methods due to their varying assumptions about inter-mediator relationships.

The single-model approach allows us to estimate how much perceived importance and PMI mediate the relationship between location and reaction using only the full model and formulas (6) and (7). For  $X = \text{location}$  and  $M = \{\text{import}, \text{PMI}\}$ , the EMCs are

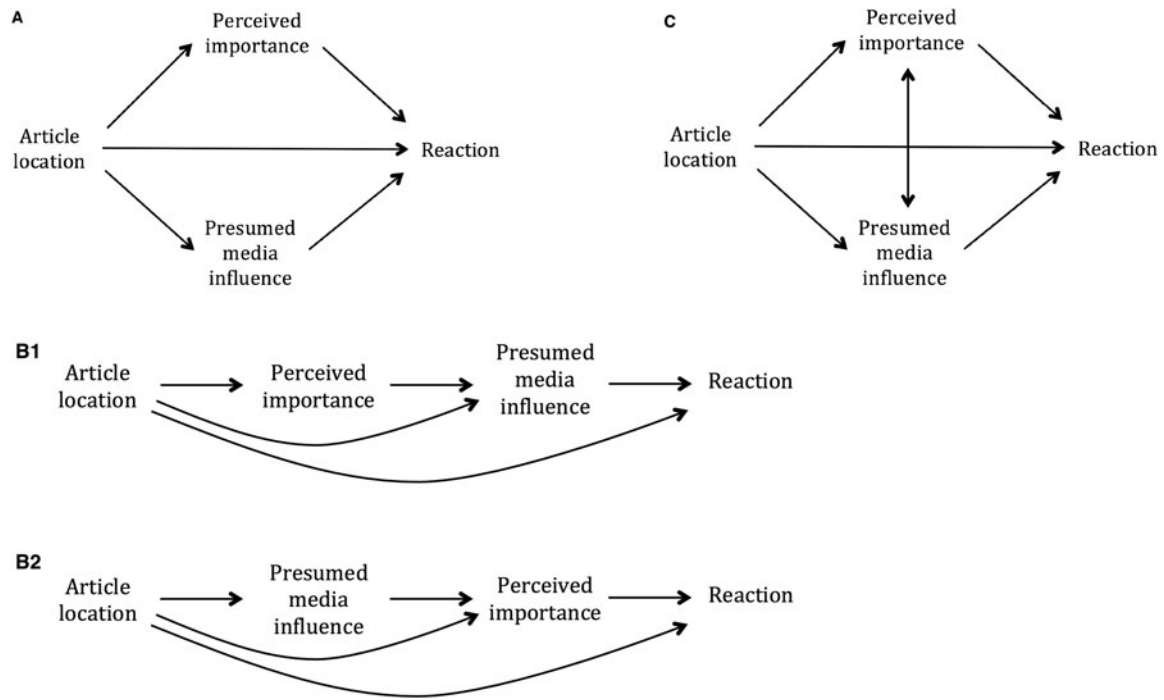
$$\Delta = -V_{XM} V_M^{-1} \beta_M = -[\text{cov}(\beta_X, \beta_{M1}) \quad \text{cov}(\beta_X, \beta_{M2})] \begin{bmatrix} \text{var}(\beta_{M1}) & \text{cov}(\beta_{M1}, \beta_{M2}) \\ \text{cov}(\beta_{M2}, \beta_{M1}) & \text{var}(\beta_{M2}) \end{bmatrix}^{-1} \begin{bmatrix} \beta_{M1} \\ \beta_{M2} \end{bmatrix}.$$

The total indirect effect through  $M$  is estimated by  $[h(x) - h(x_o)] \hat{\Delta} = 0.4053$  and its empirical variance  $\hat{V}_{XM} \hat{V}_M^{-1} \hat{V}_{MX} (x - x_o)^2 = 0.0031$  (SE = 0.0553).

To estimate how much is mediated specifically through  $M_1 = \text{import}$ , we again apply formulas (6) and (7):  $-\hat{V}_{XM1} \hat{V}_{M1}^{-1} \hat{\beta}_{M1} (x - x_o) = -\text{cov}(\hat{\beta}_X, \hat{\beta}_{M1}) \text{var}(\hat{\beta}_{M1})^{-1} \hat{\beta}_{M1} (x - x_o) = 0.1643$ . The variance follows directly:  $\hat{V}_{XM1} \hat{V}_{M1}^{-1} \hat{V}_{M1X} (x - x_o)^2 = 0.0012$  (SE = 0.0350). Similarly, to estimate how much the effect of location is mediated through  $M_2 = \text{PMI}$ , use  $-\hat{V}_{XM2} \hat{V}_{M2}^{-1} \hat{\beta}_{M2} (x - x_o) = 0.1359$ , which has an estimated variance of  $\hat{V}_{XM2} \hat{V}_{M2}^{-1} \hat{V}_{M2X} (x - x_o)^2 = 0.0010$  (SE = 0.0322). Notice that the mediator-specific indirect effects sum to 0.3002, which is less than the total indirect effect of 0.4053 (perceived importance is correlated with PMI,  $r = 0.28$ ).

The parallel approach (Hayes, 2013; MacKinnon, 2008) and the causal regression-based approach (VanderWeele & Vansteelandt, 2013) specify the same full model as above,  $\text{reaction} = \beta_0 + \beta_X \text{location} + \beta_{M1} \text{import} + \beta_{M2} \text{PMI} + \beta_C \text{gender} +$





**Figure 9.** Diagrams comparing the multiple mediator models from the example in Example: multiple mediators section. Model A depicts the single-step or parallel multiple mediator model. Models B1 and B2 depict two serial multiple mediator models. Model C depicts the proposed framework for assessing mediation with multiple mediators from the fit of a single model. The directions of arrows indicate the assumed causal relationships.

$\beta_{C2}age + \varepsilon$ , and an additional model for each mediator:  $import = \alpha_{01} + \alpha_1 location + \alpha_G gender + \alpha_A age + \varepsilon$  and  $PMI = \alpha_{02} + \alpha_2 location + \alpha_{G2} gender + \alpha_{A2} age + \varepsilon$ . This specification assumes the mediators “act in parallel” (see Figure 9 panel A). The delta method is commonly used to estimate standard errors under this approach. The estimated indirect effect through perceived importance is  $\hat{\alpha}_1 \hat{\beta}_{M_1} = 0.2184$  (SE = 0.1144) and the estimated indirect effect through PMI is  $\hat{\alpha}_2 \hat{\beta}_{M_2} = 0.1869$  (SE = 0.1039), which sum to the total indirect effect of 0.4053 (SE = 0.1510).

The serial model given by Hayes (2013) requires specifying the order in which the mediators affect each other. Suppose we assume  $location \rightarrow import \rightarrow PMI \rightarrow reaction$  (see Figure 9 panel B1). The full model is specified as  $reaction = \beta_0 + \beta_X location + \beta_1 import + \beta_2 PMI + \beta_C gender + \beta_{C2} age + \varepsilon$  (the same as above), the first reduced model is  $PMI = \kappa_{02} + \kappa_2 location + \delta_{21} import + \delta_C gender + \delta_{C2} age + \varepsilon$ , and the second reduced model is  $import = \alpha_{01} + \alpha_1 location + \alpha_C gender + \alpha_{C2} age + \varepsilon$ . There are three estimated indirect effects:  $\hat{\alpha}_1 \hat{\beta}_1 = 0.2184$  (SE = 0.1144) is the indirect effect of location through import to reaction,  $\hat{\kappa}_2 \hat{\beta}_2 = 0.1359$  (SE = 0.0982) is the indirect effect of location through PMI to reaction, and  $\hat{\alpha}_1 \hat{\delta}_{21} \hat{\beta}_2 = 0.0510$  (SE = 0.3235) is the indirect effect of location through importance to PMI to reaction.

To demonstrate how mediator-specific indirect effects depend on the specified order in a serial model, suppose we change the order of mediation to  $location \rightarrow PMI \rightarrow import \rightarrow reaction$  (see Figure 9 panel B2). The total indirect effect remains unchanged, but now the indirect effect of location through PMI to reaction is 0.1869, the indirect effect of location through importance to reaction is 0.1643, and the indirect effect of location through PMI to importance to reaction is 0.0541. Notice that in either case, the serially mediated indirect effects sum to the total indirect effect of 0.4053. The point estimates and standard errors of the mediation effects given by the proposed single-model approach, the parallel model, and the serial models are shown in Table 4.

Estimating mediator-specific indirect effects from the serial model is analogous to examining sequential sums of squares. Although the amount of mediation attributed to specific mediators depends heavily on their assumed order, the serially mediated indirect effects always sum to the total indirect effect (as shown in the example above). In contrast, estimating effects from our proposed framework is analogous to examining partial sums of squares. Just as partial sums of squares do not necessarily sum to the total, the mediator-specific indirect effects from our framework do not necessarily sum to the total indirect

**Table 4.** Results from the multiple mediator example in Example: multiple mediators section.

	Parallel (A)	Serial (B1)	Serial (B2)	Single-model (C)
Total Effect	0.5055 (0.2811)	0.5055 (0.2811)	0.5055 (0.2811)	0.5055 (0.2811)
Direct Effect	0.1002 (0.2411)	0.1002 (0.2411)	0.1002 (0.2411)	0.1002 (0.2411)
Portion Eliminated	0.4053 (0.1510)	0.4053 (0.1510)	0.4053 (0.1510)	0.4053 (0.0553)
Indirect $X \rightarrow M_1 \rightarrow Y$	0.2184 (0.1144)	0.2184 (0.1144)	0.1643 (0.1088)	0.1643 (0.0350)
Indirect $X \rightarrow M_2 \rightarrow Y$	0.1869 (0.1039)	0.1359 (0.0982)	0.1869 (0.1039)	0.1359 (0.0322)
Indirect $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$	NA	0.0510 (0.3235)	NA	NA
Indirect $X \rightarrow M_2 \rightarrow M_1 \rightarrow Y$	NA	NA	0.0541 (0.2660)	NA

This table presents mediation effect point estimates with standard errors in parentheses from the parallel, serial, and single-model approaches displayed in Figure 9, with  $M_1 = \text{import}$  and  $M_2 = \text{PMI}$ . In this example, the portion eliminated equals the total indirect effect through the set of mediators  $\{M_1, M_2\}$ . Importantly, the estimates of the total effect, direct effect, and portion eliminated (i.e., total indirect effect) are identical across methods. The mediator-specific indirect effects differ across methods, as discussed in Example: multiple mediators section and Appendix A.3.

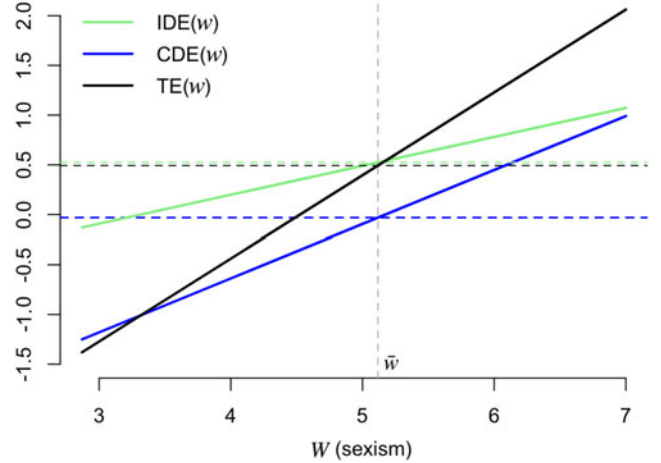
effect. If the mediators are in fact independent, the mediator-specific indirect effects will sum to the total indirect effect.

### Example: the JME

In order to quantify how article location is jointly mediated through perceived importance and PMI, we look at the JME. The standardized total indirect effect of article location through  $M$  is 0.131. The JME:  $r_{YX}\hat{\Delta} = r_{YM}^T\hat{\beta}_M - (R_{Y.XMZ}^2 - R_{Y.XZ}^2) - r_{YZ}^T\hat{\beta}_Z = 0.0210$ . The fraction of the coefficient of determination from the full model ( $R_{Y.XMZ}^2 = 0.3377$ ) accounted for by the JME is 0.0622. That is, the JME accounts for about 6% of the total variation in  $Y$  explained by the full model. Notice that  $R_{Y.XMZ}^2$  (the proportion of variance in the outcome explained by the exposure, mediator, and covariates) is small to begin with, as is the JME.

### Example: interactions

In this section, we consider a study looking at how beliefs about sexism impact women's reactions to discriminatory treatment (Garcia, Schmitt, Branscombe, & Ellemers, 2010). Female study participants ( $N=129$ ) were told that a female attorney lost a promotion to a male candidate who was less qualified due to discriminatory practices of the senior partners. The participants were told either that the attorney confronted the partners or that she did not take action. Researchers then measured how much participants "liked" the attorney, their "perceived appropriateness of the response," and their belief about how widespread sex discrimination is. The hypothesis is that whether or not the attorney protested ( $x = \text{protest}$ ) affected participants' perceptions of her ( $y = \text{liking}$ ), and that this association could be mediated by perceived appropriateness of the response ( $m = \text{appropriate}$ ). Furthermore, we will look at whether the mediated effect is moderated by



**Figure 10.** Plot of the controlled direct effect (CDE( $w$ ), blue line), the indirect effect (IDE( $w$ ), green line), and the total effect (TE( $w$ ), black line) for a unit change in the exposure ( $x - x_0 = 1$  since the exposure is binary) from a mediation model with an exposure-moderator interaction. The solid lines show the mediation effects conditional on the moderator  $W$  and the dashed lines show the effects given the average value of the moderator ( $W = \bar{w}$ ). Notice that for any value of  $W$ , the conditional direct and indirect effects sum to the total effect.

beliefs about the pervasiveness of sex discrimination ( $W = \text{sexism}$ ).

### Exposure-moderator interaction

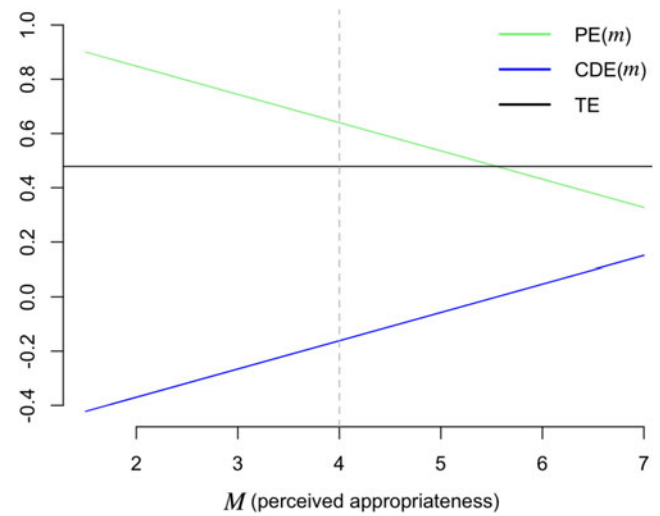
We include an exposure-moderator interaction so that the full model is  $\text{liking} = \beta_0 + \beta_X \text{protest} + \beta_M \text{appropriate} + \beta_W \text{sexism} + \beta_{XW} \text{protest} : \text{sexism} + \varepsilon$ . With  $h(X) = [X, XW]$ , the CDE is conditional on sexism:  $(\beta_X + \beta_{XW}w)(x - x_0) = (-2.8075 + 0.5426w)(x - x_0)$ . The EMCs  $\Delta^T = -[\text{cov}(\beta_X, \beta_M), \text{cov}(\beta_{XW}, \beta_M)]^T \text{var}(\beta_M)^{-1} \beta_M = [\beta_X^* - \beta_X, \beta_{XW}^* - \beta_{XW}]$  are estimated to be  $\hat{\Delta}^T = [-0.9652, 0.2910]$  and the conditional indirect effect  $(x - x_0, xw - x_0w)\Delta = (\beta_X^* - \beta_X)(x - x_0) + (\beta_{XW}^* - \beta_{XW})(xw - x_0w)$  is estimated to be  $-0.9652 + 0.2910w$ . Figure 10 shows the conditional mediation effects as a function of the moderator  $W = \text{sexism}$ .

The indirect effect marginalized over sexism is  $E[(x-x_0, xw-x_0w)\Delta|W] = (\beta_X^* - \beta_X)(x-x_0) + (\beta_{XW}^* - \beta_{XW})(x-x_0)E[W]$ . The variance is estimated using  $(x-x_0)^2\text{Var}(\Delta_1) + (x-x_0)^2E[W]^2\text{Var}(\Delta_2) + 2(x-x_0)^2E[W]\text{Cov}(\Delta_1, \Delta_2)$ . The estimated indirect effect for  $W = \bar{w}$  is  $-0.9652 + 0.2910 * 5.1170 = 0.5238$  ( $SE = 0.1029$ ). The regression-based approach by VanderWeele requires fitting the mediator model appropriate  $= \alpha_0 + \alpha_X\text{protest} + \alpha_W\text{sexism} + \beta_{XW}\text{protest}:\text{sexism} + \varepsilon$  in addition to the full model. The mediation formula estimates the indirect effect using  $E[\hat{\beta}_M(E[M|x] - E[M|x_0])|W] = \hat{\beta}_M(\hat{\alpha}_X + \hat{\alpha}_{XW}E[W])(x-x_0) = 0.5238$  ( $SE = 0.1295$ ).

### Exposure-mediator interaction

Now consider a model with an exposure-mediator interaction:  $\text{liking} = \kappa_0 + \kappa_X\text{protest} + \kappa_M\text{appropriate} + \kappa_{XM}\text{protest}:\text{appropriate} + \varepsilon$ . The  $\text{CDE}(m) = (\kappa_X + \kappa_{XM}m)(x-x_0)$  measures the effect of being told the attorney protested on how much participants liked the attorney, when participants' perceived appropriateness of the response is fixed at some level  $m$ . Using the formula for the EMCs, it is simple to estimate the PE (the difference between the total and CDEs):  $\text{PE}(m) = (\Delta - \kappa_{XM}m)(x-x_0)$ . Figure 11 shows a plot of  $\text{CDE}(m)$  and  $\text{PE}(m)$  for a unit change in  $X$ . One could also report the mediation effects for specific values of  $m$ , such as the sample mean or quartiles. The estimated CDE for the 25th percentile ( $m=4$ ) is  $-0.1616$  ( $SE = 0.2167$ ), and the estimated PE is  $0.6402$  ( $SE = 0.6469$ ).

With an exposure-mediator interaction, the controlled and natural direct effects represent distinct quantities, so the PE and the NIE differ. Estimation of the NDE and NIE requires fitting a separate mediator model:  $\text{appropriate} = \alpha_0 + \alpha_X\text{protest} + \varepsilon_M$ . The natural direct effect  $\text{NDE} = (\kappa_X + \kappa_{XM}E[M|x_0])(x-x_0) = -0.1736$  measures the effect of being told the attorney protested on how much participants liked the attorney, when each participant's perceived appropriateness is fixed to *what it would have been naturally* had they been told the attorney did *not* take action. The NIE measures the change in how much participants liked the attorney when the participants are told she protested, comparing the perceived appropriateness of her response to what it would have been if she had not taken action to what it would have been if she had confronted the attorneys:  $\text{NIE} = \alpha_X(\kappa_M + \kappa_{XM}x)(x-x_0) = 0.6522$ . The natural direct and indirect effects represent quantities that can never actually be observed, which is why they can be



**Figure 11.** Plot of the controlled direct effect ( $\text{CDE}(m)$ , blue line), the portion eliminated ( $\text{PE}(m)$ , green line), and the total effect (TE, black line) for a unit change in the exposure ( $x-x_0 = 1$  since the exposure is binary) from a mediation model with an exposure-mediator interaction. Notice that both the portion eliminated and the controlled direct effect are functions of the mediator  $M$ . At any particular value of  $M$ , the portion eliminated and the controlled direct effect sum to the total effect. The vertical dashed line is at the 25th percentile of the mediator ( $m=4$ ), at which  $\text{CDE}(m) = -0.1616$  and  $\text{PE}(m) = 0.6402$ .

difficult to interpret and arguably have less practical relevance.

### Example: nonlinear exposure effects

To demonstrate how to include nonlinear exposure effects, we use data from a study on economic stress among  $N=262$  entrepreneurs (Pollack, Vanepps, & Hayes, 2012). The hypothesis is that economic stress leads to a depressed affect, which can in turn lead business-persons to withdraw from “entrepreneurial activities.” We adjust for subjects’ age, gender, business tenure, and a self-confidence measure called entrepreneurial self-efficacy (ESE).

### Quadratic exposure effect

To allow for a *quadratic* relationship between stress and withdrawal symptoms, simply specify the full model as  $\text{withdraw} = \beta_0 + \beta_{X_1}\text{stress} + \beta_{X_2}\text{stress}^2 + \beta_M\text{affect} + \beta_Z\text{gender} + \beta_{Z_2}\text{age} + \beta_{Z_3}\text{ESE} + \beta_{Z_4}\text{tenure} + \varepsilon$ . With  $\mathbf{h}(X) = [X, X^2]$ , the EMCs  $\Delta = -\mathbf{V}_{XM}\mathbf{V}_M^{-1}\beta_M = -[\text{cov}(\beta_{X_1}, \beta_M), \text{cov}(\beta_{X_2}, \beta_M)]^T \text{var}(\beta_M)^{-1}\beta_M$ . The EMCs are estimated to be  $[-0.4620, 0.0651]^T$  and the estimated PE (which equals the NIE) is  $[(x-x_0, x^2-x_0^2)]\hat{\Delta} = -0.4620(x-x_0) + 0.0651(x^2-x_0^2) = -0.3970$  ( $SE = 0.0616$ ). Using the mediation package

gives an estimated NIE of  $-0.3970$  (exactly equal to our estimate, as expected) with  $SE = 0.2023$ .

### Splined exposure effect

Suppose instead we wish to model the effect of stress using restricted cubic splines with 4 knots  $k_1, k_2, k_3, k_4 = (2.5, 4, 5.5, 7)$  at the 5th, 35th, 65th, and 95th percentiles of stress. Consider the full model  $\text{withdraw} = \beta_0 + \mathbf{h}(\text{stress})\beta_X + \beta_M \text{affect} + \beta_Z \mathbf{Z} + \varepsilon$  and the marginal model  $\text{withdraw} = \beta_0^* + \mathbf{h}(\text{stress})\beta_X^* + \beta_Z^* \mathbf{Z} + \varepsilon$ , where  $\mathbf{Z}$  is the vector of covariates specified in the previous example and  $\mathbf{h}(X) = [S_1(X), S_2(X), S_3(X)]$  are the splined components of  $X$  given by (Harrell, 2015):

$$\begin{aligned} S_1(X) &= X \\ S_2(X) &= (X - k_1)_+^3 - \frac{(X - k_3)_+^3(k_4 - k_1)}{k_4 - k_3} + \frac{(X - k_4)_+^3(k_3 - k_1)}{(k_4 - k_3)} \\ S_3(X) &= (X - k_2)_+^3 - \frac{(X - k_3)_+^3(k_4 - k_2)}{k_4 - k_3} + \frac{(X - k_4)_+^3(k_3 - k_2)}{(k_4 - k_3)} \end{aligned}$$

To put all basis functions for  $X$  on the same scale, by default the R function divides the terms  $S_j(X)$ , ( $j > 1$ ) by  $\tau = (k_4 - k_1)^{2/3}$ . The EMCs  $\Delta = \beta_X^* - \beta_X = [\beta_1^* - \beta_1, \beta_2^* - \beta_2, \beta_3^* - \beta_3]^T$  can be estimated from the fit of only the full model using formula (6). Thus, the indirect effect is given by

$$\begin{aligned} &[\mathbf{h}(x) - \mathbf{h}(x_o)] \Delta = \\ &[S_1(x) - S_1(x_o), S_2(x) - S_2(x_o), S_3(x) - S_3(x_o)] \begin{bmatrix} \beta_1^* - \beta_1 \\ \beta_2^* - \beta_2 \\ \beta_3^* - \beta_3 \end{bmatrix}. \end{aligned}$$

To estimate the indirect effect comparing  $x =$  the 75th quantile to  $x_o =$  the median of economic stress, we have  $[\mathbf{h}(x) - \mathbf{h}(x_o)] \hat{\Delta}^T = [5.5 - 4.5, 1.3333 - 0.3951, 0.1667 - 0.0062][[-0.0609, 0.3011, -0.1526]]^T = 0.1971$ . The standard error  $([\mathbf{h}(x) - \mathbf{h}(x_o)] \text{var}(\hat{\Delta}) [\mathbf{h}(x) - \mathbf{h}(x_o)]^T)^{1/2} = 0.0266$ .

### Summary

In this paper, we defined the EMCs, provided formulas for estimating mediation effects and their variance from the fit of a single regression model, showed how to visualize mediation effects, and presented a measure of joint mediation. We highlighted situations in which using the difference and product of coefficients approaches do not yield the same estimate of the total effect of the exposure. This suggests that discrepancies between these two approaches' estimates of mediation effects depends on the specification of the marginal model and the estimation of the total effect. Last, we provided extensive examples to illustrate our approach and how it can be applied to complex mediation

hypotheses, including models with multiple mediators, interactions, and nonlinearities.

The statistical literature abounds with methods for measuring mediation in the simple setting of one exposure, one mediator, and one outcome. However, scientific mediation hypotheses typically involve a more complicated interplay between several variables. Rather than estimating the total, direct, and indirect effects from separate regression equations, one can use the simple EMC formula to estimate mediation effects and their variance. We recommend using our formula to obtain estimates of the PE (and in several settings, the NIE). This approach provides an analytical variance and reduces computation time. For estimating the pathway decompositions displayed in Figure 7, we recommend using the formulas given by (VanderWeele, 2015). When estimating mediation effects, one should thoughtfully consider the plausibility of the assumptions required for causal inference.

### Article information

**Conflict of interest disclosures:** Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

**Ethical principles:** The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding:** This work was not supported.

**Role of the funders/sponsors:** None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

**Acknowledgments:** We thank Professors Jonathan Schildcrout, Robert Johnson, and Melinda Aldrich for critical reading, helpful suggestions, and valuable feedback on prior versions of this manuscript. We are also grateful for the constructive comments from the associate editor and the reviewer, which ultimately improved the presentation of our ideas. The ideas and opinions expressed herein are those of the authors



alone, and endorsement by the authors' institutions is not intended and should not be inferred.

## ORCID

Christina T. Saunders  <http://orcid.org/0000-0003-4325-9568>

## References

- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, California: Sage Publications.
- 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
- 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), 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 17, 37–69. doi:10.2307/271028
- Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer. doi:10.1007/978-94-007-6094-3\_15
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 558–577. doi:10.1037/0021-843X.112.4.558
- de Heus, P. (2012). R squared effect-size measures and overlap between direct and indirect effect in mediation analysis. *Behavior Research Methods*, 44(1), 213–221. doi:10.3758/s13428-011-0141-5
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233–239. doi:10.1111/j.1467-9280.2007.01882.x
- Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 1, 151–182. doi:10.1023/A:1009602825894
- Garcia, D. M., Schmitt, M. T., Branscombe, N. R., & Ellemers, N. (2010). Women's reactions to ingroup members who protest discriminatory treatment: The importance of beliefs about inequality and response appropriateness. *European Journal of Social Psychology*, 40(5), 733–745. doi:10.1002/ejsp.644
- Gelfand, L. A., Mensinger, J. L., & Tenhave, T. (2009). Mediation analysis: A retrospective snapshot of practice and more recent directions. *Journal of General Psychology*, 136(2), 153–176. doi:10.3200/GENP.136.2.153-178
- Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55(292), 708–713. doi:10.1080/01621459.1960.10483369
- Goodnight, J. H. (1979). A tutorial on the SWEEP operator. *The American Statistician*, 33, 149–158. doi:10.1080/00031305.1979.10482685
- Gunzler, D., Chen, T., Wu, P., & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai Archives of Psychiatry*, 25(6), 390–394. doi:10.3969/j.issn.1002-0829.2013.06.009
- Harrell, F. E. (2015). *Regression Modeling Strategies* (2nd ed.). Springer. doi:10.1007/978-3-319-19425-7
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4), 408–420. doi:10.1080/03637750903310360
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press. doi:10.1111/jedm.12050
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). Guilford Press.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960. doi:10.2307/2289064
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334. doi:10.1037/a0020761
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: estimating mediation in treatment evaluations. *Evaluation Review*, 5(5), 602–619. doi:10.1177/0193841X8100500502
- Kraemer, H., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, 158(6), 848–856. doi:10.1176/appi.ajp.158.6.848
- Last, J. M. (1988). *In a dictionary of epidemiology* (2nd ed.). New York: Oxford University Press. doi:10.1093/aje/154.4.389
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Taylor & Francis Group. doi:10.4324/9780203809556
- MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, 39(3), 384. doi:10.3758/BF03193007
- 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. doi:10.1207/s15327906mbr3901\_4
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, 30(1), 41. doi:10.1207/s15327906mbr3001\_3
- McDonald, R. P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research*, 32(1), 1–38. doi:10.1207/s15327906mbr3201\_1
- Meinert, C. L. (1986). *In clinical trials: Design, conduct, and analysis* (p. 285). Oxford University Press.
- Morgan-Lopez, A. A., & MacKinnon, D. P. (2006). Demonstration and evaluation of a method for assessing



- mediated moderation. *Behavior Research Methods*, 38(1), 77–87. doi:[10.3758/BF03192752](https://doi.org/10.3758/BF03192752)
- Naimi, A. I., Kaufman, J. S., & MacLehose, R. F. (2014). Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. *International Journal of Epidemiology*, 43(5), 1656–1661. doi:[10.1093/ije/dyu107](https://doi.org/10.1093/ije/dyu107)
- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th conference in uncertainty in artificial intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4), 426–436. doi:[10.1007/s11121-011-0270-1](https://doi.org/10.1007/s11121-011-0270-1)
- Pearl, J. (2012b). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, & L. Bernard (Eds.), *Causality: Statistical perspectives and applications* (pp. 151–179). John Wiley and Sons. doi:[10.1007/s11121-011-0270-1](https://doi.org/10.1007/s11121-011-0270-1)
- Pollack, J. M., Vanepps, E. M., & Hayes, A. F. (2012). The moderating role of social ties on entrepreneurs' depressed affect and withdrawal intentions in response to economic stress. *Journal of Organizational Behavior*, 33(6), 789–810. doi:[10.1002/job.1794](https://doi.org/10.1002/job.1794)
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *The Annual Review of Psychology*, 66, 4.1–4.28. doi:[10.1146/annurev-psych-010814-015258](https://doi.org/10.1146/annurev-psych-010814-015258)
- Preacher, K. J., & Hayes, A. F. (2008a). 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](https://doi.org/10.3758/BRM.40.3.879)
- Preacher, K. J., & Hayes, A. F. (2008b). Contemporary approaches to assessing mediation in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *Advanced data analysis methods for communication research* (pp. 13–54). SAGE Publications, Inc. doi:[10.4135/9781452272054.n2](https://doi.org/10.4135/9781452272054.n2)
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1), 185–227. doi:[10.1080/002731170701341316](https://doi.org/10.1080/002731170701341316)
- 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](https://doi.org/10.1097/00001648-199203000-00013)
- Saunders, C. T., & Blume, J. D. (2017). A classical regression framework for mediation analysis: fitting one model to estimate mediation effects. *Biostatistics*. doi:[10.1093/biostatistics/kxx054](https://doi.org/10.1093/biostatistics/kxx054)
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290–312. doi:[10.2307/270723](https://doi.org/10.2307/270723)
- Springer, M. D., & Thompson, W. E. (1966). The distribution of products of independent random variables. *SIAM Journal on Applied Mathematics*, 14(3), 511–526. doi:[10.1137/0114046](https://doi.org/10.1137/0114046)
- Stone, C. A., & Sobel, M. E. (1990). The robustness of estimates of total indirect effects in covariance structure models estimated by maximum likelihood. *Psychometrika*, 55(2), 337–352. doi:[10.1007/BF02295291](https://doi.org/10.1007/BF02295291)
- Tal-Or, N., Cohen, J., Tsfaty, Y., & Gunther, A. (2010). Testing causal direction in the influence of presumed media influence. *Communication Research*, 37(6), 801–824. doi:[10.1177/0093650210362684](https://doi.org/10.1177/0093650210362684)
- 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](https://doi.org/10.1177/1094428107300344)
- Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5), 1–38. doi:[10.18637/jss.v059.i05](https://doi.org/10.18637/jss.v059.i05)
- Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, 43(3), 692–700. doi:[10.1037/0003-066X.54.8.594](https://doi.org/10.1037/0003-066X.54.8.594)
- Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2), 137–150. doi:[10.1037/a0031034](https://doi.org/10.1037/a0031034)
- VanderWeele, T., & Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1), 95–115. doi:[10.1515/em-2012-0010](https://doi.org/10.1515/em-2012-0010)
- VanderWeele, T. J. (2013). Policy-relevant proportions for direct effects. *Epidemiology*, 24(1), 175–176. doi:[10.1097/EDE.0b013e3182781410](https://doi.org/10.1097/EDE.0b013e3182781410)
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.
- Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation Strategies for the Estimation of Natural Direct and Indirect Effects. *Epidemiologic Methods*, 1(1). doi:[10.1515/2161-962x.1014](https://doi.org/10.1515/2161-962x.1014)
- Vinokur, A. D., & Schul, Y. (1997). Mastery and inoculation against setbacks as active ingredients in the jobs intervention for the unemployed. *Journal of Consulting and Clinical Psychology*, 65(5), 867–877. doi:[10.1037/0022-006X.65.5.867](https://doi.org/10.1037/0022-006X.65.5.867)
- Woodworth, R. S. (1928). Dynamic Psychology. In *Psychologies of 1925* (pp. 111–126). Clark University Press. doi:[10.1037/11020-005](https://doi.org/10.1037/11020-005)
- Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197–206. doi:[10.1086/651257](https://doi.org/10.1086/651257)

## A. Appendix

### A.1. Approaches to estimating the indirect effect

#### A.1.1. Baron and Kenny's causal steps

Baron and Kenny published their landmark paper on assessing mediation from the simple mediation model using the *causal steps approach* in (1986). Their approach says that before estimating the indirect effect and its variance, the variables must be significant in a series of hypothesis tests:  $X$  must affect  $M$  in Equation (2),  $X$  must affect  $Y$  in Equation (3), and  $M$  must affect  $Y$  in Equation (1). If the causal steps are established, one estimates the indirect effect and tests for

its significance using the Sobel test (Sobel, 1982). If the exposure has no effect when the mediator is controlled (if  $\beta_X = 0$ ), then there is “strong evidence for a single, dominant mediator,” or so-called perfect mediation. If  $\beta_X \neq 0$ , this is termed *partial mediation* and indicates “the operation of multiple mediating factors” (Baron & Kenny, 1986).

Although Baron and Kenny’s (1986) article is considered a cornerstone of mediation analysis (three decades later, researchers have cited their method over 70,000 times), flaws in the causal steps approach have been presented (Fritz & MacKinnon, 2007; Gelfand et al., 2009; Hayes, 2009; 2013; MacKinnon et al., 2002; Preacher & Hayes, 2008b; Vansteelandt, Bekaert, & Lange, 2012; Zhao et al., 2010). If an indirect effect exists and there is *inconsistent mediation* (the direct effect and indirect effects of  $X$  on  $Y$  have similar magnitudes and opposite signs, leading to a total effect near zero), using the causal steps approach would stop the analysis at the second step. Plus, a nonzero association between  $X$  and  $Y$  reducing to zero when a third variable  $Z$  is covaried does not necessarily mean that  $Z$  mediates the effect of  $X$  on  $Y$ . Furthermore, the terms “partial” and “complete” mediation are defined in terms of statistical significance (Preacher & Hayes, 2008b) and concluding “complete mediation” may inhibit future research into other possible mediators. We recommend disregarding the causal steps approach (because it places too much emphasis on statistical significance); instead, one should carefully construct a mediation hypothesis, consider the assumptions required for causal inference, and report mediation effects and their confidence intervals.

### A.1.2. The product and difference of coefficients approaches

For a unit change in the exposure, the estimated total effect of  $X$  is  $\hat{\beta}_X^*$ , the coefficient for  $X$  in Equation (3), and the estimated direct effect of  $X$  is  $\hat{\beta}_X$ , the coefficient for  $X$  in Equation (1). The difference of coefficients approach estimates the indirect effect for a unit change in  $X$  by subtracting the direct effect from the total effect:  $\hat{\beta}_X^* - \hat{\beta}_X$ . The product of coefficients approach estimates the indirect effect by multiplying the coefficient for  $X$  in Equation (2) by the coefficient for  $M$  in Equation (1):  $\hat{\alpha}_X \hat{\beta}_M$ . For the simple mediation model with continuous  $M$  and  $Y$ , the product and difference of coefficients approaches agree and  $\hat{\beta}_X^* - \hat{\beta}_X = \hat{\alpha}_X \hat{\beta}_M$ . This leads to a nice interpretation of  $\hat{\beta}_X^* = \hat{\beta}_X + \alpha_X \hat{\beta}_M$ : the total effect of  $X$  on  $Y$  equals the sum of the direct and indirect effects. Although the product and difference of coefficients approaches agree for linear models, in general the two approaches and their interpretation may differ (Pearl, 2012b) and there is disagreement as to which approach is

preferable (Alwin & Hauser, 1975; Imai et al., 2010; Preacher & Hayes, 2008b).

### A.1.3. The PO framework

A formal approach to mediation analysis based on the PO framework has been developed (Holland, 1986; Pearl, 2001; Robins & Greenland, 1992). Causal mediation effects are defined as contrasts in average PO that depend on both the exposure and mediator variables. Let  $Y(x, m)$  be the PO that would be observed if the exposure  $X$  were equal to  $x$  and the mediator  $M$  were equal to  $m$ . Let  $Y(x, M_{x_o})$  be the PO that would be observed if the exposure were equal to  $x$  but the mediator  $M$  were equal to the value it would have been if the exposure were equal to  $x_o$ . Note that the PO  $Y(x, M(x))$  and  $Y(x_o, M(x_o))$  are observable, but  $Y(x, M(x_o))$  and  $Y(x_o, M(x))$  can never be observed, and thus are always counterfactual. The counterfactual definitions of causal mediation effects are listed in Table 1. By using the  $(x, x_o)$  notation, we make explicit that causal mediation effects are generally defined for any two levels of the exposure. When  $X$  is a binary exposure, the only possible pair of values is (0, 1).

Until now, we have discussed total, direct, and indirect effects. However, the causal mediation literature distinguishes between *controlled* and *natural* effects. The CDE measures the effect of  $X$  on  $Y$  while holding the mediator fixed at level  $m$  for everyone in the population:  $CDE(x, x_o, m) \equiv Y(x, m) - Y(x_o, m)$ . The natural direct effect measures the effect of the exposure on the outcome when each individual’s mediator is fixed to  $M(x_o)$ , what it would have been “naturally” had the exposure been absent (or equal to some referent value):  $NDE(x, x_o) \equiv Y(x, M(x_o)) - Y(x_o, M(x_o))$ . The natural indirect effect represents the difference in the outcome if one holds the exposure at level  $x$  and changes the mediator from the value that would have been observed under the referent exposure,  $M(x_o)$ , to the value that would have been observed under treatment,  $M(x)$ :  $NIE(x, x_o) \equiv Y(x, M(x)) - Y(x, M(x_o))$ . Regardless of how the direct and indirect effects are defined, the total effect of  $X$  on  $Y$  is  $TE(x, x_o) \equiv Y(x) - Y(x_o)$ . Causal effects cannot be estimated at the individual level, but one may estimate average causal effects by taking the expectation of the causal contrasts.

The controlled and natural direct effects diverge in the presence of exposure-mediator interactions. From  $E[Y|X, M, C] = \beta_0 + \beta_X X + \beta_M M + \beta_{XM} XM + \beta_C C$ , the CDE of  $X$  is estimated by  $E[Y(x, m) - Y(x_o, m)|C] = (\beta_X + \beta_{XM} m)(x - x_o)$ . The natural direct effect is estimated by  $E[Y(x, M(x_o)) - Y(x_o, M(x_o))|C] = (\beta_X + \beta_{XM} E[M|x_o])(x - x_o) = (\beta_X + \beta_{XM}(\alpha_0 + \alpha_X x^* + \alpha_C C))(x - x_o)$ . To estimate the natural indirect effect, one uses  $E[Y(x, M(x)) - Y(x, M(x_o))|C] = \alpha_X(\beta_M + \beta_{XM} x)(x - x_o)$ .

Because the total effect can always be broken down into the natural direct and indirect effects, the natural indirect effect can be written as the difference between the total and natural direct effects:  $NIE = TE - NDE$ . Another important quantity is the PE, which is the difference between the total and CDEs:  $PE = TE - CDE$  (VanderWeele, 2015). Note that the PE should not be confused with the *proportion or percentage mediated* (PM). Whereas the PE is the difference between the total and controlled direct effects ( $PE = TE - CDE$ ), the proportion mediated is the ratio of the natural indirect effect to the total effect ( $PM = NIE/TE$ ). A large sample size is required to obtain stable estimates of the proportion mediated (MacKinnon et al., 1995). Additionally, when the indirect and total effects have opposite signs, the proportion mediated will be negative (in which case it is not actually a percentage). For these reasons, we recommend the proportion mediated be used with caution.

For the simple mediation model without an exposure–mediator interaction, the CDE equals the natural direct effect; as a result, the portion eliminated ( $PE = TE - CDE$ ) equals the natural indirect effect ( $NIE = TE - NDE$ ). In general, however, the difference between the total effect and the CDE is not equal to the natural direct effect (i.e., the PE is not equal to the natural indirect effect). If the full model includes an exposure–mediator interaction, the PE may be nonzero due to an interaction effect, not due to mediation. As shown in Figure 7, the PE equals the PIE (pure indirect effect) + PAI (portion attributable to interaction), or alternatively, TIE (total indirect effect) +  $INT_{ref}$  (reference interaction). Even if the exposure does not affect the mediator (such that there is no mediation of  $X$  by  $M$ ), there could be interaction effects that the PE captures.

The *mediation formula* is a generalization of the product of coefficients approach and can be used to estimate causal mediation effects from any type of model (Imai et al., 2010; Pearl, 2001; 2012a). Let  $P[\cdot]$  denote the probability mass function. The mediation formula estimates the natural direct effect, natural indirect effect, and the total effect using:

$$\begin{aligned} NDE(x, x_o) &= \sum_{c,m} (E[Y|x, m, c] - E[Y|x_o, m, c])P[m|x_o, c]P[c] \\ NIE(x, x_o) &= \sum_{c,m} E[Y|x, m, c](P[m|x, c] - P[m|x_o, c])P[c] \\ TE(x, x_o) &= \sum_c (E[Y|x, c] - E[Y|x_o, c])P[c] \end{aligned}$$

#### A.1.4. The structural equation modeling framework

The language of structural equation modeling is often used by social scientists for conducting mediation analysis. We briefly mention a few basic characteristics of SEM; a thorough and technical treatment of using SEM for mediation analysis is given in (Bollen, 1987).

Structural equation models distinguish between observed and latent (unobserved) variables, as well as endogenous and exogenous variables. Endogenous

variables are affected by other variables, whereas exogenous variables only affect other variables, without being affected themselves. Furthermore, SEMs make use of a measurement model and a structural model, from which effects are estimated simultaneously. The measurement model specifies the relationship between latent variables and measured indicator variables, and the structural model specifies the causal relationships among the variables and their covariance structure.

SEM uses path diagrams to graphically display the theoretical causal relationships: rectangles represent observed variables, ovals represent latent variables, straight unidirectional arrows show causal effects between variables, and curved bidirectional arrows represent covariance between two variables. The absence of a link between two variables is important – it represents an assumed *lack* of a causal relationship. The simple mediation model is an example of a SEM with observed exposure, mediator, and outcome variables and uncorrelated errors. The exposure is exogenous, the mediator is endogenous with respect to the exposure and exogenous with respect to the outcome, and the outcome variable is endogenous.

Structural equation modeling often represents the mediation model using a matrix equation  $\eta = \beta\eta + \gamma\xi + \zeta$ , where  $\eta$  is the vector of dependent (endogenous) variables,  $\xi$  is the vector of independent (exogenous) variables,  $\zeta$  is the vector of residuals, and  $\beta$  and  $\gamma$  are matrices of coefficients. For complex models with multiple mediators and multiple exposures, a matrix equation represents “the same information” as a system of equations, but it helps “simplify the organization” (MacKinnon, 2008). For example, using the SEM notation of MacKinnon (2008), the following is a matrix representation of the simple mediation model with one exposure  $\xi$ , one mediator  $\eta_1$ , and one outcome  $\eta_2$ :  $\eta = \beta\eta + \gamma\xi + \zeta$

$$\zeta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \xi + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}. \quad \text{Notice}$$

that this matrix equation can be written as two separate equations:  $\eta_1 = \gamma_1\xi + \zeta_1$  and  $\eta_2 = \beta_{21}\eta_1 + \gamma_2\xi + \zeta_2$ . Rewriting the matrix equation using our notation

$$\begin{bmatrix} M \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \beta_M & 0 \end{bmatrix} \begin{bmatrix} M \\ Y \end{bmatrix} + \begin{bmatrix} \alpha_X \\ \beta_X \end{bmatrix} X + \begin{bmatrix} \varepsilon_M \\ \varepsilon_Y \end{bmatrix} \quad \text{makes it}$$

clear that this is simply another way to write (1–2):  $M = \alpha_X X + \varepsilon_M$  and  $Y = \beta_X X + \beta_M M + \varepsilon_Y$ . Whether the system is written as separate regression equations or as a single matrix equation, more than one equation must still be fit.

#### A.2. Existing approaches to estimating the variance of mediation effects

The methods commonly used for approximating the variance of the estimated indirect effect are based on



the multivariate delta method, bootstrapping, and Monte Carlo simulation. Now that an analytical solution for the variance exists, it is of interest to reexamine the behavior of these approximations. Simulations in Saunders and Blume (2017) shed light on the efficiency gains inherent in avoiding conservative approximations.

### A.2.1. Delta method approximations

Sobel (1982) proposed the multivariate delta method (or first order Taylor series) approximation to the variance of the indirect effect for the simple mediation model:  $\text{Var}(\hat{\alpha}_X \hat{\beta}_M)_{\text{Sobel}} = \hat{\alpha}_X^2 s_{\beta_M}^2 + \hat{\beta}_M^2 s_{\alpha_X}^2$ . The second order Taylor series approximation is  $\text{Var}(\hat{\alpha}_X \hat{\beta}_M)_{\text{Exact}} = \hat{\alpha}_X^2 s_{\beta_M}^2 + \hat{\beta}_M^2 s_{\alpha_X}^2 + s_{\alpha_X}^2 s_{\beta_M}^2$ , but the  $s_{\alpha_X}^2 s_{\beta_M}^2$  term tends to be trivially small in practice and is often omitted from the standard error calculation (MacKinnon et al., 1995). An unbiased variance estimator subtracts rather than adds  $s_{\alpha_X}^2 s_{\beta_M}^2$  in the equation above (Goodman, 1960) but can result in a negative value for the standard error (so it is not recommended) (MacKinnon et al., 2002). The derivation of these three methods assumes independence of  $\alpha_X$  and  $\beta_M$ . Valeri and VanderWeele (2013) derived delta method variance approximations for several more complex mediation models.

A disadvantage of the above delta method approximations is their reliance on the sampling distribution of  $\hat{\alpha}_X \hat{\beta}_M$  being normal. In practice,  $\hat{\alpha}_X \hat{\beta}_M$  tends to be skewed and highly leptokurtic (MacKinnon, Lockwood, & Williams, 2004; MacKinnon et al., 2002). The Sobel test of the indirect effect compares the statistic  $(\hat{\alpha}_X \hat{\beta}_M) / \text{SE}(\hat{\alpha}_X \hat{\beta}_M)$  to a standard normal distribution. The Sobel confidence intervals tend to lie to the left of the true value for positive indirect effects, and to the right for negative indirect effects (MacKinnon et al., 1995; Stone & Sobel, 1990). As a result, the Sobel test has less power than expected to detect a true indirect effect because the 95% CI will often improperly include zero (MacKinnon et al., 2004).

### A.2.2. Distribution of the product method

Rather than assuming the product  $\hat{\alpha}_X \hat{\beta}_M$  is normally distributed, the *distribution of the product* method assumes  $\hat{\alpha}_X \sim N(\alpha_X, \sigma_{\alpha_X}^2)$  and  $\hat{\beta}_M \sim N(\beta_M, \sigma_{\beta_M}^2)$ , and uses the analytical distribution of the product of two normal random variables (MacKinnon et al., 2004). The distribution of the product method yields confidence intervals that are asymmetric (MacKinnon, Fritz, Williams, & Lockwood, 2007; Tofighi & MacKinnon, 2011). The assumption of normality of the sampling distributions of  $\hat{\alpha}_X$  and  $\hat{\beta}_M$  is arguably more realistic than the assumption of normality of the distribution of their product. After all, the coefficient estimates properly scaled have a  $t$ -distribution. The

form of the distribution of the product is highly complex, but values of the function under the null condition that  $\hat{\alpha}_X = \hat{\beta}_M = 0$  are tabulated in (Springer & Thompson, 1966). Although there are tables that do not assume both  $\hat{\alpha}_X$  and  $\hat{\beta}_M$  are zero, for hypothesis testing their use still requires assumptions about their true value, information that is not usually available. The distribution of the product approach relies on large samples for accurate approximation.

### A.2.3. Bootstrapping

The bootstrap method estimates the variance of mediation effects from the empirical sampling distribution of the estimates. Bootstrapping handles asymmetric sampling distributions better than the delta method and thus improves the accuracy of confidence limits (Preacher & Hayes, 2008b; Valeri & VanderWeele, 2013). There are two approaches to bootstrap resampling in regression, observation resampling and residual resampling. Observation resampling is not model dependent and treats the design matrix as random by resampling cases (i.e., the rows in a design matrix). By contrast, residual resampling treats the design matrix as fixed, is model dependent, and does not maintain the  $(X, M, Y)$  association. Bootstrapping cases usually gives a larger estimate of the variance since it allows for more sources of variation from the randomness in the design matrix. As the sample size grows, both methods become similar, assuming the model is correctly specified.

The case-based bootstrap approach to estimating the variance of the indirect effect proceeds as follows. From the data  $(X, M, Y)$  of sample size  $N$ , draw with replacement  $N$  observations to create a bootstrap sample  $B^* = (X^*, M^*, Y^*)$ . From  $B^*$ , estimate the indirect effect using either the product or difference of coefficients approach. Repeat this process  $M > 5000$  times. The distribution of the  $M$  bootstrap estimates of the indirect effect provides an empirical, nonparametric approximation to the sampling distribution of the indirect effect. Obtain the point estimate and the standard deviation of the indirect effect from the mean and standard error of the  $M$  mediation effect estimates, respectively. A 95% percentile confidence interval is constructed from the 2.5th and 97.5th percentiles of the empirical distribution.

Under the three-equation system of the simple mediation model, bootstrapping residuals is complicated. Since there are three equations, one might think to bootstrap the residuals from each model separately. This, however, leads to inconsistent results. To bootstrap residuals, fit the full model and save the fitted values  $\hat{Y}$  and residuals  $e$ . Sample with replacement from the residuals  $e$  to get  $e^*$  and a new outcome variable  $Y^* = \hat{Y} + e^*$ . To estimate the indirect effect using the difference of coefficients approach, re-fit the

full and reduced models as follows:  $Y^* = \beta_0 + \beta_X X + \beta_M M$  and  $Y^* = \gamma_0 + \gamma_X X$  and store  $\gamma_X - \beta_X$ . Use the distribution of  $\gamma_X - \beta_X$  for inference. To estimate the indirect effect using the product of coefficients approach, fit  $M = \alpha_0 + \alpha_X X$  and multiply  $\alpha_X$  by  $\beta_M$  from the bootstrapped full model. For the simple mediation model, the residual bootstrap distributions of the estimated indirect effect will be identical under both approaches.

#### A.2.4. The Monte Carlo method

Monte Carlo methods estimate the variance by simulating the sampling distribution of mediation effects (MacKinnon et al., 2004). First, estimate the coefficients used in calculating the indirect effect and their standard errors. For example, if using the product of coefficients approach to estimate the indirect effect, obtain the estimates  $\hat{\alpha}_X, \hat{\beta}_M, s_{\alpha_X}^2$  and  $s_{\beta_M}^2$ . Next, generate  $S > 5,000$  random samples of the product  $\alpha_X^* \beta_M^*$  based on  $\alpha_X^* \sim N(\hat{\alpha}_X, s_{\alpha_X}^2)$  and  $\beta_M^* \sim N(\hat{\beta}_M, s_{\beta_M}^2)$ . To allow  $\hat{\alpha}_X$  and  $\hat{\beta}_M$  to covary, specify a bivariate normal distribution with some covariance. Obtain the lower and upper confidence limits for the indirect effect from the percentiles of the simulated sampling distribution of the indirect effect. The same general procedure holds for the difference of coefficients approach. We do not recommend using the Monte Carlo approach to estimate the variance unless one has the coefficient and standard error estimates but the raw data are unavailable.

### A.3. Existing approaches to multiple mediator models

#### A.3.1. Parallel (or single-step) models

The parallel mediator model specifies a separate model for each mediator in which they independently affect the outcome (Hayes, 2013; MacKinnon, 2008):

$$\begin{aligned} E[Y|X, \mathbf{M}] &= \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2 + \dots + \beta_j M_j \\ E[M_i|X] &= \alpha_{0i} + \alpha_i X, i = 1, \dots, j \\ E[Y|X] &= \beta_0^* + \beta_X^* X \end{aligned}$$

Analogous to the simple mediation model, the total and direct effects for a unit change in  $X$  are given by the coefficients  $\beta_X^*$  and  $\beta_X$ , respectively. This approach assumes no mediators affect each other. The specific indirect effect of  $X$  on  $Y$  through  $M_i$  is quantified as  $\alpha_i \beta_i$  (MacKinnon, 2008). If the independence of mediators assumption holds, the total indirect effect of  $X$  on  $Y$  is the sum of the specific indirect effects,  $\sum_i (\alpha_i \beta_i)$ ,  $i = 1 \dots j$ , which equals  $\beta_X^* - \beta_X$  for ordinary least squares regression with continuous  $M$  and  $Y$ . In this case, the total effect of  $X$  on  $Y$  can be written as the sum of the direct effect and all  $j$  mediator-specific indirect effects:  $\beta_X^* = \beta_X + \sum_i (\alpha_i \beta_i)$ ,  $i = 1 \dots j$ . This approach traditionally uses delta method

approximations to estimate the variance of mediator-specific indirect effects and the total indirect effect. The formulas for a two-mediator model are  $\hat{\alpha}_1^2 s_{\beta_{M_1}}^2 + \hat{\beta}_{M_1}^2 s_{\alpha_1}^2$ ,  $\hat{\alpha}_2^2 s_{\beta_{M_2}}^2 + \hat{\beta}_{M_2}^2 s_{\alpha_2}^2$ , and  $s_{\alpha_1}^2 \hat{\beta}_{M_1}^2 + s_{\beta_{M_1}}^2 \hat{\alpha}_1^2 + s_{\alpha_2}^2 \hat{\beta}_{M_2}^2 + s_{\beta_{M_2}}^2 \hat{\alpha}_2^2 + 2\hat{\alpha}_1 \hat{\alpha}_2 s_{\beta_{M_1} \beta_{M_2}}$ , respectively.

VanderWeele and Vansteelandt (2013) provide a regression-based approach for multiple mediators that is similar in spirit to the single-step multiple mediator model. This approach specifies one regression for the outcome  $Y$  (regress  $Y$  on  $X, M$ , and  $Z$ ), and a separate model for each mediator and each mediator-mediator interaction. Both the natural and CDEs are given by the coefficient for the  $X$  in the full model. The natural indirect effect is equal to the sum over the  $j$  mediators of the product of the coefficient for the exposure in the model for the  $i$ th mediator and the coefficient for the  $i$ th mediator in the full model. Including covariates  $C$  can lead to compatibility issues between the models for  $M_i$ ,  $M_k$ , and their product  $M_i M_k$ . Their alternative inverse probability weighting approach circumvents this issue in settings with mediator-mediator interactions.

#### A.3.2. Weighting approach

The weighting approach does not require modeling the mediators, allows the mediators to affect each other, and can be used for essentially any type of outcome and mediators, although it performs best when the exposure has only a few levels (e.g., binary or discrete) (VanderWeele & Vansteelandt, 2013). Obtaining the weights requires fitting several logistic regression models to estimate the probabilities  $P[X = x]$ ,  $P[X = x_o]$ ,  $P[X = x|C = c]$ ,  $P[X = x_o|C = c]$ . They recommend bootstrapping the variance for both the regression-based and weight-based approaches.

#### A.3.3. Serial models

The *serial multiple mediator model* requires the researcher to specify the order in which the mediators affect each other. Like the single-step multiple mediator model, this approach specifies a separate model for each mediator, although now each mediator depends on those that precede them temporally in the causal chain (Hayes, 2013; MacKinnon, 2008; Taylor, MacKinnon, & Tein, 2008). The model is specified as

$$\begin{aligned} E[Y|X, \mathbf{M}] &= \beta_0 + \beta_X X + \beta_1 M_1 + \beta_2 M_2 + \dots + \beta_j M_j \\ E[M_i|X, M_1, \dots, M_{i-1}] &= \kappa_{0i} + \kappa_i X + \sum_{k=1}^{i-1} \delta_{ik} M_k, i = 2, \dots, j \\ E[M_1|X] &= \alpha_{01} + \alpha_1 X \\ E[Y|X] &= \beta_0^* + \beta_X^* X \end{aligned}$$

As before, the total indirect effect for a unit change in  $X$  is given by  $\beta_X^* - \beta_X$ . The indirect effect through  $M_1$  is given by  $\alpha_1 \beta_1$ . For  $i = 2, \dots, j$ , the indirect



effect through  $M_i$  only is  $\kappa_i\beta_i$ , and the indirect effect through  $M_1 \rightarrow \dots \rightarrow M_{i-1}$  in serial is  $\alpha_1 \times \delta_{21} \dots \delta_{i-1, i-2} \delta_{i, i-1} \times \beta_i$ . If these relationships are correctly specified, then the total indirect effect can be written as the sum of the serially mediated indirect effects. For the two-mediator example, we have  $\beta_x^* - \beta_x = \alpha_1\beta_1 + \kappa_2\beta_2 + \alpha_1\delta_{21}\beta_2$ . The variances of  $\hat{\alpha}_1\hat{\beta}_1$  and  $\hat{\kappa}_2\hat{\beta}_2$  are estimated using Sobel's formula and  $\hat{\text{Var}}(\hat{\alpha}_1\hat{\delta}_{21}\hat{\beta}_2) = \hat{\alpha}_1^2\hat{\delta}_{21}^2s_{\beta_2}^2 + \hat{\alpha}_1^2\hat{\beta}_2^2s_{\delta_{21}}^2 + \hat{\delta}_{21}^2\hat{\beta}_2^2s_{\alpha_1}^2$  (Hayes, 2013). When the ordering of the mediators is known, VanderWeele and Vansteelandt (2013) provide a PO approach in which indirect effects are estimated sequentially, similar to the serial multiple mediator model. Table 3 shows the regression equations that correspond to the parallel model, serial

model, and the proposed single-model framework for the setting of two mediators  $M_1$  and  $M_2$ . The parallel model assumes the mediators affect the outcome independently, and the serial model assumes that  $M_1$  precedes (and causally affects)  $M_2$ . Both the parallel and serial approaches require fitting several regression equations. In contrast, our proposed single-model approach implicitly incorporates the covariance between the mediators but requires fitting only the equation for the full outcome model. To help the reader gain intuition for how this approach implicitly allows all mediators to covary, we include the implied mediator models (note that the models in gray font do not need to be fit under the single-model approach).