Behavior Research Methods
https://doi.org/10.3758/s13428-023-02095-4

Causal moderated mediation analysis: Methods and software

Xu Qin[1] · Lijuan Wang[2]

Accepted: 16 February 2023

Abstract

Research questions regarding how, for whom, and where a treatment achieves its effect on an outcome have become increasingly valued in substantive research. Such questions can be answered by causal moderated mediation analysis, which assesses the heterogeneity of the mediation mechanism underlying the treatment effect across individual and contextual characteristics. Various moderated mediation analysis methods have been developed under the traditional path analysis/structural equation modeling framework. One challenge is that the definitions of moderated mediation effects depend on statistical models of the mediator and the outcome, and no solutions have been provided when either the mediator or the outcome is binary, or when the mediator or outcome model is nonlinear. In addition, it remains unclear to empirical researchers how to make causal arguments of moderated mediation effects due to a lack of clarifications of the underlying assumptions and methods for assessing the sensitivity to violations of the assumptions. This article overcomes the limitations by developing general definition, identification, estimation, and sensitivity analysis for causal moderated mediation effects under the potential outcomes framework. We also developed a user-friendly R package `moderate.mediation` (https://cran.r-project.org/web/packages/moderate.mediation/index.html) that allows applied researchers to easily implement the proposed methods and visualize the initial analysis results and sensitivity analysis results. We illustrated the application of the proposed methods and the package implementation with a re-analysis of the National Evaluation of Welfare-to-Work Strategies (NEWWS) Riverside data.

Keywords Causal · Moderation · Mediation · Sensitivity analysis · R package implementation

Introduction

In social science research, many questions involve *how*, *for whom*, and *where* a treatment achieves its effect on an outcome. To answer such questions, it is necessary to unpack the underlying mediation mechanism and assess the heterogeneity across individual and contextual characteristics. Such evidence is crucial for advancing in-depth scientific understanding. *Mediation* analysis and *moderation* analysis play essential roles in this line of research (e.g., Baron & Kenny, 1986; Hayes, 2017; Hong, 2015; VanderWeele, 2015).

Mediation analysis answers the question of *how* by uncovering the pathways through which a treatment effect is generated. In the basic mediation framework (e.g., Baron & Kenny, 1986), a treatment affects a focal *mediator*, which in turn affects an outcome. The total treatment effect can be decomposed into an indirect effect transmitted via the mediator and a direct effect that works directly or through other unspecified pathways. Mediation analysis plays a key role in developing causal theories and testing causal hypotheses.

Moderation analysis answers the question of *for whom* or *where* by evaluating the heterogeneity in the direction and/or strength of the treatment effect across different subpopulations or settings defined by a *moderator*. In other words, moderation analysis assesses the treatment effect as a function of the moderator. Moderation analysis is key to understanding the generalizability of causal relations.

A combination of mediation analysis and moderation analysis integrates the investigations of *how*, *for whom*, and *where* and thus enables a unified and deeper understanding of the mediation mechanism and heterogeneity. In particular, *moderated mediation* assesses the heterogeneity of the mechanism, while *mediated moderation* assesses the mechanism of the heterogeneity (e.g., Edwards & Lambert, 2007; Fairchild & MacKinnon, 2009; Hayes, 2015; James & Brett, 1984; Morgan-Lopez & MacKinnon, 2006; Muller et al., 2005; Preacher et al., 2007; Wang &

✉ Xu Qin
xuqin@pitt.edu

[1] Department of Health and Human Development
at the School of Education, University of Pittsburgh,
5312 Wesley W. Posvar Hall, 230 South Bouquet Street,
Pittsburgh, PA 15260, USA

[2] University of Notre Dame, Notre Dame, IN, USA

Published online: 16 October 2023

Preacher, 2015). In other words, moderated mediation answers the question of whether and how the direction and/or strength of the pathways that transmit the treatment effect differ across different individuals and contexts. In the presence of heterogeneity of a treatment effect, mediated moderation uncovers the pathways through which such a heterogeneity is generated by decomposing the moderated treatment effect. This study focuses on moderated mediation.

Methodological development of moderated mediation analysis under the path analysis/ structural equation modeling (SEM) framework has been arising in the past two decades. The basic model for mediation analysis under this framework regresses a continuous mediator M on treatment T (the mediator model) and regresses a continuous outcome Y on T and M (the outcome model):

$$\begin{aligned} M &= \beta_0^m + \beta_t^m T + \varepsilon_M \\ Y &= \beta_0^y + \beta_t^y T + \beta_m^y M + \varepsilon_Y. \end{aligned} \tag{1}$$

The product of β_t^m and β_m^y quantifies the indirect effect of T on Y via M, and β_t^y quantifies the direct effect. To determine if mediation is moderated by a moderator W, one may adopt a multiple-group analysis approach by fitting separate models within each of the subgroups defined by W. However, this procedure sacrifices power, which should be avoided in psychology research that generally has modest levels of power, and it does not test how mediation differs between subgroups (Edwards & Lambert, 2007). Instead, a widely adopted solution is to incorporate W and its interaction with T in the mediator model and/or M in the outcome model. With this interaction modeling approach, a conditional indirect effect is defined as the indirect effect at given values of one or more moderators (Preacher et al., 2007). Various approaches have been developed to determine if the conditional indirect effect significantly varies across different values of the moderator(s), which reflects if mediation is significantly moderated. Some researchers claimed significant moderated mediation based on the significant interaction of W with path $T \rightarrow M \left(\beta_t^m\right)$ or path $M \rightarrow Y \left(\beta_m^y\right)$. However, Fairchild and MacKinnon (2009) and Hayes (2015) showed that the indirect effect could vary significantly by W (i.e., the mediation could be significantly moderated) even if neither β_t^m nor β_m^y significantly varies by W, and a significant interaction of W with β_t^m or β_m^y does not guarantee that the indirect effect significantly varies by W. Hayes (2015) proposed a single test based on an interval estimate of an index of moderated mediation, which quantifies the relationship between the indirect effect and the moderator and varies across different moderated mediation scenarios. Although the test allows for an overall assessment of moderated mediation, it applies only if the indirect effect is a linear function of the moderator. When the indirect effect and the moderator are nonlinearly associated, one can assess moderated mediation by testing if the indirect effects conditional on two given values of the moderator are significantly different (Edwards & Lambert, 2007; Wang & Preacher, 2015). While Edwards and Lambert (2007) focused on a single moderator, Wang and Preacher (2015) considered various scenarios when multiple moderators are involved.

Despite the rich literature on moderated mediation analysis, two major problems remain unsolved under this traditional framework. First, although the ultimate goal of moderated mediation analysis is to reveal causal relations among variables, researchers (Edwards & Lambert, 2007; Fairchild & MacKinnon, 2009; Hayes, 2015) acknowledged that the methods do not offer a formal test of causality. Ignoring confounders of the treatment–mediator, treatment–outcome, or mediator–outcome relationships would generate biased results, and failure to assess the influence of unmeasured confounders may result in misleading conclusions. Second, no general definition of the conditional indirect effect is available beyond specific mediator and outcome models. As enumerated by Preacher et al. (2007) and Wang and Preacher (2015), the conditional indirect effect is defined differently under different scenarios (e.g., only β_t^m or β_m^y is moderated; both β_t^m and β_m^y are moderated; β_t^m is moderated by one moderator while β_m^y is moderated by another).

The limitations can be overcome with the advancement of causal mediation analysis (e.g., Hong, 2010; Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b; VanderWeele & Vansteelandt, 2009, 2010) under the counterfactual causal framework, which is also known as the potential outcomes framework (Neyman & Iwaszkiewicz, 1935; Rubin, 1978). However, there has been sparse research on causal moderated mediation analysis. Extending the SEM framework, Tingley et al. (2014, pp.10–12) very briefly introduced an implementation in the R package `mediation` for the estimation and inference of indirect and direct effects at given values of a moderator and the difference in each effect between two given values of the moderator. Steen et al. (2017, pp. 24-25) provided a short description of an additional R implementation in the package `medflex` by extending an imputation-based causal mediation analysis method. Hong (2015, pp. 325-327) introduced in several paragraphs an extension of a propensity score-based weighting strategy for causal moderated mediation analysis. The imputation-based approach imputes potential outcomes based on an outcome model and observed values of the mediator while not requiring a mediator model (Vansteelandt et al., 2012). In contrast, the weighting-based approach predicts potential outcomes by transforming observed outcomes via weights that are constructed based on propensity scores of the mediator. It relies on a mediator model while not requiring an outcome model (Hong et al., 2015). Hence, both methods are more robust to model misspecifications but less efficient than the extended SEM framework that relies on both a mediator model and an outcome model. None of these studies assessed sensitivity of analysis results to unmeasured confounding.

We have evaluated the performance of `mediation` and `medflex` through simulations (Appendix H). `mediation` cannot handle more than one moderator. When there is only one moderator, the 95% confidence interval coverage rates of the moderated indirect and direct effects were close to 1 under the studied conditions. For `medflex`, the point estimates of the moderated indirect effect under the studied conditions had nonignorable bias (e.g., relative bias could be as high as 50%) when the true effect is nonzero, and the 95% confidence interval coverage rates of the moderated indirect and direct effects were as low as 0.2%. In addition, `medflex` required that a moderator should interact with both the treatment and mediator in affecting the outcome, which can be restrictive. Muthén et al. (2017) incorporated in Mplus the traditional moderated mediation analysis methods and the causal moderated mediation analysis method under the extended SEM framework but did not offer a test of moderated mediation or a sensitivity analysis for the moderated mediation effects.

There has been no study that articulates the definition, identification, estimation, and sensitivity analysis of causal moderated mediation effects for various types of mediator, outcome, and moderator variables. Hence, this article aims to fill this important gap in the literature, with a focus on one single mediator. First, we offered a general definition of moderated mediation effects under the potential outcomes framework, independent of any statistical models. Second, we identified the effects under the assumption of sequential ignorability within subgroups defined by moderator(s). Third, we introduced three estimation and inference methods, including bootstrapping, Monte Carlo, and Bayesian methods, for continuous and binary mediator and outcome and various types of moderators. We also compared their performance through simulations. Fourth, we developed a visual tool that presents how the point estimates and confidence intervals of the conditional mediation effects change with a moderator. Fifth, we developed a sensitivity analysis strategy that allows an evaluation of the extent to which causal inference about moderated mediation would be affected by unmeasured pretreatment confounders (i.e., confounders preceding the treatment). Sixth, we developed a user-friendly R package that enables empirical researchers to implement all the proposed methods. We also illustrated the applications of the proposed methods and implementation of the package by analyzing data from the National Evaluation of Welfare-to-Work Strategies (NEWWS) study.

Application example

To ease the introduction of the proposed methods, we first introduce the NEWWS, conducted before the nationwide welfare reform in the mid-1990s. Participants of the study consist of applicants to the Aid to Families with Dependent Children (AFDC) program and current AFDC recipients who were not working for 30 or more hours per week. They were randomly assigned to the labor force attachment (LFA) program, which aimed at moving low-income parents from welfare to work by providing employment-focused incentives and services, and the control group, which received aid from AFDC without requirement for working.

Although the LFA program was found to increase the likelihood of employment (Michalopoulos et al., 2001), concerns have been raised about its impact on the psychological well-being of the participants, who were mostly low-income single mothers with young children (e.g., Morris, 2008). However, no significant total effect of LFA on maternal depression was detected (Hamilton & Freedman, 2001). To understand why, Hong et al. (2015) studied how employment mediated the impact of LFA on maternal depression. The indirect effect was estimated to be negative, indicating that the LFA-induced increase in employment relieved one's depression. In contrast, the direct effect was positive, indicating that other components of the program (e.g., financial penalties for noncompliance in program activities or failure to secure employment) might stress participants out and thus adversely affect one's mental health. The counteracting indirect and direct effects explained the null total effect. In addition, they examined the treatment-by-mediator interaction and found that employment would be more beneficial to psychological well-being under the LFA condition than under the control condition. In this study, we further assessed if such mediation mechanism is generalizable across different subpopulations, e.g., participants who received different amount of welfare prior to the randomization; or participants who had different numbers of children.

This article uses data from the Riverside sample (a subsample of NEWWS), same as Hong et al. (2015). It includes 694 participants with preschool-age children, 208 randomly assigned to the LFA program and 486 randomly assigned to the control group. Treatment $T=1$ if a participant was assigned to LFA and $T=0$ otherwise. Mediator $M=1$ if one was ever employed during the two-year period after randomization and $M=0$ if not. Outcome Y, maternal depression at the end of the second year after randomization, is a summary score of 12 items measuring depressive symptoms during the past week on a 0–3 scale. A higher score indicates more severe depression. The outcome ranged from 0 to 34, with a mean of 7.50 and a standard deviation of 7.74. Rich baseline covariates were collected shortly before randomization, including depressive symptoms, attitudes toward employment and training, employment status, marital status, race, welfare amount, and number of children, etc. Details can be found in Table 2.

Definitions of causal moderated mediation effects

Using the NEWWS example as an illustration, we offer general definitions of causal moderated mediation effects by extending the definitions of causal mediation effects (Pearl, 2001; Robins & Greenland, 1992) under the potential outcomes framework (Neyman & Iwaszkiewicz, 1935; Rubin, 1978). Individual i has two potential mediators, $M_i(1)$ and $M_i(0)$, which respectively denote one's potential employment experience if assigned to the LFA program and that if assigned to the control group. Only the one under the actual treatment condition is observed. Similarly, individual i has two potential outcomes, $Y_i(t)$, for $t=0$, 1, which represents one's potential depression level under treatment condition t, and $Y_i(t)$ is observed only if the individual was actually assigned to group t. In causal mediation analysis, to reflect the fact that treatment affects the mediator, which subsequently affects the outcome, one may equivalently view $Y_i(t)$ as a function of both the treatment and the potential mediator under the same treatment condition, i.e., $Y_i(t, M_i(t))$. These are defined under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980, 1986, 1990), which assumes that (1) there is only one version of each treatment condition, (2) individuals do not interfere with each other.

Correspondingly, the *total treatment effect* can be defined as a contrast of the two potential outcomes for each individual:

$$\delta_{TEi} = Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(0)\big).$$

The decomposition of the total treatment effect involves two additional potential outcomes that are unobservable, $Y_i(t, M_i(1-t))$, for $t=0$, 1, indicating one's potential depression level if assigned to treatment group t yet counterfactually having the same employment experience as he or she would have had under the alternative treatment condition $1-t$. Therefore, Robins and Greenland (1992) decomposed the total treatment effect into the sum of the *total indirect effect*,

$$\delta_{TIEi} = Y_i\big(1, M_i(1)\big) - Y_i\big(1, M_i(0)\big),$$

representing the effect of LFA on individual i's depression transmitted solely through the LFA-induced change in employment experience, when the treatment is held at the *LFA* condition, and the *pure direct effect*,

$$\delta_{PDEi} = Y_i\big(1, M_i(0)\big) - Y_i\big(0, M_i(0)\big),$$

representing the effect of LFA on individual i's depression if his or her mediator were held constant at the level that would have been realized under the *control* condition. Alternatively, the total treatment effect can be decomposed into the sum of the *pure indirect effect*,

$$\delta_{PIEi} = Y_i\big(0, M_i(1)\big) - Y_i\big(0, M_i(0)\big),$$

representing the effect of LFA on individual i's depression solely attributable to the LFA-induced change in employment experience, when the treatment is held at the *control* condition, and the *total direct effect* ,

$$\delta_{TDEi} = Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(1)\big),$$

representing the effect of LFA on individual i's depression if his or her mediator were held constant at the level that would have been realized under the *LFA* condition.

The total indirect effect and pure direct effect are also known as the natural indirect effect and natural direct effect (Pearl, 2001). *TIE* and *PDE* are used here for a better contrast with *TDE* and *PIE*. δ_{TIEi} may not be equal to δ_{PIEi}, and similarly, δ_{PDEi} may be unequal to δ_{TDEi}. The discrepancy exists if the treatment interacts with the mediator when affecting the outcome,

$$\delta_{INTi} = \delta_{TIEi} - \delta_{PIEi} = \delta_{TDEi} - \delta_{PDEi},$$

which is known as the *natural treatment-by-mediator interaction effect* (Hong et al., 2015) and represents the effect of LFA on individual i's depression transmitted through a change in the relationship between employment experience and depression level.

Hence, in addition to the two-way decomposition of the total treatment effect:

$$\delta_{TEi} = \delta_{TIEi} + \delta_{PDEi} = \delta_{PIEi} + \delta_{TDEi},$$

a three-way decomposition also applies in the presence of the treatment-by-mediator interaction:

$$\delta_{TEi} = \delta_{PIEi} + \delta_{PDEi} + \delta_{INTi}.$$

Above we illustrate the definitions with a binary treatment. If the treatment has more than two categories or is continuous, one may replace $T=1$ and $T=0$ with any two different values of T, t and t' (VanderWeele & Vansteelandt, 2009). A summary of the general definitions of each potential outcome and each causal effect for individual i can be found in Table 1.

By taking an average of each individual-specific effect as defined above over all the individuals, one can define the population average of the effect. Filling an important gap in

Table 1 Definitions of individual-specific potential outcomes and causal effects

	Notation	Definition (for individual i)
	$Y_i(t, M_i(t))$	Potential outcome if $T_i = t$
	$Y_i(t', M_i(t'))$	Potential outcome if $T_i = t'$
	$Y_i(t, M_i(t'))$	Potential outcome if $T_i = t$ yet the mediator takes the value that would result if $T_i = t'$
	$Y_i(t', M_i(t))$	Potential outcome if $T_i = t'$ yet the mediator takes the value that would result if $T_i = t$
Total effect	$\delta_{TEi} = Y_i(t, M_i(t)) - Y_i(t', M_i(t'))$	The total treatment effect on the outcome
Total indirect effect (natural indirect effect)	$\delta_{TIEi} = Y_i(t, M_i(t)) - Y_i(t, M_i(t'))$	The treatment effect on the outcome transmitted solely through the treatment-induced change in the mediator, while the treatment status is held at t
Pure direct effect (natural direct effect)	$\delta_{PDEi} = Y_i(t, M_i(t')) - Y_i(t', M_i(t'))$	The treatment effect on the outcome if the mediator is held at the level that would be realized under $T_i = t'$
Pure indirect effect	$\delta_{PIEi} = Y_i(t', M_i(t)) - Y_i(t', M_i(t'))$	The treatment effect on the outcome transmitted solely through the treatment-induced change in the mediator, while the treatment status is held at t'
Total direct effect	$\delta_{TDEi} = Y_i(t, M_i(t)) - Y_i(t', M_i(t))$	The treatment effect on the outcome if the mediator is held at the level that would be realized under $T_i = t$
Natural treatment-by-mediator interaction effect	$\delta_{INTi} = \delta_{TIEi} - \delta_{PIEi} = \delta_{TDEi} - \delta_{PDEi}$	The difference in how the treatment-induced change in the mediator affects the outcome between the treatment conditions t and t'

the literature, we formalize the definitions of the moderated mediation effects under the potential outcomes framework. Let $\boldsymbol{W}$ denote a vector of moderators, which define subpopulations and/or settings across which the mediation mechanism may differ. Same as Muller et al. (2005) and Hong (2015), we assume that a moderator is *pretreatment* in nature.[1] In other words, a moderator measures a stable individual or contextual difference prior to the treatment. A key question is how to choose moderators. The commonly recommended approach is built upon substantive knowledge about which pretreatment covariates are most likely to show evidence for heterogeneity (VanderWeele, 2015) of the mediation mechanism. In the NEWWS example, we hypothesize that $\boldsymbol{W}$ consists of amount of welfare and number of children at baseline.

Averaging each individual-specific effect over individuals within given levels of $\boldsymbol{W}$, we define the conditional average of each component of the total treatment effect as

$$\delta_{\boldsymbol{w}} = E\left[\delta_i \mid \boldsymbol{W}_i = \boldsymbol{w}\right], \tag{2}$$

where δ stands for δ_{TIE}, δ_{PDE}, δ_{PIE}, δ_{TDE}, and δ_{INT}. This definition reflects the essence of moderated mediation, i.e., mediation mechanism varies as a function of moderators. Correspondingly, each moderated effect can be defined as a contrast of the conditional effect between subpopulations or settings defined by two different levels of $\boldsymbol{W}$:

$$\delta_{MOD} = \delta_{\boldsymbol{w}_1} - \delta_{\boldsymbol{w}_2}. \tag{3}$$

These definitions intuitively formalize mediation and moderation without relying on any statistical models and apply to all the possible moderated mediation scenarios.

Identification of causal moderated mediation effects

As elaborated above, $Y_i(t, M_i(t))$ (or $Y_i(t', M_i(t'))$) is observed only if individual i received treatment level t (or t′), while $Y_i(t, M_i(t'))$ (or $Y_i(t', M_i(t))$) is never observable if $t \neq t'$. The key to identifying the population average of the causal mediation effects is to relate the counterfactual quantities to the observed data, which relies on the sequential ignorability assumption (e.g., Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b; Ten Have et al., 2004). To identify the conditional and moderated mediation effects, we assume that sequential ignorability holds within levels of moderators:

Assumption 1 (Conditional ignorability of treatment) Within levels of moderators $\boldsymbol{W}$ that are of theoretical interest, the treatment is independent of all the potential mediators and potential outcomes for those sharing the same observed pretreatment covariates $\boldsymbol{X}$,[2]

$$\left\{Y_i\left(t', m\right), M_i(t)\right\} \perp\!\!\!\perp T_i \mid \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x},$$

[1] If a variable that moderates the mediation mechanism is posttreatment (i.e., affected by the treatment), it is essentially an additional mediator that interacts with the treatment and/or the focal mediator.

[2] $Y_t(t', m)$ stands for the potential outcome if the treatment condition is at t' and the potential mediator takes the value of m. A proof of why Assumptions 1 and 2 are needed for the identification of the conditional and moderated mediation effects can be found in Appendix A.

where $0 < \Pr(T_i = t \mid \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}) < 1$. Here and throughout the rest of the paper, t and t' can be either equal or unequal, so that the following discussions on the identification and estimation apply to all the four potential outcomes used to define the effects in Table 1. This assumption implies no omitted confounders of the treatment-mediator or treatment-outcome relationship within levels of moderators. It is guaranteed by design in a randomized experiment such as NEWWS.

Assumption 2 (Conditional ignorability of mediator) Within levels of moderators $\boldsymbol{W}$ that are of theoretical interest, the potential mediator under either treatment condition is independent of the potential outcomes for those sharing the same observed pretreatment covariates $\boldsymbol{X}$,

$$Y_i\left(t', m\right) \perp\!\!\!\perp M_i(t) \mid T_i = t, \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x},$$

where $0 < \Pr(M_i(t) = m \mid T_i = t, \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}) < 1$. This assumption implies no omitted pretreatment confounders (i.e., confounders preceding the treatment) and no posttreatment confounders (i.e., confounders affected by the treatment) of the mediator-outcome relationship within a treatment condition or across treatment conditions, given levels of moderators. It may not hold even in a randomized experiment because mediator values are typically generated through a natural process rather than being experimentally manipulated. For example, among the participants with the same welfare amount and number of children prior to the randomization, those who had a higher salary expectation before participating in the study might be more likely to be employed because they were more motivated to secure employment in the job market, but they might suffer from more severe depression due to their higher work pressure. Failures to account for such a pretreatment confounder of the mediator-outcome relationship within levels of moderators would bias the conditional and moderated mediation effects. Although it is almost impossible to account for all the confounders, one can conduct sensitivity analysis to assess if the analytic results are robust to omitted confounders, as explained in the sensitivity analysis section.

Under the above conditional sequential ignorability assumption, the conditional average of the potential outcomes given values of moderators can be identified as

$$\begin{aligned} &E\left[Y_i\left(t, M_i\left(t'\right)\right) \mid \boldsymbol{W}_i = \boldsymbol{w}\right] \\ &= \iint E\left[Y_i \mid \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}, T_i = t, M_i = m\right] dF_{M_i \mid \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}, T_i = t'}(m)\, dF_{\mathbf{X}_i \mid \boldsymbol{W}_i = \boldsymbol{w}}(\mathbf{x}), \end{aligned} \tag{4}$$

where $F_{M_i \mid \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}, T_i = t'}(m)$ represents the conditional distribution of M given $\boldsymbol{W}$, $\mathbf{X}$, and T, and $F_{\mathbf{X}_i \mid \boldsymbol{W}_i = \boldsymbol{w}}(\mathbf{x})$ represents the conditional distribution of $\mathbf{X}$ given $\boldsymbol{W}$. The proof can be found in Appendix A. Correspondingly, we are able to identify the conditional and moderated mediation effects as defined in Eqs. (2) and (3).

In summary, the identification of the moderated mediation effects relies on Assumptions 1 and 2 that, within levels of moderators, there are no omitted pretreatment confounders of the treatment–mediator, treatment–outcome, and mediator–outcome relationships, and there are no posttreatment confounders of the mediator–outcome relationship. Loeys et al. (2016) showed that the index of moderated mediation proposed by Hayes (2015) can be estimated without bias in the presence of omitted pretreatment confounders under some restricted conditions. However, the Hayes index is constructed upon specific linear regressions of the mediator and outcome and applies when the indirect effect is a linear function of the moderator, as discussed in the introduction section.

In the following section, we will discuss the estimation of the effects based on Assumptions 1 and 2.

Estimation

The above identification result allows one to develop a general estimation procedure for conditional and moderated mediation effects using any mediator and outcome models, including linear, nonlinear, and nonparametric models, etc. We illustrate the estimation procedure with linear models for a continuous mediator and a continuous outcome. Extensions to the cases when the mediator and/or outcome are binary can be found in Appendix B.

To assess the population average causal mediation mechanisms, researchers (e.g., Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b; VanderWeele & Vansteelandt, 2009) have extended the mediator and outcome models under the traditional SEM framework in Eq. (1) to account for the treatment-by-mediator interaction and observed pretreatment confounders:[3]

$$\begin{aligned} M &= \beta_0^m + \beta_t^m T + \mathbf{X}\boldsymbol{\beta}_x^m + \varepsilon_m, \\ Y &= \beta_0^y + \beta_t^y T + \beta_m^y M + \beta_{tm}^y TM + \mathbf{X}\boldsymbol{\beta}_x^y + \varepsilon_y, \end{aligned} \tag{5}$$

where β_t^m, $\boldsymbol{\beta}_x^m$, β_t^y, β_m^y, β_{tm}^y, and $\boldsymbol{\beta}_x^y$ respectively represent the paths $T \rightarrow M$, $X \rightarrow M$, $T \rightarrow Y$, $M \rightarrow Y$, $TM \rightarrow Y$ (i.e., moderation of $M \rightarrow Y$ by T), and $X \rightarrow Y$. Under the sequential ignorability assumption and based on Eq. (5), we can derive the estimands as

$$\begin{aligned} \delta_{TIE} &= \left(\beta_m^y + \beta_{tm}^y t\right)\beta_t^m\left(t - t'\right), \\ \delta_{PDE} &= \left(\beta_t^y + \beta_{tm}^y\left(\beta_0^m + \beta_t^m t' + E[\mathbf{X}]\boldsymbol{\beta}_x^m\right)\right)\left(t - t'\right), \\ \delta_{PIE} &= \left(\beta_m^y + \beta_{tm}^y t'\right)\beta_t^m\left(t - t'\right), \\ \delta_{TDE} &= \left(\beta_t^y + \beta_{tm}^y\left(\beta_0^m + \beta_t^m t + E[\mathbf{X}]\boldsymbol{\beta}_x^m\right)\right)\left(t - t'\right), \\ \delta_{INT} &= \beta_t^m \beta_{tm}^y\left(t - t'\right)^2, \end{aligned} \tag{6}$$

where $\boldsymbol{E}[\boldsymbol{X}] = \mathbf{0}$ if $\boldsymbol{X}$ is mean centered. A proof can be found in the appendix of VanderWeele and Vansteelandt (2009).

[3] For the notations, the subscript indicates the predictor that the coefficient corresponds to, and the superscript indicates the model that the predictor belongs to. This can avoid misreading especially when there are a lot of predictors and/or moderators in the model. Here β_0^m, β_t^m, $\boldsymbol{\beta}_x^m$, β_0^y, β_t^y, β_m^y, β_{tm}^y, and $\boldsymbol{\beta}_x^y$ respectively correspond to β_0, β_1, $\boldsymbol{\beta}_2$, θ_0, θ_1, θ_2, θ_3, and $\boldsymbol{\theta}_4$ in VanderWeele (2015) and α_2, β_2, $\boldsymbol{\xi}_2$, α_3, β_3, γ, κ, and $\boldsymbol{\xi}_3$ in Imai et al. (2010a).

When the mediation mechanism differs across different subpopulations or settings, we need to incorporate pretreatment moderators of the mediation mechanism in the mediator and outcome models. Model specification and interpretation would become increasingly challenging as more moderated relationships are considered. Such concerns can be greatly alleviated by specifying the models in a hierarchical form (Wang & Preacher, 2015). Based on the main models in Eq. (5), we can further express each coefficient as a function of moderator W,

$$\begin{aligned}
\beta_0^m &= \beta_{00}^m + \beta_{0w}^m W \\
\beta_t^m &= \beta_{t0}^m + \beta_{tw}^m W \\
\boldsymbol{\beta}_x^m &= \boldsymbol{\beta}_{x0}^m + \boldsymbol{\beta}_{xw}^m W \\
\beta_0^y &= \beta_{00}^y + \beta_{0w}^y W \\
\beta_t^y &= \beta_{t0}^y + \beta_{tw}^y W \\
\beta_m^y &= \beta_{m0}^y + \beta_{mw}^y W \\
\beta_{tm}^y &= \beta_{tm0}^y + \beta_{tmw}^y W \\
\boldsymbol{\beta}_x^y &= \boldsymbol{\beta}_{x0}^y + \boldsymbol{\beta}_{xw}^y W,
\end{aligned} \tag{7}$$

where β_{0w}^m, β_{t0}^m, and $\boldsymbol{\beta}_{x0}^m$ respectively represent the main effects of W, T, and $\boldsymbol{X}$ on the mediator; β_{tw}^m equals the average change in β_t^m with one unit increase of W and thus reflects the extent to which W moderates the path T➔M; and each element of $\boldsymbol{\beta}_{xw}^m$ indicates how W moderates each of the paths $\boldsymbol{X}$➔M. All the coefficients in the outcome model can be interpreted in a similar way. For illustration purposes, Eq. (7) focuses on one single moderator and includes the moderator in the models of all the main model coefficients. We could decide whether to remove the moderator or add more moderators in each model based on theories or empirical findings of how each path is moderated. Nevertheless, we must consider the main effects of all the moderators by incorporating all the moderators in the mediator (outcome) model into the model of β_0^m (β_0^y).

Combing Eqs. (5) and (7) yields

$$\begin{aligned}
M &= \beta_{00}^m + \beta_{0w}^m W + \beta_{t0}^m T + \boldsymbol{X}\boldsymbol{\beta}_{x0}^m + \beta_{tw}^m WT + W\boldsymbol{X}\boldsymbol{\beta}_{xw}^m + \varepsilon_M \\
Y &= \beta_{00}^y + \beta_{0w}^y W + \beta_{t0}^y T + \beta_{m0}^y M + \boldsymbol{\beta}_{x0}^y \boldsymbol{X} + \beta_{tw}^y WT + \beta_{mw}^y WM \\
&\quad + W\boldsymbol{X}\boldsymbol{\beta}_{xw}^y + \beta_{tm0}^y TM + \beta_{tmw}^y WTM + \varepsilon_Y
\end{aligned} \tag{8}$$

Compared to the combined form in Eq. (8), the hierarchical form in Eqs. (5) and (7) facilitates the specification and interpretation of moderation. The above ways of model specification can cover various scenarios, including but not limited to those discussed in Wang and Preacher (2015) and Preacher et al. (2007). In particular, it can account for possible heterogeneity of treatment-by-mediator interaction across subpopulations or settings, which has not been discussed in the existing literature, to the best of our knowledge.

Based on the identification result in Eq. (4) and the mediator and outcome models that incorporate the moderator W in Eqs. (5) and (7), which is equivalent to the combined form in Eq. (8), we can obtain the estimands of the conditional effects at $W = w$ as

$$\begin{aligned}
\delta_{TIE,w} &= \left(\beta_{m|w}^y + \beta_{tm|w}^y t\right)\beta_{t|w}^m (t - t'), \\
\delta_{PDE,w} &= \left(\beta_{t|w}^y + \beta_{tm|w}^y (\beta_{0|w}^m + \beta_{t|w}^m t' + E[\mathbf{X}|W = w]\boldsymbol{\beta}_{x|w}^m)\right)(t - t'), \\
\delta_{PIE,w} &= \left(\beta_{m|w}^y + \beta_{tm|w}^y t'\right)\beta_{t|w}^m (t - t'), \\
\delta_{TDE,w} &= \left(\beta_{t|w}^y + \beta_{tm|w}^y (\beta_{0|w}^m + \beta_{t|w}^m t + E[\mathbf{X}|W = w]\boldsymbol{\beta}_{x|w}^m)\right)(t - t'), \\
\delta_{INT,w} &= \beta_{t|w}^m \ \beta_{tm|w}^y (t - t')^2,
\end{aligned} \tag{9}$$

where each β takes the value at $W = w$, e.g., $\beta_{m|w}^y = \beta_{m0}^y + \beta_{mw}^y w$. Eq. (9) is essentially a straightforward extension of the population average effect estimands in Eq. (6) through a further conditioning on W. It reveals that TIE, PIE, and INT vary with only W, while the PDE and TDE vary with not only W but also $\mathbf{X}$. The variation of PDE and TDE by $\mathbf{X}$ is purely due to the treatment-by-mediator interaction. Within a given level of W, we can directly estimate $\delta_{TIE,w}$, $\delta_{PIE,w}$ and $\delta_{INT,w}$ based on point estimates of model coefficients. It is similar for PDE and TDE if the treatment and the mediator do not interact in affecting the outcome (i.e., $\beta_{tm}^y = 0$), or if there is no $\mathbf{X}$. Otherwise, we need to account for $E[\mathbf{X}|W = w]$ by predicting the expectation of $\mathbf{X}$ as a function of W, or directly predicting $\delta_{PDE,w}$ and $\delta_{TDE,w}$ as functions of W.

The above estimands are illustrated with a single moderator that moderates all the coefficients in the main models of continuous mediator and outcome. Below we propose a more general estimation procedure that can incorporate multiple moderators and be easily extended to binary mediator and/or outcome and to various causal mediation analysis methods as explained in the discussion section. In the presence of treatment-by-mediator interaction or nonlinearities, some effects vary with not only the hypothesized moderators of the mediation mechanisms, $\boldsymbol{W}$, but also pretreatment confounders within levels of the moderators, $\mathbf{X}$. Hence, the basic idea of the general estimation procedure is to first predict each effect as a function of $\boldsymbol{W}$ and $\mathbf{X}$ based on the mediator and outcome models and then fit a multivariate adaptive regression spline (Friedman, 1991) of the effect estimates on $\boldsymbol{W}$. The two-step procedure shares the essence with the existing methods for investigating the total treatment effect heterogeneity (Carvalho et al., 2019). A multivariate adaptive regression spline is preferred to a parametric regression in the second step because it can flexibly capture nonlinear relationships between the effects and moderators. In some cases, as illustrated above and explicated in Step 2.5 below, there is no need to fit a spline in the second step because the first step itself can parametrically determine how the effects vary with $\boldsymbol{W}$ without conditioning on $\mathbf{X}$.

Below we consider applications of various commonly used estimation and inference methods, including bootstrapping, Monte Carlo, and Bayesian method, for estimating and making inference about conditional and moderated mediation effects.

Bootstrapping method

To make statistical inference of mediation analysis or moderated mediation analysis results, researchers derived the standard errors of the effect estimates using delta method and assumed the sampling distributions of the effect estimates to be approximately normal. However, a growing literature under the traditional framework for mediation analysis and moderated mediation analysis has been advocating the use of bootstrapping for inference because the sampling distribution of the indirect effect estimate or conditional indirect effect estimate, which involves products of regression coefficient estimates, may not be normal (e.g., Preacher et al., 2007; Preacher & Hayes, 2004). The same applies to the counterfactual causal framework.

Following the identification result, we propose an algorithm based on bootstrapping for the estimation and inference of conditional and moderated mediation effects.[4]

Algorithm 1 (Bootstrapping method)

Step 1. Generate Q (e.g., 1,000) bootstrapping samples by resampling raw data with replacement, for each of which repeat Step 2.

Step 2.1. Fit mediator and outcome models (e.g., Eqs. (5) and (7), the combination of which is equivalent to Eq. (8)).

Step 2.2. Predict $M(t^{'})$ given $T_i = t^{'}$ and the observed values of $\boldsymbol{W}_i$ and $\mathbf{X}_i$ for each individual i.

Step 2.3. Predict $Y_i(t, M_i(t^{'}))$ given $T_i = t, M_i = M_i(t^{'})$, and the observed values of $\boldsymbol{W}_i$ and $\mathbf{X}_i$ for each individual i.

Step 2.4. Predict δ_i as a function of $\boldsymbol{W}_i$ (and $\mathbf{X}_i$) by taking a contrast of the corresponding predicted potential outcomes for each individual i.

Step 2.5. Estimate the conditional effect at a given level of $\boldsymbol{W}$, $\delta_{\boldsymbol{w}} = E[\delta_i | \boldsymbol{W}_i = \boldsymbol{w}]$.

If δ_i is predicted as a function of only $\boldsymbol{W}_i$ in Step 2.4, we can estimate $\delta_{\boldsymbol{w}}$ by setting $\boldsymbol{W}$ at the given level $\boldsymbol{w}$ for each individual in Steps 2.2 and 2.3. This is the case when (a) there is no $\mathbf{X}$; when (b) the mediator and outcome are continuous, parameteric models of the mediator and outcome are fitted in Step 2.1, and one of the following conditions is met: (b.1) $\delta_{\boldsymbol{w}}$ stands for conditional TIE, PIE, or INT, (b.2) $\delta_{\boldsymbol{w}}$ stands for conditional TDE or PDE, and there is no treatment-by-mediator interaction; or when (c) the mediator is binary, the outcome is continuous, parameteric models of the mediator and outcome are fitted in Step 2.1, $\delta_{\boldsymbol{w}}$ stands for conditional TDE or PDE, and there is no treatment-by-mediator interaction.

If δ_i is predicted as a function of not only $\boldsymbol{W}_i$ but also $\mathbf{X}_i$, which is due to treatment-by-mediator interaction or nonlinearities, estimate $\delta_{\boldsymbol{w}}$ by fitting a multivariate adaptive regression spline of δ_i estimated in Step 2.4 on $\boldsymbol{W}_i$.

Sometimes researchers may prefer to assess how δ_i varies with a subset of $\boldsymbol{W}_i$, $\boldsymbol{W}_{si}$, such as how the pure indirect effect varies with the number of children without conditioning on the amount of welfare. In this case, we can estimate $\delta_{\boldsymbol{w}_s}$ by fitting a multivariate adaptive regression spline of estimated δ_i on $\boldsymbol{W}_{si}$.

The uncertainty of estimated δ_i is taken into account by the resampling procedure as described in Step 1.

Step 2.6. Estimate each moderated effect δ_{MOD} by taking a contrast of the conditional effect between two different levels of $\boldsymbol{W}$.

Step 3. Obtain the final point estimates of the conditional and moderated effects by applying Step 2 to the original sample. Compute standard errors and confidence intervals respectively based on standard deviations and percentiles of the estimates obtained from the Q bootstrapped samples.

Monte Carlo method

The Monte Carlo method (also known as the Monte Carlo confidence interval method) is an alternative method that does not assume normality for the causal effect estimates that involve products of regression coefficient estimates. It was proposed by King et al. (2000), which has been extended to traditional mediation analysis in the single-level setting (MacKinnon et al., 2004; Preacher & Selig, 2012) and multilevel setting (Bauer et al., 2006) as well as causal mediation analysis (the quasi-Bayesian Monte Carlo method in Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b; implementation in R package `mediation` by Tingley et al., 2014). The idea is to simulate the sampling distribution of each effect estimate based on random draws of regression coefficient estimates from their multivariate sampling distributions.

Following the identification result for the conditional and moderated mediation effects, we propose an alternative algorithm based on the Monte Carlo method.

Algorithm 2 (Monte Carlo Method)

Step 1. Fit mediator and outcome models.

Step 2. Simulate Q (e.g., 1,000) random draws of model parameter estimates from their multivariate normal sampling distribution[4]. Based on each draw of model parameter estimates, follow Steps 2.2 to 2.6 in Algorithm 1.

Step 3. Compute the final point estimates, their standard errors, and confidence intervals for the conditional

[4] This is because the sampling distribution of the regression coefficient estimates is asymptotically multivariate normal (King et al., 2000).

and moderated effects respectively based on the means[5], standard deviations, and percentiles of the Q sets of point estimates.

The above two algorithms share the essence with the existing regression-based causal mediation analysis methods that focus on population average effects. If Steps 2.5 and 2.6 are changed to calculating the mean of δ_i over all the individuals in the sample, Algorithm 1 is equivalent to the nonparametric inference algorithm proposed by Imai et al. (2010a) and provides the same point estimates of the population average total (natural) indirect and pure (natural) direct effects as those derived from the same mediator and outcome models by VanderWeele and colleagues (Valeri & VanderWeele, 2013; VanderWeele & Vansteelands, 2009, 2010). Algorithm 2 is equivalent to the parametric inference algorithm in Imai et al. (2010a).

Bayesian method

Bayesian procedures have been proposed for estimating and testing mediation effects (Miočević et al., 2018; Yuan & MacKinnon, 2009) and conditional and moderated mediation effects (Wang & Preacher, 2015) under the traditional framework. The Bayesian approaches have several advantages. First, they allow an analyst to draw inference without assuming the posterior distributions of the conditional or moderated mediation effects to be normal. Second, by incorporating prior information into the analysis, the Bayesian approaches improve estimation efficiency if useful prior information is available. Third, Bayesian credible intervals support direct probabilistic statements about parameters because parameters are treated as random from the Bayesian perspective.

Following the identification result for the conditional and moderated mediation effects, we propose an alternative algorithm based on the Bayesian method.

Algorithm 3 (Bayesian Method)

Step 1. Specify mediator and outcome models for the likelihood function and specify prior distributions for the model parameters.

Step 2. Obtain the posterior distributions of the conditional and moderated mediation effects through Gibbs sampling. Specifically, (1) assign initial values to the model parameters; (2) for the mediator and outcome models separately, sample each model parameter from its conditional distribution given the current values of the other model parameters and the data; (3) based on the sampled model parameters, follow Steps 2.2 to 2.6 in Algorithm 1. Repeat (2) and (3) for a number of iterations (e.g., 20,000). To check whether the sampled values converge to a stationary posterior distribution, one may assess convergence via a trace plot, which represents the sequence of sample values across iterations, in combination with the Gelman–Rubin potential scale reduction (PSR) statistic (Gelman & Rubin, 1992) or the Geweke statistic (Geweke, 1992). Convergence is reached if the trace plot appears stationary, and the Gelman-Rubin statistic is below 1.05, or the Geweke statistic is between -1.96 and 1.96. To make inference, one needs to discard a number of samples (e.g., 10,000) at the beginning of the iterations (i.e., burn-in period) and use the remaining Q (e.g., 10,000) samples at convergence. The more complex the model and data are, the longer the burn-in period is.

Step 3. Compute the final point estimates, their standard errors, and credible intervals for the conditional and moderated effects respectively based on the means[6], standard deviations, and percentiles of the post-burn-in samples.

In summary, all the three methods share the same advantage that they do not rely on the normality assumption for any conditional or moderated mediation effect, and that they well account for the uncertainty of the estimated δ_i in the final estimation of δ_w. We further conducted Monte Carlo simulations to compare the performance of the three algorithms in terms of empirical bias, mean squared error (MSE), confidence/credible interval (CI) coverage rate, power, and type I error rate for the estimation and inference of conditional and moderated effects under various scenarios. Details can be found in Appendix C. When both the mediator and the outcome are continuous, all the three algorithms can estimate the conditional and moderated effects with ignorable bias, while the Monte Carlo method achieves the highest power for all the effects at a sample size ($n = 50$). However, when the mediator or the outcome is binary, the Monte Carlo method generated the largest bias, especially when the sample size is small. This is because the Monte Carlo method relies on the asymptotic multivariate normal assumption for the sampling distribution of the regression coefficient estimates, which can be violated at a small sample size.

[5] We compute the final point estimates based on the mean rather than the median of the Q sets of point estimates as Imai et al. (2010a) did. This is because we found through simulations that the mean estimators are mostly less biased than the median estimators across different scenarios, especially when the sample size is relatively small. Simulations were conducted under the settings described in Appendix C. Simulation results are available upon request.

[6] We compute the final point estimates based on the mean rather than the median of the post-burn-in samples. This is because, similar to what Wang and Preacher (2015) reported, we found through simulations that the mean estimators are mostly less biased than the median estimators across different scenarios, especially when the sample size is relatively small. Simulations were conducted under the settings described in Appendix C. Simulation results are available upon request.

Sensitivity analysis

The conditional and moderated mediation effect estimates would be biased if the conditional sequential ignorability assumption (i.e., Assumptions 1 and 2) were violated. Hence, it is essential to conduct sensitivity analysis to assess the extent to which causal conclusions would be invalidated by potential violations of the identification assumptions. As explicated in the identification section, the conditional sequential ignorability assumption would be violated in the presence of posttreatment confounders or omitted pretreatment confounders of the mediator-outcome relationship within levels of moderators. This section focuses on assessing the influence of omitted pretreatment confounders, while a discussion of posttreatment confounding can be found in the discussion section.

There are two types of omissions of pretreatment confounding. Some pretreatment confounders are observed but omitted to avoid model overfitting, while others are unmeasured. To evaluate the influence of the former, one may simply compare the results before and after including the omitted pretreatment confounders in the analysis. To assess the influence of the latter, we extended a simulation-based sensitivity analysis strategy for causal mediation analysis (Qin & Yang, 2022). The idea is to simulate an unmeasured pretreatment confounder from its conditional distribution at a given strength and compare the results before and after adjusting for it in the analysis. The strategy allows one to evaluate how unmeasured confounding affects both the point estimates of causal effects and their standard errors. The original analysis results would be sensitive if the signs or significance of the effects can be altered by a slight violation of the identification assumption, i.e., by an omitted confounder that is merely weakly associated with the treatment, mediator, and outcome (e.g., Hong et al., 2018; Imai, Keele, & Tingley, 2010a; Imai, Keele, & Yamamoto, 2010b). The degree to which the assumption is violated, i.e., the strength of unmeasured confounding, can be gauged by prior knowledge, theoretical reasoning, or the observed pretreatment confounders in the data, as illustrated in the section of application with implementation in R.

Assume that the treatment is randomized and the conditional ignorability of the mediator holds given $\mathbf{X}$ and U,

$$Y_i\left(t', m\right) \perp\!\!\!\perp M_i(t) \mid T_i = t, \boldsymbol{W}_i = \boldsymbol{w}, \mathbf{X}_i = \mathbf{x}, U_i = u,$$

where U is independent of $\mathbf{X}$ and $\boldsymbol{W}$ and thus represents the unique part of unmeasured pretreatment confounding of the mediator-outcome relationship given $\boldsymbol{W}$ that remains unexplained by $\mathbf{X}$. Assume that the true main mediator and outcome models that adjust for U are

$$M = \beta_0^m + \beta_t^m T + \mathbf{X}\boldsymbol{\beta}_x^m + \beta_u^m U + \varepsilon_m, \varepsilon_m \sim N\left(0, \sigma_m^2\right),$$
$$Y = \beta_0^y + \beta_t^y T + \beta_m^y M + \beta_{tm}^y TM + \mathbf{X}\boldsymbol{\beta}_x^y + \beta_u^y U + \varepsilon_y, \varepsilon_y \sim N\left(0, \sigma_y^2\right), \tag{10}$$

where the intercepts and the coefficients of the observed variables are each a function of moderator(s) $\boldsymbol{W}$. β_u^m (representing the $U-M$ relationship conditional on T, $\mathbf{X}$, and $\boldsymbol{W}$) and β_u^y (representing the $U-Y$ relationship conditional on T, M, $\mathbf{X}$, and $\boldsymbol{W}$) reflect the confounding role of U, or in other words, the degree to which the identification assumption is violated. By standardizing the mediator and outcome models, we use the standardized β_u^y and β_u^m as *sensitivity parameters* to measure the extent to which U confounds the $M-Y$ relationship, independent of measurement scales. Should the confounding role of U vary by subpopulations or settings, each of β_u^m and β_u^y needs to be expressed as a function of the moderators, as in Eq. (7). This would introduce more sensitivity parameters. For simplicity, we assume that β_u^m and β_u^y stay the same across all the individuals, while this assumption can be relaxed. Based on the pre-specified sensitivity parameter values, the assumption of the marginal distribution of U, and Eq. (10), one can derive the conditional distribution of U given Y, M, T, $\mathbf{X}$, and $\boldsymbol{W}$ and simulate U through a stochastic expectation–maximization (EM) algorithm (Qin & Yang, 2022). More details can be found in Appendix D.

A sensitivity analysis aims to assess how strong unmeasured confounding needs to be for the original findings to be altered. To reach this goal, we estimate and test the conditional and moderated mediation effects with adjustment for simulated U at various strengths.

Step 1. Specify a plausible value range for each sensitivity parameter, i.e., standardized β_u^y and β_u^m, and divide them into a grid. Conduct the following steps for each cell of the grid.

Step 2.1. Simulate U from its conditional distribution through a stochastic EM algorithm based on S (e.g., 10) iterations.

Step 2.2. Follow Steps 1 and 2 of the chosen estimation method[7] to estimate the conditional and moderated mediation effects by adjusting for U in the mediator and outcome models given the sensitivity parameter values. This step generates Q (e.g., 1000) sets of adjusted point estimates of the effects.

Step 2.3. Repeat Steps 2.1 and 2.2 K (e.g., 5) times to account for the uncertainty of U.

Step 3. Compute the final adjusted point estimates and confidence intervals based on the mean and percentiles of the $Q \times K$ sets of the adjusted point estimates.

The idea of the above strategy is similar to multiple imputation with bootstrap (Schomaker & Heumann, 2018). The

[7] Because U is generated from the frequentist perspective, this sensitivity analysis algorithm applies to the bootstrapping method and Monte Carlo method but does not directly apply to the Bayesian method. To implement the sensitivity analysis with the Bayesian method, U should be generated from the Bayesian perspective in Step 2.1.

construction of confidence intervals accounts for both the between- and within-imputation uncertainties.

Above we illustrate the simulation-based sensitivity analysis strategies for a continuous mediator and a continuous outcome when the treatment is randomized. It can be extended to a binary mediator and/or outcome, in which case β_u^m (β_u^y) becomes the coefficient of U in the probit regression of M and Y. It can also be extended to observational studies by introducing an additional sensitivity parameter that captures the $U-T$ relationship conditional on $\mathbf{X}$ and $\boldsymbol{W}$, as explicated in Appendix D. In the section of application with implementation in R, we further introduce how to visualize the sensitivity analysis results for researchers to determine whether the initial analysis results are sensitive.

Comparison of traditional and causal moderated mediation analysis approaches

In the traditional moderated mediation analysis, as reviewed in the introduction section, one has to redefine the conditional and moderated mediation effects as the mediator and outcome models change. The definitions of the effects become more complex as more moderated relationships are involved, making the estimation and inference of the effects more daunting. In contrast, under the potential outcomes framework, no matter how mediator and outcome models are specified, the definitions in Eqs. (2) and (3) and identification in Eq. (4) always apply and thus enable unified estimands that are applicable to various scenarios.

Another major advantage of causal moderated mediation analysis is that it allows for the definition and estimation of conditional indirect and direct effects in the presence of treatment-by-mediator interaction and nonlinearities. When both the mediator and the outcome are continuous and there is no treatment-by-mediator interaction in the outcome model, the estimands of the conditional indirect and direct effects under the causal framework, as illustrated in Eq. (9), are equivalent to those under the traditional framework. When the treatment interacts with the mediator in affecting the outcome, Preacher et al. (2007) defined the indirect effect, but the direct effect remains undefined under the traditional framework. When either the mediator or the outcome is binary, or when the mediator or outcome model is nonlinear, moderated mediation analysis has not been developed under the traditional framework, to the best of our knowledge.

In addition, unlike the traditional moderated mediation analysis, causal moderated mediation analysis clarifies the reliance on the assumption of no unmeasured confounding for causal interpretations and emphasizes the importance of controlling for confounders in the analysis. A sensitivity analysis, as an essential part of causal moderated mediation analysis, facilitates the assessment of robustness to unmeasured confounding.

Application with implementation in R

We developed a user-friendly R package `moderate.mediation` (https://cran.r-project.org/web/packages/moderate.mediation/index.html) for researchers to estimate and test the conditional and moderated mediation effects, assess their sensitivity to unmeasured pretreatment confounding, and visualize the results. We built the package based on the Monte Carlo method (default) and the bootstrapping method. When the sample size is small (e.g., n = 50 or 100 under the studied conditions in the simulations), and the mediator or the outcome is binary, we recommend the bootstrapping method, because it generates point estimates with smaller bias than the Monte Carlo method, as shown in Appendix C. Otherwise, we recommend the Monte Carlo method, because its running speed is much faster and its statistical power is slightly higher at a small sample size when both the mediator and outcome are continuous. Because a relatively large sample size is usually required to detect a significant moderated mediation effect, we set the Monte Carlo method to be the default method. The package is applicable to a binary or continuous treatment, a binary or continuous mediator, a binary or continuous outcome, and one or more moderators of any scale. We illustrate its implementation with an application to the NEWWS example. The package can be loaded through the following syntax,

```
R> library("moderate.mediation")
```

Estimation and inference

The estimation and inference are conducted via the `modmed` function. A challenge of moderated mediation analysis lies in the specification of the mediator and outcome models in the form of Eq. (8), which can be very complex as more interactions are involved. To ease model building, the function allows users to specify the main model predictors and the moderators of each main model coefficient following the hierarchical form in Eqs. (5) and (7). In the NEWWS example, with a binary treatment, a binary mediator, and a continuous outcome, we assume that paths $T \rightarrow M$, $T \rightarrow Y$, $M \rightarrow Y$, and $TM \rightarrow Y$ are moderated by the number of children and welfare amount at baseline, while $\boldsymbol{X} \rightarrow M$ and $\boldsymbol{X} \rightarrow Y$ are not moderated. We are interested in the difference in the mediation mechanism between those who had three or more children and those who had two children at baseline, while baseline welfare amount is set at its median, 5050. Hence, we specified the function as follows:

```
R> results = modmed(data = newws,
+  treatment = "treat",
+  mediator = "emp",
+  outcome = "depression",
+  covariates.disc = c("emp_prior",
+  "nevmar", "hispanic", "nohsdip"),
+  covariates.cont = c("workpref",
+  "attitude", "depress_prior"),
+  moderators.disc= "CHCNT",
+  moderators.cont = "ADCPC",
+  m.model = list(intercept = c("ADCPC",
+  "CHCNT"), treatment = c("ADCPC",
+  "CHCNT"), emp_prior = NULL,
+  nevmar = NULL, hispanic = NULL,
+  nohsdip = NULL, workpref = NULL,
+  attitude = NULL, depress_prior = NULL),
+  y.model = list(intercept = c("ADCPC",
+  "CHCNT"), treatment = c("ADCPC",
+  "CHCNT"), mediator = c("ADCPC",
+  "CHCNT"), tm = c("ADCPC", "CHCNT"),
+  emp_prior = NULL, nevmar = NULL,
+  hispanic = NULL, nohsdip = NULL,
+  workpref = NULL, attitude = NULL,
+  depress_prior = NULL),
+  comp.treatment.value = 1,
+  ref.treatment.value = 0,
+  comp.mod.disc.values = 3,
+  ref.mod.disc.values = 2,
+  comp.mod.cont.values = 5050,
+  ref.mod.cont.values = 5050,
+  m.scale = "binary",
+  y.scale = "continuous",
+  method = "mc",
+  nmc = 1000,
+  conf.level = 0.95,
+  seed = 1)
```

Variable names Users are required to specify the names of the data, treatment, mediator, and outcome via the arguments, `data`, `treatment`, `mediator`, and `outcome`. If there are any missing values in the data, users need to impute them before running the function. Whenever applicable, names of the discrete and continuous moderators $\boldsymbol{W}$ and other pretreatment covariates $\mathbf{X}$ are specified separately via `moderators.disc`, `moderators.cont`, `covariates.disc`, and `covariates.cont`, whose default values are `NULL`. For example, if there are no discrete moderators, users do not need to specify `moderators.disc`. Users do not need to reformat discrete variables. No matter if the discrete variables are coded as string or numerical values, the program can automatically factorize them. This eases data cleaning and avoids mistakes.

$\boldsymbol{W}$ contains pretreatment covariates that moderate the mediation mechanism. As clarified in the definition and identification sections, $\boldsymbol{W}$ is chosen based on theoretical interest. $\mathbf{X}$ contains other pretreatment covariates that are confounders within levels of $\boldsymbol{W}$. If treatment is randomized, $\mathbf{X}$ should contain confounders of the mediator–outcome relationship within levels of $\boldsymbol{W}$. If treatment is not randomized, $\mathbf{X}$ should also contain confounders of the treatment–mediator and treatment–outcome relationships within levels of $\boldsymbol{W}$. If $\boldsymbol{W}$ is not specified, only the population average effects are estimated and tested.

Table 2 lists the covariates collected prior to randomization in NEWWS. More details of the data, treatment, mediator, and outcome can be found in the section of application example

Model specification Users can specify the mediator and outcome models via `m.model` and `y.model`, each as a list. The names of the elements in each list include the intercept and the predictors in the corresponding main model in Eq. (5). Specifically, in `m.model`, the names must include `intercept`, `treatment`, and the covariates in `covariates.disc` and `covariates.cont` that predict the mediator (i.e., a covariate that only confounds the treatment-outcome relationship does not need to be included). In `y.model`, the names must include `intercept`, `treatment`, `mediator`, and the covariates in `covariates.disc` and `covariates.cont` that predict the outcome (i.e., a covariate that only confounds the treatment-mediator relationship does not need to be included). If the treatment is assumed to interact with the mediator when affecting the outcome, an additional element should be added to `y.model`, named as `tm`. We suggested users always include `tm` unless it barely changes the final estimates. As VanderWeele (2015) wrote on page 46, “An investigator might be tempted to only include such exposure-mediator interactions in the model if the interaction is statistically significant. This approach is problematic. It is problematic because power to detect interaction tends to be very low unless the sample size is very large. Such exposure-mediator interaction may be important in capturing the dynamics of mediation. A better approach is perhaps to include them by default and only exclude them if they do not seem to change the estimates of the direct and indirect effects very much".

Each element of each list is equal to a vector of moderators in the function of the corresponding main model coefficient as in Eq. (7). Each moderator specified in `moderators.disc` and `moderators.cont` must moderate at least one slope in either the main mediator model or the main outcome model. The vector corresponding to the intercept must contain all the moderators in the corresponding model because their coefficients represent the main effects of the moderators. If a main

Table 2 Pretreatment covariates used in the NEWWS analysis

Type	Variable	Description
X	emp_prior	1 if employed and 0 otherwise
X	nevmar	1 if never married and 0 otherwise
X	hispanic	1 if Hispanic and 0 otherwise
X	nohsdip	1 if had never obtained a high school diploma or a General Educational Development certificate and 0 otherwise
X	workpref	One's level of preference for taking care of family full time than working on the scale of 1–4
X	attitude	A composite score of two attitude items – "so many family problems that I cannot work at a full time or part time job"; "so much to do during the day that I cannot go to a school or job training program – measured on the scale of 1–4"
X	depress_prior	A composite score of three depressive symptom items – sad, depressed, blues, and lonely – in the week before randomization measured on the scale of 1–4
W	CHCNT	1 if had 1 child, 2 if had 2 children, and 3 if had 3 or more children before randomization
W	ADCPC	Welfare amount in the year before randomization

model coefficient in Eq. (7) is not moderated, the corresponding vector of moderators needs to be specified as `NULL`. The set of moderators in `m.model` and that in `y.model` are not necessarily the same. Specifically, for a moderator that moderates coefficient(s) in one model, it is possible that it does not occur in the other model; it is also possible that it is included in the other model as a predictor rather than a moderator, in which case users will only include it in the vector corresponding to the intercept when specifying the list for the other model. The moderators that are included in both the mediator model and the outcome model essentially also confound the mediator–outcome relationship. The union of the two sets of moderators in `m.model` and `y.model` should be the same as the union of `moderators.disc` and `moderators.cont`.

In the NEWWS example, it is assumed that the paths *T*➔*M*, *T*➔*Y*, *M*➔*Y,* and *TM*➔*Y* are moderated by both `"CHCNT"` and `"ADCPC"`.

Variable values and scales Users can specify t and t', respectively, via `comp.treatment.value` and `ref.treatment.value`, where "`comp`" stands for comparison, while "`ref`" stands for reference. $t=1$ and $t'=0$ by default. In other words, if treatment is binary, one does not need to specify t. Similarly, users can specify vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ in Eq. (3), respectively, through `comp.mod.disc.values` (and/or `comp.mod.cont.values`) and `ref.mod.disc.values` (and/or `ref.mod.cont.values`), whose default values are `NULL`. To be specific, if the focal interest is in the conditional effects rather than the moderated effects, users do not need to specify `comp.mod.disc.values` or `comp.mod.cont.values`; if there are no discrete moderators, users do not need to specify `comp.mod.disc.values` or `ref.mod.disc.values`; if there are no continuous moderators, users do not need to specify `comp.mod.cont.values` or `ref.mod.cont.values`. If not `NULL`, the length and order of each value vector should be consistent with those of the corresponding name vector.

If one does not want to condition some moderators on specific values, one may specify their values to be `NA`. In the NEWWS example, if we want to check how each effect varies by CHCNT over all the individuals rather than those whose ADCPC is at 5050, we could replace `5050` with `NA` in the above syntax.

The mediator and the outcome are continuous by default. If the mediator (outcome) is binary, `m.scale` (`y.scale`) needs to be specified as `"binary"`, and a probit regression is fitted.

In the NEWWS example, we evaluated the difference in the mediation mechanism between `CHCNT = 3` vs. `CHCNT = 2` while `ADCPC` is set at its median (5050).

Other arguments There are additional arguments that users do not need to specify unless they want to change the default. Users can specify the estimation method via `method`, which is set to `"mc"` if the Monte Carlo method is chosen and to `"boot"` if the bootstrapping method is chosen (default = `"mc"`). If `method = "mc"`, one can specify the number of simulations that the Monte Carlo algorithm takes via `nmc` (default = `1000`); If `method = "boot"`, one can specify the number of bootstrapped samples via `nboot` (default = `1000`). `conf.level`, which indicates the confidence level, is set to 0.95 by default. Users can also specify `seed` (default = `NULL`).

Numerical summary of analysis results

The output of the `modmed` function can be passed via `object` to the `summary_modmed` function for numerical summary of the analysis results.

```
R> summary_modmed(object = results)
```

The output includes the estimation and inference results of the population average, conditional, and/or moderated effects. In addition, it also includes the fitting results of the mediator and outcome models, which allow one to assess the moderation of not only each mediation effect but also each specific path. The model fitting results are shown in the hierarchical form and thus better demonstrate how each specific path is moderated. A complete output for the NEWWS example can be found in Appendix E.

The results for the NEWWS data show that the total LFA effect is estimated to be 0.48 (SE = 0.67, 95% CI = [-0.86, 1.78]), which can be decomposed into a pure (natural) direct effect, estimated to be 1.24 (SE = 0.72, 95% CI = [–0.20, 2.62]), about 16.07% of a standard deviation of the outcome in the control group[8]; and a total (natural) indirect effect, estimated to be -0.76 (SE = 0.35, 95% CI = [-1.47, -0.10]), about -9.88% of a standard deviation of the outcome in the control group. The pure direct effect indicates that, LFA would have increased maternal depression if one's employment experience is held at the level under the control condition, but not by a statistically significant amount. In contrast, the total indirect effect reflects that the LFA-induced increase in employment rate significantly reduced one's maternal depression, when the treatment is held at the LFA condition. The counteracting indirect and direct effects explained the insignificant total effect.

The total indirect effect can be further decomposed into a pure indirect effect, which is estimated to be -0.02 (SE = 0.23, 95% CI = [-0.45, 0.45]); and a natural treatment-by-mediator interaction effect, which is estimated to be -0.75 (SE = 0.42, 95% CI = [-1.62, 0.03]). It indicates that the LFA-induced increase in employment rate reduced maternal depression more under the LFA condition (i.e., total indirect effect) than under the control condition (i.e., pure indirect effect). Equivalently, the natural treatment-by-mediator interaction effect can also be viewed as the difference between the total direct effect and the pure direct effect, which indicates that LFA would have increased maternal depression to a smaller extent if holding one's employment experience under the LFA condition than if holding that under the control condition.

Through the moderated mediation analysis, we further detected a significantly positive pure direct effect among those who had two children at baseline and received median level ($5050) of welfare in the year prior to randomization, which is estimated to be 3.15 (SE = 1.27, 95% CI = [0.71, 5.59]). This conditional pure direct effect is significantly higher than that of those with three children and $5,050 welfare in the year prior to randomization. The magnitude of the difference is estimated to be 3.71 (SE = 1.68, 95% CI = [0.34, 6.92]).

The above numerical summary provides important information about the conditional effects at a given set of values of the moderators and how they differ between two given sets of values of the moderators. To further assess how the effects vary across the whole value range of a moderator, the `R` package offers a graphical tool.

Graphical summary of analysis results

The output of the `modmed` function can be passed via `object` to the `modmed.plot` function for generating a whole picture of how the mediation mechanism differs by a moderator.

```
R> modmed.plot(object = results,
+  effect = "PDE",
+  moderator = "CHCNT",
+  other.mod.cont.values = 5050)
```

Plot of a conditional causal effect versus a moderator Each time users choose one focal effect (specified via `effect`, which can be `"TE"`, `"TIE"`, `"PDE"`, `"PIE"`, `"TDE"`, or `"INT"`) to be plotted against a focal moderator (specified via `moderator`), while conditioning on given values of the other moderators (specified via `other.mod.disc.values` for discrete moderators and/or `other.mod.cont.values` for continuous moderators). The length and order of `other.mod.disc.values` must match those of `moderators.disc` (specified in the `modmed` function) with the focal moderator removed if it is discrete. If one does not want to condition some moderators on specific values, one may specify their values to be `NA`. The same is true for `other.mod.cont.values`.

Applying the above syntax to the NEWWS data, we obtained Fig. 1, which represents how the pure direct effect varied with the number of children that participants had at baseline, among those who received $5050 welfare in the

[8] To obtain mediation effect estimates that are comparable across studies, some researchers reported mediation effect estimates in the standard deviation of the outcome in the control group (Kraft, 2020), or the standard deviation of the outcome in the whole sample (e.g., Hong et al.,), while some fully standardized the mediation effects by standardizing all the variables (e.g., Preacher & Hayes, 2008). If a researcher would like to report such effect size measures in the following graphical summary and sensitivity analysis, they may run the analysis based on the data with the corresponding variables standardized in the desired way. We chose to report the estimates in the standard deviation of the outcome in the control group to eliminate the influence of the treatment on the standard deviation of the outcome. As Kraft (2020) argued on page 245, "it is preferable to use the standard deviation of the control group outcome rather than the pooled sample because the intervention may have affected the variation in outcomes among the treatment group".

year before randomization. For each category of the moderator, a boxplot is generated to represent the Q estimates of the conditional PDE obtained from the Monte Carlo or bootstrapping algorithm. Each blue dot in the middle represents the final effect estimate, while each blue interval denotes the confidence interval at the confidence level specified in the `modmed` function.

From Fig. 1, among those who received $5050 welfare in the year before randomization, the pure direct effect of LFA on maternal depression is estimated to be positive among those with one or two children at baseline, while significant only among those with two children. In contrast, among those with three or more children, LFA would have reduced maternal depression if not changing one's employment experience, but not by a statistically significant amount. This matches and supplements the numerical summary obtained from the `summary_modmed` function.

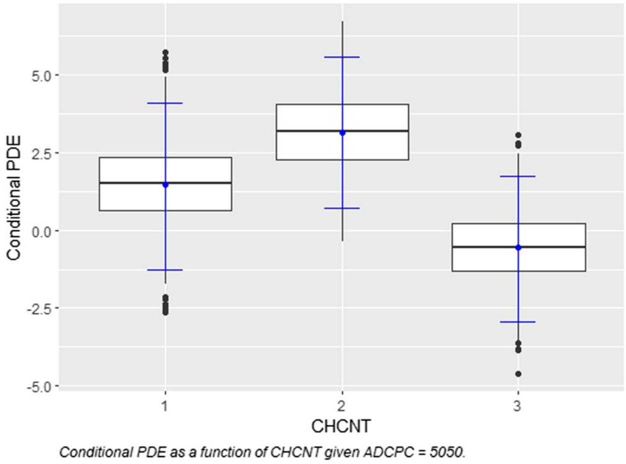


Fig. 1 Plot of pure direct effect versus CHCNT given ADCPC = 5050

Conditional distribution of a continuous moderator If the focal moderator is continuous, one can choose to add its sample distribution given values of the other moderators via `is.dist.moderator = TRUE`. It will not be plotted if the subsample within the given levels of the other moderators has fewer than 30 observations. Five percentile lines are added to the distribution of the moderator, which facilitates the evaluation of conditional effects at given percentiles of each continuous moderator. The percentiles can be specified via `probs`. The default is `c(0.1, 0.25, 0.5, 0.75, 0.9)`.

By running the following syntax, we further assessed how the pure direct effect varied with welfare amount in the year before randomization and how the welfare amount is distributed, among those who had two children at baseline.

```
R> modmed.plot(object = results,
+  effect = "PDE",
+  moderator = "ADCPC",
+  other.mod.disc.values = 2,
+  is.dist.moderator = TRUE,
+  probs = c(0.1, 0.25, 0.5, 0.75, 0.9),
+  ncore = 8)
```

As shown in Fig. 2, the estimates of the conditional PDE are represented by the blue curve. To obtain such a curve, the function estimates the conditional PDE at each of 15 evenly spaced values of the moderator and fits a loess line to the 15 estimates. Similarly, we obtained the grey band, which represents the confidence band.[9] The dashed vertical lines divide the significant and insignificant regions. The limit of the x-axis is determined by the range of the focal moderator in the whole sample.

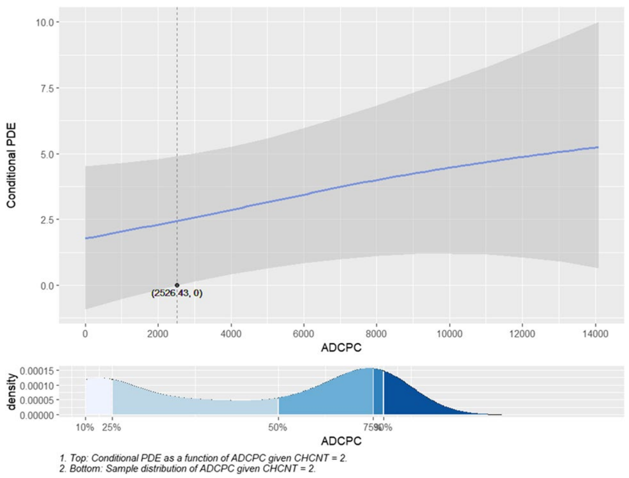
From Fig. 2, among those who had two children at baseline, the pure direct effect of LFA on maternal depression increased as one received more welfare amount in the year prior to randomization. The effect is significant when the welfare amount is larger than $2526. The sample distribution of the welfare amount indicates that about 70% of the participants with two children are in the significant region.

The figures of all the effects can be found in Appendix F. These figures provide researchers with a general picture of how each effect changes with a moderator given values of the other moderators, as well as the moderator values at which the sign or significance of the effect is flipped. If particularly interested in a visualized difference in an effect between two values of the focal moderator, one may use the `modmed` function to estimate the difference and test its significance. If the figure shows that an effect is linearly associated with a continuous moderator, one may estimate and test the difference in the effect between two arbitrarily chosen moderator values that are one unit apart, which implies the average change in the effect with one unit increase in the moderator.

Sensitivity analysis

In a causal mediation analysis, it is important to theoretically discuss which unmeasured variables may be confounders, assess how plausible it is, and evaluate how additional

[9] It is inappropriate to simply plot $\hat{\delta}_i$ predicted from Step 2.4 based on one replication against the moderator, because it ignores the uncertainty of $\hat{\delta}_i$ when constructing the confidence band. The cost of the proposed approach is that it is more time-consuming. To speed up the calculation, the package allows parallel computing by setting the number of CPU cores via `ncore`. Its default value is 2. A progress bar is displayed as the program runs. An application of the above syntax to the NEWWS data took 2.5 minutes. In the cases when Step 2.5 does not involve a spline regression, it usually takes around 30 seconds with only one core.



1. Top: Conditional PDE as a function of ADCPC given CHCNT = 2.
2. Bottom: Sample distribution of ADCPC given CHCNT = 2.

Fig. 2 Plot of pure direct effect versus ADCPC given CHCNT = 2

adjustment for these potential unmeasured confounders would have changed the sign or significance of the effects. The sensitivity of the analysis results to unmeasured pretreatment confounding can be assessed by passing the output of the `modmed` function via `object` to the `modmed.sens` function.

```
R> sens.results = modmed.sens(object = results,
+  sens.effect = c("TIE", "PIE", "TDE", "PDE",
+  "INT","TIE.ref", "PIE.ref", "PDE.ref",
+  "TDE.ref", "INT.ref", "TIE.dif", "PIE.dif",
+  "PDE.dif", "TDE.dif", "INT.dif"),
+  range.b.m = NULL,
+  range.b.y = NULL,
+  grid.b.m = 10,
+  grid.b.y = 10,
+  U.scale = "binary",
+  p.u = 0.5,
+  t.rand = TRUE,
+  t.model = NULL,
+  t.scale = "binary",
+  b.t = NULL,
+  iter = 10,
+  nsim = 5,
+  ncore = 8)
```

The effects whose sensitivity will be evaluated can be specified via `sens.effect`. It is a vector, c(`"TIE", "PIE", "TDE", "PDE", "INT","TIE.ref", "PIE.ref", "PDE.ref", "TDE.ref", "INT.ref", "TIE.dif", "PIE.dif", "PDE.dif", "TDE.dif", "INT.dif")`, by default (i.e., if `sens.effect` is not specified). `sens.effect` can also be specified as a subvector of the default. It does not matter how the effects are ordered. For example, if a researcher is mainly interested in the sensitivity of moderated indirect effects, `sens.effect` can be specified as `c("PIE.dif", "TIE.dif")`. Following the algorithm for sensitivity analysis, users can specify value ranges for the sensitivity parameters via `range.b.m` and `range.b.y` (e.g., `c(-2, 2)`); the values are automatically generated if `NULL` is specified[10]. They are divided into a `grid.b.m`-by-`grid.b.y` grid, which is 10-by-10 by default. The finer the grid, the smoother the sensitivity analysis plot in Fig. 3. The sensitivity parameters are the slopes of U in the standardized mediator and outcome models, in which both the independent and dependent variables are standardized. If the dependent variable is binary, its latent index is standardized instead (Grace et al., 2018). The variance of the latent index is the sum of the predictor variance and the assumed error variance (McKelvey & Zavoina, 1975).

As clarified in the sensitivity analysis section, the conditional distribution of U also relies on the marginal distribution of U, which is assumed to be binary (`U.scale = "binary"`) with $\Pr(U=1)=0.5$ (`p.u = 0.5`) by default. One can change `U.scale` to be `"continuous"` for a continuous U, which is assumed to follow a standard normal distribution (`sigma.u = 1`) by default. `p.u` and `sigma.u` can be adjusted.

By default, the treatment is randomized (`t.rand = TRUE`). In other words, there is no confounder of the treatment–mediator or treatment–outcome relationship. Hence, by default, `t.model = NULL`. If treatment is not randomized (`t.rand = FALSE`), one needs to specify `t.model` in the same way as specifying `m.model` and `y.model`, because a treatment model is required for the derivation of the conditional distribution of the unmeasured pretreatment confounder U as shown in Appendix D. The names of the elements in `t.model` must include `intercept` and the covariates in `covariates.disc` and `covariates.cont` that predict the treatment (i.e., a covariate that only confounds the mediator-outcome relationship does not need to be included). Each element of the list is equal to a vector of moderators in the function of the main treatment model coefficient. The moderators of the intercept must contain all the moderators in the treatment model because their slopes represent the main effects of the moderators. If a main model coefficient is not moderated, the corresponding vector of moderators needs to be specified as `NULL`. A moderator in the mediator and outcome models does not necessarily moderate any coefficient in the treatment model. If it is only a predictor rather than a moderator in the treatment model, users will only include it in the vector corresponding to the intercept when specifying `t.model`. For example, suppose the treatment is not randomized, we may specify that `t.model` in the `modmed.sens` function as `t.model = list(intercept = c("ADCPC", "CHCNT"), emp_prior = "ADCPC", nevmar = "CHCNT", hispanic = NULL, nohsdip = NULL, workpref = NULL, attitude = NULL, depress_prior = NULL)`.

[10] The default range is symmetric, and the upper limit of `range.b.m` (`range.b.y`) is twice the magnitude of the largest coefficient of the observed covariates in the standardized mediator (outcome) model.

Besides, a value for sensitivity parameter β_u^t needs to be specified via `b.t`. One may use the standardized coefficient estimates of the observed covariates in the treatment model as referent values for the specification of `b.t`. The treatment is binary by default. If the treatment is continuous, `t.scale` needs to be specified as `"continuous"`. The sensitivity analysis is conducted given one single value of β_u^t. If there are multiple values of β_u^t to be assessed, one may conduct the sensitivity analysis multiple times to assess the influence of each value of β_u^t, one at a time.

In addition, users can specify the number of iterations that the stochastic EM algorithm takes for simulating *U* (i.e., *S*) and the number of simulations of *U* (i.e., *K*) respectively via `iter` and `nsim`, which are 10 and 5 by default. Each run of the `modmed` function only takes seconds, while the `modmed.sens` function involves `nsim` runs of the `modmed` function, each of which involves `iter` iterations, for each cell of the grid. To speed up calculation, the package allows parallel computing by setting the number of CPU cores via `ncore.` Its default value is 2. A user may use more cores to speed up the program. `detectCores()` can be used to detect the number of all cores available on a computer. It is not recommended to use up all the cores. At least one core should be saved for users to run other programs on the computer while running the R program. A progress bar is displayed as the program runs. An application of the above syntax to the NEWWS data took about 1 hour. It takes less time if `sens.effect` is specified as a subvector of all the effects. For example, it takes 15 minutes if `sens.effect = "TIE.ref"`. We reduced the size of the grid to 5-by-5 and obtained similar results for all the effects in 15 minutes, while the visualized results are less smooth than those in Fig. 3. We would suggest users use a coarse grid (e.g., 5-by-5) to get a preliminary sense of sensitivity and then use a fine grid (e.g., 10-by-10) for the final report of the sensitivity analysis results. In the cases when Step 2.5 does not involve a spline regression, it takes less than 10 minutes for a 10-by-10 grid with only one core.

The output includes a `grid.b.m`-by-`grid.b.y` table separately for the point estimates, CI lower bounds, and CI upper bounds of each effect after adjusting for *U* at the strengths reflected by the column and row names. Each column (row) name corresponds to one of the `grid.b.y` (`grid.b.m`) sensitivity parameter values that evenly divide `range.b.y` (`range.b.m`). The table can be extracted via `sens.results$results.new`. The output also includes slopes of each observed pretreatment confounder in the standardized mediator and outcome models, which can be used as referent values to calibrate the strength of unmeasured pretreatment confounding. The slopes can be extracted via `sens.results$X.coef.plot`.

In the NEWWS example, as illustrated in the identification section, one's salary expectation at baseline is a potential unmeasured confounder of the relationship between employment and depression within levels of baseline welfare amount and number of children. To assess how an additional adjustment for such an unmeasured confounder would affect the initial results, one may specify the sensitivity parameter values based on prior knowledge about its association with employment and depression or a comparable observed pretreatment covariate. For example, it may be reasoned that the unique confounding role of baseline salary expectation is similar to that of baseline employment. Hence, one may use the standardized coefficient estimates of baseline employment in the mediator and outcome models as referent values when specifying the sensitivity parameters, by setting `range.b.m` and `range.b.m` to the values and setting `grid.b.m` and `grid.b.m` to 1:

```
R> modmed.sens(object = results,
+  range.b.m = 0.31, range.b.y = 0.04,
+  grid.b.m = 1, grid.b.y = 1)
```

Correspondingly, we obtained the adjusted estimates of the causal effects, as shown in the right panel of Table 3. A comparison of the adjusted results with the original results in the left panel of Table 3 reveals little influence of the additional adjustment on the estimation and inference results. Therefore, the original conclusions may be robust to the omission of baseline salary expectation from the analysis.

Visualization of sensitivity analysis results

The output of the `modmed` function and the output of the `modmed.sens` function can be passed respectively via `object` and `sens.results` to the `sens.plot` function for visualizing the sensitivity analysis results for each effect specified via `effect`, which must be included in `sens.effect` of the `modmed.sens` function.

```
R> sens.plot(object = results,
+  sens.results = sens.results,
+  effect = "PIE")
```

Applying the above syntax to the NEWWS data, we obtained Fig. 3[11]. The *X* and *Y* axes are respectively the slopes of *U* in the standardized mediator and outcome models. The blue number on the origin indicates the original estimate of the pure indirect effect, without adjustment for *U*. Each black contour represents the combinations of sensitivity

[11] The sensitivity analysis plot looks similar to the L.O.V.E. method (Mauro, 1990; Cox et al., 2013), while we allow more scenarios (e.g., inclusion of the treatment-by-mediator interaction) for assessing sensitivity of conditional and moderated causal mediation effects.

Table 3 The effect estimates before and after adjustment for *U*

	Before Adjustment for *U*			After Adjustment for *U*		
	Estimate	95% CI Lower 2.5%	95% CI Upper 2.5%	Estimate	95% CI Lower 2.5%	95% CI Upper 2.5%
TIE	–0.765	–1.472	–0.098	–0.823	–1.535	–0.167
PIE	–0.016	–0.449	0.455	–0.052	–0.516	0.395
PDE	1.244	–0.196	2.619	1.309	–0.054	2.667
TDE	0.496	–0.914	1.906	0.538	–0.758	1.878
INT	–0.749	–1.619	0.027	–0.771	–1.613	0.033
TIE.ref	–0.956	–2.499	0.305	–1.039	–2.500	0.251
PIE.ref	0.166	–0.730	1.052	0.136	–0.770	1.072
PDE.ref	3.155	0.706	5.587	3.251	0.848	5.779
TDE.ref	2.032	–0.405	4.448	2.076	–0.289	4.469
INT.ref	–1.122	–2.894	0.509	–1.175	–2.919	0.340
TIE.dif	0.437	–0.951	2.108	0.490	–0.985	2.126
PIE.dif	–0.164	–1.057	0.744	–0.158	–1.107	0.823
PDE.dif	–3.706	–6.921	–0.335	–3.746	–7.130	–0.370
TDE.dif	–3.106	–6.442	0.120	–3.098	–6.407	0.201
INT.dif	0.601	–1.099	2.534	0.648	–1.182	2.551

parameters that result in the adjusted pure indirect effect estimate as indicated by the number at the end of the contour. The sensitivity parameters on the red dashed contours reduce the PIE estimate to zero. Each blue dotted contour indicates the boundary at which the significance of the PIE is changed at the confidence level specified in the `modmed` function. The adjusted PIE is insignificant in the region that contains the zero lines. The results are more sensitive if the magnitudes of the sensitivity parameters that change the sign or significance of PIE are smaller. The strengths of the sensitivity parameters can be calibrated by the black dots. Each dot is plotted based on the slopes of each observed pretreatment confounder in the standardized mediator and outcome models and thus represents an unmeasured confounder comparable

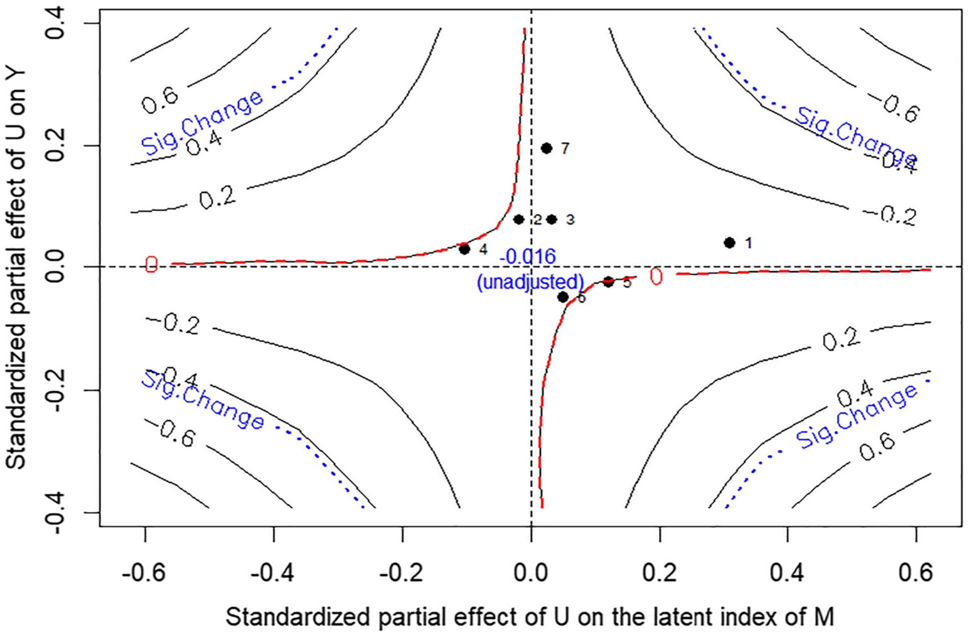


Fig. 3 Sensitivity analysis plot for the pure indirect effect (The range of X (Y) axis depends on the width of range.b.m (range.b.y) specified in the modmed.sens function; 1–7 are respectively comparable to `emp_prior`, `nevmar`, `hispanic`, `nohsdip`, `workpref`, `attitude`, and `depress_prior`

to the observed confounder. A message of which observed confounder each dot is comparable to is displayed after each run of the function.

Fig. 3 shows that an additional adjustment of a binary unmeasured pretreatment confounder U with $\Pr(U=1)=0.5$ would reverse the sign of PIE as long as the confounding role of U is as strong as one's baseline educational attainment (`nohsdip`) or work preference (`workpref`). Nevertheless, for the significance of PIE to be altered, U must be much stronger than the strongest observed pretreatment confounder. Because we have controlled for the most important pretreatment confounders in theory, it is almost impossible for the remaining confounders to reverse the significance of PIE even collectively. Hence, the sign of PIE may be sensitive but its significance is more robust. All the sensitivity plots can be found in Appendix G. The plots show that the other effects are mostly robust to unmeasured pretreatment confounding.

In addition, we assessed sensitivity to a continuous U that follows a standard normal distribution and obtained very similar results. It is more computationally intensive because the derivation of the conditional distribution of a continuous U relies on adaptive numerical integration. Hence, if U is expected to follow a standard normal distribution, one may assess its influence by specifying U to be binary with $\Pr(U=1)=0.5$ to reduce computing time.

It is worth noting that, if the treatment is not randomized and if β_u^t is specified to be nonzero, the contours would cross the X and Y axes, because even though U is not associated with M (Y), it can still cause bias as long as U is associated with Y (M), given that U is associated with T. In addition, as shown in Fig. 3, the sensitivity analysis plot is symmetric with respect to the origin, indicating that the bias caused by unmeasured confounder U_1 and that caused by U_2 would be the same when $\beta_{u1}^m = -\beta_{u2}^m$ and $\beta_{u1}^y = -\beta_{u2}^y$. Similarly, when the treatment is not randomized, the bias would be the same only if $\beta_{u1}^t = -\beta_{u2}^t$ in addition to $\beta_{u1}^m = -\beta_{u2}^m$ and $\beta_{u1}^y = -\beta_{u2}^y$. Therefore, at a given nonzero value of β_u^t, the sensitivity analysis plot is asymmetric about the origin.

Discussion

In this article, we proposed general definitions of causal conditional and moderated mediation effects under the potential outcomes framework, independent of any statistical models. We identified the effects under the conditional sequential ignorability assumption and compared through simulations three widely used estimation methods. We also extended a simulation-based sensitivity analysis method, which accounts for the influence of unmeasured pretreatment confounding on both estimation bias and estimation efficiency. The overall analytic procedure is applicable to a binary or continuous treatment, a binary or continuous mediator and outcome, one or more moderators of any scale, and a wide range of scenarios of moderated mediation, while the sensitivity analysis assumes a continuous treatment, mediator, outcome, or unmeasured confounder to be normally distributed.

In addition, we developed an easy-to-implement R package, which greatly facilitates the application of the proposed methods. The package has its unique advantages. First, the argument specification for the estimation function guides users through a systematic process of incorporating moderators in the model building, and the output allows one to assess the moderation of not only each mediation effect but also each specific path. Different from a standard regression output, the output of mediator and outcome models is shown in the hierarchical form and thus better demonstrates how each specific path is moderated. Second, the graphical representations of the estimation and inference results allow users to visually assess how the point estimate and significance of a mediation effect vary with a moderator, and the change pattern can be linear or nonlinear. Third, the sensitivity analysis plots enable applied researchers to conclude how sensitive the analysis results are to unmeasured pretreatment confounding, which fills an important gap in the literature of traditional moderated mediation analysis.

Researchers have discussed various ways of assessing whether a mediation effect is moderated, as reviewed in the introduction section. A common practice in psychology research in particular is to dichotomize a continuous moderator and test the difference in the mediation effect between the two subgroups. However, it has been criticized for decades because it loses information, results in biased estimates, and reduces power (e.g., Cohen, 1983; Edwards & Lambert, 2007; MacCallum et al., 2002; Maxwell & Delaney, 1993). We recommend that researchers utilize the `modmed.plot` function for an overview of the influence of a moderator on the magnitude, sign, and significance of a mediation effect. If the mediation effect linearly changes with a continuous moderator, one can assess if the mediation effect is significantly moderated by testing the difference in the effect between any two moderator values that are one unit apart via the `modmed` function. Otherwise, one can locally assess moderated mediation by testing the difference in the effect between two given values of the moderator.

Future research can make various extensions of the proposed approach. First, the proposed approach relies on correct specifications of the parametric mediator and outcome models for the estimation and inference of the causal effects. The causal effect estimates would be biased if a confounder interacts with the treatment, mediator, treatment-by-mediator interaction, or other confounders but such interactions are ignored in the model specifications. The potential for bias increases as the number of covariates increases. One may test whether the analysis results are robust to model-based assumptions by examining whether alternative model specifications would generate similar results. One may also fit nonparametric models or machine learning models of the mediator and outcome in Step 2.1 to allow for more robust and flexible modeling. Alternatively, Steps 2.1 – 2.4 can be replaced with weighting-based or imputation-based approach for predicting the effects. In addition, Step 2.5 can be replaced with machine learning-based methods for more flexible investigations of how the effects vary by moderators. Second, in general, a relatively large sample size is required to detect a significant moderated mediation effect. Qin (2023) developed an R shiny app (https://xuqin.shinyapps.io/CausalMediationPowerAnalysis/) to calculate power for detecting significant population average causal mediation effects at a given sample size or calculate the sample size required to achieve an adequate power. Straightforward extensions can be made for causal moderated mediation analysis. Third, in this study, we assume no posttreatment confounding of the mediator-outcome relationship. Future research can be done to relax this assumption by extending Daniel et al. (2015) and Hong, Yang, and Qin (2023). Fourth, extensions can be made for longitudinal or multilevel moderated mediation analysis, to account for measurement error, or to incorporate multiple mediators that are parallel or interact with each other.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02095-4.

Acknowledgements This study received support from Pitt Momentum Funds, a U.S. Department of Education Institute of Education Sciences Grant (R305D200031) an National Institutes of Health Grant (R01AG080590), and the 2022 National Academy of Education (NAEd)/Spencer Postdoctoral Fellowship Program. Lijuan Wang is grateful for the support from IES grant R305D210023 during the study. The authors thank Guanglei Hong and Fan Yang for their insightful comments on previous versions of the manuscript. In addition, comments from the Associate Editor and the reviewers have led to major improvements in this article.

References

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.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies, 5*(2), 21–35.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013). Sensitivity plots for confounder bias in the single mediator model. *Evaluation Review, 37*(5), 405–431.

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

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science, 10*(2), 87–99.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling–based approaches to the calculation of posterior moments. In J. O. B. J. M. Bernado, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Clarendon.

Grace, J. B., Johnson, D. J., Lefcheck, J. S., & Byrnes, J. E. (2018). Quantifying relative importance: computing standardized effects in models with binary outcomes. *Ecosphere, 9*(6), e02283.

Hamilton, G., & Freedman, S. (2001). How effective are different welfare-to-work approaches? Five-year adult and child impacts for eleven programs. National Evaluation of Welfare-to-Work Strategies.

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research, 50*(1), 1–22.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proceedings of the American Statistical Association, Biometrics Section* (pp. 2401–2415). merican Statistical Association.

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

Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics, 40*, 307–340.

Hong, G., Qin, X., & Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics, 43*(1), 32–56.

Hong, G., Yang, F., & Qin, X. (2023). Post-treatment confounding in causal mediation studies: A cutting-edge problem and a novel solution via sensitivity analysis. *Biometrics , 79*(2), 1042-1056.

Imai, K., Keele, L., & Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods, 15*, 309.

Imai, K., Keele, L., & Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science, 25*, 51–71.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69*(2), 307.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science, 44*(2), 347–361.

Loeys, T., Talloen, W., Goubert, L., Moerkerke, B., & Vansteelandt, S. (2016). Assessing moderated mediation in linear models requires fewer confounding assumptions than assessing mediation. *British Journal of Mathematical and Statistical Psychology, 69*(3), 352–374.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Erlbaum.

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.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin, 108*(2), 314.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*(1), 103–120.

Michalopoulos, C., Schwartz, C., & Adams-Ciardullo, D. (2001). National evaluation of welfare-to-work strategies. *What Works Best for Whom: Impacts of, 20*.

Miočević, M., Gonzalez, O., Valente, M. J., & MacKinnon, D. P. (2018). A tutorial in Bayesian potential outcomes mediation analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(1), 121–136.

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.

Morris, P. A. (2008). Welfare program implementation and parents' depression. *Social Service Review, 82*(4), 579–614.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2017). *Regression and mediation analysis using Mplus*. Muthén & Muthén.

Neyman, J., & Iwaszkiewicz, K. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society, 2*, 107–180.

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420). Morgan Kaufmann.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news, 6*(1), 7–11.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Contemporary approaches to assessing mediation in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 13–54). Sage.

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.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77–98.

Qin, X. (2023). *Sample Size and Power Calculations for Causal Mediation Analysis*. Behavior Research Methods: In press.

Qin, X., & Yang, F. (2022). Simulation-based sensitivity analysis for causal mediation studies. *Psychological Methods.* Advance online publication.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3*, 143–155.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association, 75*, 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25*, 279–292.

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine, 37*(14), 2252–2266.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017). medflex: An R Package for flexible mediation analysis using natural effect models. *Journal of Statistical Software, 76*, 1–46.

Ten Have, T. R., Elliott, M. R., Joffe, M., Zanutto, E., & Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association, 99*(465), 16–25.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*(5).

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.

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

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface, 2*, 457–468.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology, 172*, 1339–1348.

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods, 1*(1), 131–158.

Wang, L., & Preacher, K. J. (2015). Moderated mediation analysis using Bayesian methods. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 249–263.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*(4), 301.

Open practices statement The data and R package are available on CRAN: https://cran.r-project.org/web/packages/moderate.mediation/index.html. The data were not preregistered.