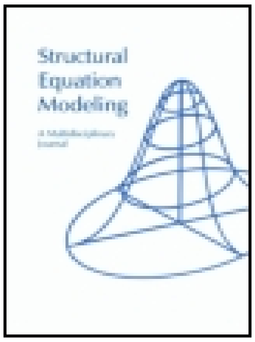

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction

Judith J.M. Rijnhart , Matthew J. Valente , David P. MacKinnon , Jos W.R. Twisk & Martijn W. Heymans

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2020.1811709

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

The use of traditional and causal estimators for mediation models with a binary outcome and exposure-mediator interaction

Judith J.M. Rijnhart[1], Matthew J. Valente[2], David P. MacKinnon[3], Jos W.R. Twisk[1], and Martijn W. Heymans[1]

[1]Department of Epidemiology & Data Science, Amsterdam Public Health Institute, Amsterdam University Medical Center; [2]Center for Children and Families, Department of Psychology, Florida International University; [3]Department of Psychology, Arizona State University

ABSTRACT

An important recent development in mediation analysis is the use of causal mediation analysis. Causal mediation analysis decomposes the total exposure effect into causal direct and indirect effects in the presence of exposure-mediator interaction. However, in practice, traditional mediation analysis is still most widely used. The aim of this paper is to demonstrate the similarities and differences between the causal and traditional estimators for mediation models with a continuous mediator, a binary outcome, and exposure-mediator interaction. A real-life data example, analytical comparisons, and a simulation study were used to demonstrate the similarities and differences between the traditional and causal estimators. The causal and traditional estimators provide similar indirect effect estimates, but different direct and total effect estimates. Traditional mediation analysis may only be used when conditional direct effect estimates are of interest. Causal mediation analysis is the generally preferred method as its casual effect estimates help unravel causal mechanisms.

KEYWORDS
Mediation analysis; potential outcomes; interaction; binary outcome

Introduction

Mediating variables are central to structural equation modeling (SEM) methodology, serving as a motivation for methods development (Judd & Kenny, 1981; Sobel, 1982) and the substantive applications of SEM (Bollen, 1989; Kline, 2015; Little, 2013). SEM developments for mediating variables include effect decomposition (Alwin & Hauser, 1975), general formulas for standard errors (Sobel, 1982), resampling methods (Bollen & Stine, 1990), methods for non-normal distributions (Browne, 1984), and most recently, modern causal inference methods (Bollen & Pearl, 2013). The purpose of this paper is to elucidate the application of modern causal inference methods for mediation models with a binary outcome variable.

Mediation analysis disentangles the total exposure effect into direct and indirect effects (Judd & Kenny, 1981; MacKinnon, 2008, 2020). However, the underlying causal mechanisms of exposure effects are often more complex than this. In addition to mediating the exposure effect, a mediator can also moderate the exposure effect, resulting in exposure-mediator (XM) interaction (Holland, 1988; Judd & Kenny, 1981). When there is XM interaction, the magnitude of the direct effect depends on mediator values and the magnitude of the indirect effect depends on exposure values (Judd & Kenny, 1981). However, in the traditional mediation analysis literature, limited guidance is provided for estimating direct and indirect effects in the presence of XM interaction (Judd & Kenny, 1981; MacKinnon, 2008, 2020). Reporting average direct and indirect effect estimates is therefore common practice. The reporting of average effect estimates is problematic when the XM interaction is present, because average effect estimates ignore important information on the direct and indirect effects at specific exposure and mediator values and therefore do not provide complete insight into the causal mechanisms underlying the total effect (Pearl, 2001).

An important recent development in mediation analysis is the application of causal mediation analysis (Imai et al., 2010; Pearl, 2001; Vanderweele & Vansteelandt, 2010). Causal mediation analysis provides definitions and estimators of direct and indirect effects that naturally incorporate the XM interaction (Pearl, 2001; Robins & Greenland, 1992). Causal mediation analysis is based on the potential outcomes framework and the more general counterfactual framework, and defines causal effects as the difference between two potential outcomes or counterfactuals (Holland, 1986; Pearl, 2001).

A recent study demonstrated how, in the presence of XM interaction, traditional mediation analysis can be used to estimate causal effects for models with a continuous mediator and a continuous outcome variable (MacKinnon et al., 2020). However, the traditional mediation analysis methodology is less suited for the estimation of causal mediation effects when the outcome is binary. This is due to differences in the types of effects estimated by causal and traditional mediation analysis when the underlying models are non-linear, i.e., average effects versus conditional effects, respectively. Since in practice both causal and traditional mediation analysis are used for the analysis of mediation models with binary outcomes (Vo et al., 2020), it is important that researchers know when they can expect traditional mediation analysis to yield causal effect estimates. When the traditional and causal methods provide

CONTACT Judith J.M. Rijnhart j.rijnhart@amsterdamumc.nl Department of Epidemiology and Data Science, Amsterdam UMC, Location VU University Medical Center, Amsterdam, MB 1007, The Netherlands.

Supplemental data for this article can be accessed on the publisher's website.

different effect estimates, it is important that researchers know which of the two methods is better suited to answer their research questions.

The aim of this paper is to demonstrate the similarities and differences between the causal and traditional estimators for mediation models with a continuous mediator, a binary outcome, and XM interaction. We first provide a short overview of the basic concepts of statistical mediation analysis, followed by an introduction of the causal and traditional mediation analysis methodologies. To assess the similarities and differences between the causal and traditional estimators, we perform analytical comparisons. Then, we use a simulation study to demonstrate when researchers should expect to observe differences between the causal and traditional direct and total effect estimates. Subsequently, the estimation and interpretation of the causal and traditional effects are demonstrated using a real-life data example. Finally, we provide recommendations for the estimation of effects for models with a continuous mediator, a binary outcome, and XM interaction.

Statistical mediation analysis

Mediation analysis was developed based on path analysis and decomposes the total exposure effect (the *c* path in Figure 1) into an indirect and direct effect (Judd & Kenny, 1981; MacKinnon, 2008, 2020; Wright, 1923). The indirect effect is the part of the total effect that is explained by the mediator (the *a* and *b* paths in Figure 1), and the direct effect is the part of the total effect that is not explained by the mediator (the *c'* path in Figure 1).

When the outcome is binary and the mediator is continuous, the paths in Figure 1 can be estimated using three equations (Judd & Kenny, 1981; MacKinnon et al., 2007):

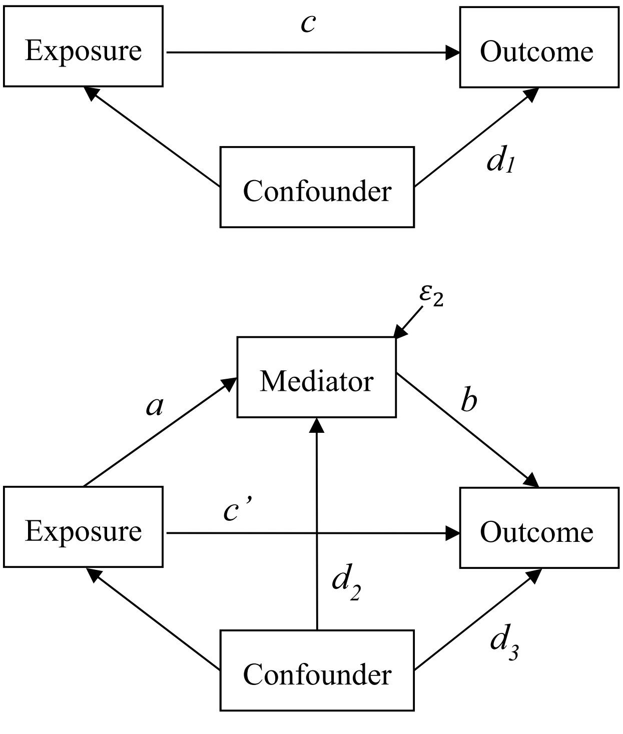

Figure 1. Path diagram of the single mediator model, including a confounder.

$$logit(Pr(Y = 1|x, z)) = i_1 + cX + d_1Z \qquad (1)$$

$$M = i_2 + aX + d_2Z + \varepsilon_2 \qquad (2)$$

$$logit(Pr(Y = 1|x, m, z)) = i_3 + c'X + bM + d_3Z \qquad (3)$$

where Equations 1 and 3 are estimated using logistic regression, yielding path coefficients on the log-odds scale, and Equation 2 is estimated using linear regression. In Equation 1, the *c* coefficient represents the effect of the exposure (X) on the outcome (Y). In Equation 2, the *a* coefficient represents the effect of the exposure on the mediator (M), and ε_2 represents the linear regression residual term. In Equation 3, the *c'* coefficient represents the effect of the exposure on the outcome when adjusted for the mediator, and the *b* coefficient represents the effect of the mediator on the outcome when adjusted for the exposure. The i_1, i_2, and i_3 terms represent intercepts, and the d_1, d_2, and d_3 terms represent the effects of a confounder on the outcome and mediator. In the absence of confounding the d_1, d_2, and d_3 terms drop out of Equations 1, 2, and 3.

Equation 3 assumes the absence of an XM interaction, but the XM interaction can be investigated by extending Equation 3 with an XM-interactioin term (Judd & Kenny, 1981; MacKinnon, 2008, 2020; Vanderweele & Vansteelandt, 2010), yielding the following equation:

$$logit(Pr(Y = 1|x, m, z)) = i_4 + c'X + bM + hXM + d_4Z \qquad (4)$$

where the *h* coefficient represents the effect of the exposure-mediator interaction on the outcome. In the following two sections we describe how the path coefficients from Equations 1, 2, and 4 can be used to estimate the causal and traditional direct, indirect, and total effects.

Causal mediation analysis

Causal mediation analysis distinguishes between causal effect definitions and causal effect estimation (Holland, 1988; Robins & Greenland, 1992). A strength of causal mediation analysis is that its effect definitions are general and can be applied to any mediation model (Pearl, 2001). Various estimation methods can be used to estimate the causal effects, including a simulation-based approach, a numerical integration approach, and a regression-based approach (Imai et al., 2010; Muthén et al., 2017; VanderWeele, 2009). In this paper, we focus on the regression-based approach, as this approach is most similar to traditional mediation analysis.

Causal effect definitions

Causal mediation analysis defines causal effects as the difference between two potential outcomes (Holland, 1988; Pearl, 2001; Robins & Greenland, 1992). Consider that we are interested in the effect of an intervention X, where *x* represents the intervention group and *x** represents the control group, on an outcome Y. In this situation, two potential outcome values can be observed for one participant; the outcome value when in the intervention group, i.e., Y(*x*), and the outcome value when in the control group, i.e., Y(*x**) (Holland, 1986). The causal intervention effect is defined as the comparison of the participant's outcome values simultaneously observed under both groups, i.e., $Y(x) - Y(x^*)$.

In mediation analysis, the potential outcome values are not only dependent on exposure values, but also on mediator values (Holland, 1988; Pearl, 2001). The potential outcomes notation is therefore extended to also include the mediator value, i.e., $Y(x, m)$ and $Y(x^*, m)$, where m represents a mediator value of interest. Suppose that we are interested in physical activity as a mediator of the relation between an intervention (where X = 1 represents intervention and X = 0 represents control) and hypertension (where Y = 1 represents hypertension and Y = 0 represents no hypertension). A participant's risk of hypertension had the participant been in the intervention group (X = 1), while holding physical activity constant at, for instance, 3 hours, is denoted as $Y(x = 1, m = 3)$. The participant's risk of hypertension had the participant been in the control group (X = 0), while holding physical activity constant at 3 hours, is denoted as $Y(x = 0, m = 3)$. The difference between the two potential outcomes $Y(x, m)$ and $Y(x^*, m)$ is the controlled direct effect (CDE) (Pearl, 2001). The CDE is therefore the direct exposure effect when holding the mediator constant at a predetermined value. In other words, the CDE is the direct exposure effect when we intervene on the mediator value. In our theoretical example the CDE is the difference between $Y(x = 1, m = 3)$ and $Y(x = 0, m = 3)$, which is interpreted as the direct effect of the intervention on hypertension when the participant is forced to adhere to 3 hours of physical activity per week.

Instead of holding the mediator constant at a predetermined value, e.g., 3 hours of physical activity, the mediator can also be set to values that would naturally be observed in the control and intervention group (Holland, 1988; Pearl, 2001). The potential outcomes then take the natural relation between the exposure and mediator into account. This extends the potential outcomes notation to $Y(x, M(x))$, $Y(x, M(x^*))$, $Y(x^*, M(x))$, and $Y(x^*, M(x^*))$, where $M(x)$ represents the participant's naturally observed mediator value when in the intervention group, and $M(x^*)$ represents the participant's naturally observed mediator value when in the control group. These potential outcomes are referred to as nested potential outcomes, as the potential mediator values are nested within the potential outcome values (Pearl, 2001). Whereas in a non-nested potential outcome the m value indicates that the researcher intervenes on the mediator value, in a nested potential outcome $M(x^*)$ and $M(x)$ indicates that the mediator takes on the value that would naturally be observed for a specific participant had that participant been exposed to values x^* and x, respectively.

We illustrate the interpretation of the nested potential outcomes with a theoretical example. Suppose that we would naturally observe 4 hours of physical activity had a participant been in the intervention group, i.e., $M(x = 1) = 4$, and 2 hours had the same participant been in the control group, i.e., $M(x = 0) = 2$. The nested potential outcome $Y(x, M(x))$ then represents the participant's risk of hypertension when in the intervention group, while holding physical activity constant at the value that would naturally have been observed had that participant been in the intervention group, i.e., $Y(x = 1, M(x = 1) = 4)$. The nested potential outcome $Y(x, M(x^*))$ represents the participant's risk of hypertension in the intervention group, while holding physical activity constant at the value that would naturally have been observed in the control group, i.e., $Y(x = 1, M(x = 0) = 2)$. The nested potential outcome $Y(x^*, M(x))$ represents the participant's risk of hypertension when in the control group, while holding physical activity constant at the value that would naturally have been observed had that participant been in the intervention group, i.e., $Y(x = 0, M(x = 1) = 4)$. Finally, the nested potential outcome $Y(x^*, M(x^*))$ represents the participant's risk of hypertension when in the control group, while holding physical activity constant at the value that would naturally have been observed had that participant been in the control group, i.e., $Y(x = 0, M(x = 0) = 2)$.

Based on the differences between the nested potential outcome values, two natural direct effects, two natural indirect effects, and the total effect are defined (Pearl, 2001; Robins & Greenland, 1992). The pure natural direct effect (PNDE) is the direct effect when the participant's mediator is held constant at the value naturally observed had that participant been in the control group, and therefore equals the difference between $Y(x, M(x^*))$ and $Y(x^*, M(x^*))$. The total natural direct effect (TNDE) is the direct effect when the participant's mediator is held constant at the value naturally observed had that participant been in the intervention group, and therefore equals the difference between $Y(x, M(x))$ and $Y(x^*, M(x))$. The pure natural indirect effect (PNIE) is the indirect effect of changing the participant's mediator from that participant's naturally observed value in the treatment group to that participant's naturally observed value in the control group, while the direct exposure effect is held constant at the control group level, and therefore equals the difference between $Y(x^*, M(x))$ and $Y(x^*, M(x^*))$. The total natural indirect effect (TNIE) is the indirect effect of changing the participant's mediator from that participant's naturally observed value in the treatment group to that participant's naturally observed value in the control group, while the direct exposure effect is held constant at the intervention group level, and therefore equals the difference between $Y(x, M(x))$ and $Y(x, M(x^*))$. The total effect (TE) is the total effect of changing the exposure from the intervention group value to the control group value, and therefore equals the difference between $Y(x, M(x))$ and $Y(x^*, M(x^*))$.

The definitions of the natural direct and indirect effects allow for XM interaction, as the PNDE and TNDE are estimated based on different mediator values and the PNIE and TNIE are estimated based on different exposure values. That is, the PNDE is based on the comparison of two potential outcomes for which the mediator is held constant at the participant's value naturally observed in the control group, i.e., $M(x^*)$, while the TNDE is based on the comparison of two potential outcomes for which the mediator is held constant at the participant's value naturally observed in the treatment group, i.e., $M(x)$. The PNIE is based on the comparison of two potential outcomes for which the exposure is held constant at x^*, while the TNIE is based on the comparison of two potential outcomes for which the exposure is held constant at x.

Causal effect estimation

In practice it is not feasible to simultaneously observe two potential outcomes for the same participant (Holland, 1986). It is also not feasible to observe the two nested potential outcomes Y(x, M(x*)) and Y(x*, M(x)) for the same participant, as these nested potential outcomes assume that one participant is simultaneously exposed to the intervention and control condition (Pearl, 2001). Instead of observing the (nested) potential outcomes at the individual level, the (nested) potential outcomes are estimated at the population-average level (Holland, 1988; Pearl, 2001; Robins & Greenland, 1992). The notation of the potential outcomes changes slightly when estimated at the population-average level. For example, at the population-average level the potential outcomes Y(x, m) and Y(x*, m) are denoted as E[Y(x, m)] and E[Y(x*, m)], respectively.

At the population-average level, all participants take on the same mediator value for the non-nested potential outcomes, while the mediator values for the nested potential outcomes differ from participant to participant depending on each participant's naturally observed mediator values under x^* and x (Holland, 1988; Pearl, 2001; Robins & Greenland, 1992). For example, when we estimate the average potential outcome E[Y(x = 0, m = 3)], the mediator is held constant at a value of three for all participants in the population. In contrast, when we estimate the average nested potential outcome E[Y(x = 0, M (x = 0))], we allow the mediator value to vary across participants. For example, for one participant the mediator value naturally observed when in the control group might be two, i.e., M(x = 0) = 2, while for another participant the mediator value naturally observed when in the control group might be one, i.e., M(x = 0) = 1. Since the four nested potential outcomes allow every participant to take on their own mediator values, a distribution of mediator values is observed for participants in the control and intervention group. The four nested potential outcomes are subsequently averaged over the distribution of mediator values observed in either the control or intervention group, rather than estimated conditional on a fixed mediator value.

Four no-confounding assumptions are needed to ensure the causal interpretation of the average (nested) potential outcomes and the average causal effect estimates (Pearl, 2001; Vanderweele & Vansteelandt, 2010):

1. no unmeasured confounding of the exposure-mediator effect, i.e., the a path;
2. no unmeasured confounding of the mediator-outcome effect, i.e., the b path;
3. no unmeasured confounding of the exposure-outcome effect, i.e., the c' path;
4. there are no confounders of the mediator-outcome effect, i.e., b path, that are affected by the exposure.

In the regression-based approach, the average potential outcomes are estimated based on the path coefficients from Equations 2 and 4 (Vanderweele & Vansteelandt, 2010). Both multiple regression analysis and SEM may be used to estimate Equations 2 and 4 (Miočević et al., 2018; Muthén & Asparouhov, 2015). The path coefficients from these equations are subsequently used to estimate the average potential outcomes for a model with a continuous mediator and a binary outcome:

$$E[Y(x^*,m)|z] = i_4 + bm + d_4z \tag{5}$$

$$E[Y(x,m)|z] = i_4 + c' + bm + hm + d_4z \tag{6}$$

$$E[Y(x,M(x^*))|z] = i_4 + c' + d_4z + (b+h)(i_2+d_2z) + 0.5(b+h)^2\sigma^2_{\varepsilon_2} \tag{7}$$

$$E[Y(x^*,M(x^*))|z] = i_4 + d_4z + b(i_2+d_2z) + 0.5b^2\sigma^2{}_{\varepsilon_2} \tag{8}$$

$$E[Y(x,M(x))|z] = i_4 + c' + d_4z + (b+h)(i_2+a+d_2z) + 0.5(b+h)^2\sigma^2_{\varepsilon_2} \tag{9}$$

$$E[Y(x^*,M(x))|z] = i_4 + d_4z + b(i_2+a+d_2z) + 0.5b\sigma^2_{\varepsilon_2} \tag{10}$$

The two potential outcomes E[Y(x*, m)] and E[Y(x, m)] are both estimated conditional on a predetermined mediator value m (Pearl, 2001), hence the multiplication of the b and h coefficients with the predetermined mediator value m. The four nested potential outcomes E[Y(x, M(x*))], E[Y(x*, M (x*))], E[Y(x, M(x*))], and E[Y(x, M(x))] are averaged over the distribution of mediator values in the control and intervention group. The regression-based approach approximates this process by estimating the potential outcomes conditional on the mean mediator value observed in the control and intervention group, i.e., M(x*) and M(x) respectively (Vanderweele & Vansteelandt, 2010). When the mediator follows a normal distribution, the mean mediator values in the control and intervention group provide summary estimates of the mediator distributions in these groups. The intercept in Equation 2, i.e., i_2, represents the mean mediator value in the control group, and the summation of i_2 and a coefficient in Equation 2 represents the mean mediator value in the intervention group. If the outcome is continuous and Equation 4 is estimated with linear regression, the average-nested potential outcomes can be estimated by plugging these path coefficients from Equation 2 into Equation 4 (VanderWeele & Vansteelandt, 2009). However, when the outcome is binary, the average-level potential outcomes are defined on the odds ratio (OR) scale (Vanderweele & Vansteelandt, 2010). When estimating the nested potential outcomes on the OR scale, the mediator variable is assumed to follow a *log-normal* distribution instead of a normal distribution. The mean of a log-normal distribution is a function of the mean *and* variance of the respective normally distributed variable. Therefore, the residual variance from Equation 2 is introduced in the equations of the nested potential outcomes to approximate averaging over the log-normal distribution of the mediator values in the control and intervention group (Vanderweele & Vansteelandt, 2010).

The average causal effects are estimated as the difference between two average (nested) potential outcomes (Vanderweele & Vansteelandt, 2010). Table 1 provides an overview of the logistic-regression-based causal estimators on the log-odds scale for

Table 1. Overview of the regression-based causal estimators on the log-odds scale for models with a continuous mediator, a binary outcome, and XM interaction.

Causal effect	Definition	Estimator
CDE	$E[Y(1, m) - Y(0, m)\|z]$	$c' + hm$
PNDE	$E[Y(1, M(0)) - Y(0, M(0))\|z]$	$c' + h(i_2 + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2}$
TNDE	$E[Y(1, M(1)) - Y(0, M(1))\|z]$	$c' + h(i_2 + a + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2}$
PNIE	$E[Y(0, M(1)) - Y(0, M(0))\|z]$	ab
TNIE	$E[Y(1, M(1)) - Y(1, M(0))\|z]$	$ab + ha$
TE	$E[Y(1, M(1)) - Y(0, M(0))\|z]$	$c' + h(i_2 + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2} + ab + ha$

Abbreviations: CDE, controlled direct effect; PNDE, pure natural direct effect; TNDE, total natural direct effect; PNIE, pure natural indirect effect; TNIE, total natural indirect effect; TE, total effect.

models with a continuous mediator and a binary outcome. The m value in the CDE estimator can be fixed to any mediator value of substantive interest (Pearl, 2001). The CDE estimator is, therefore, conditional on one specific mediator value and provides insight in the average direct effect when for all participants the mediator is held constant at this specific value. In contrast with the conditional CDE estimator, the PNDE and TNDE estimators provide population-average effect estimates. To estimate the PNDE and TNDE the conditional direct effect estimates for each observed mediator value observed in the control and intervention group are averaged over the mediator distribution in the control and intervention group, respectively (Pearl, 2001). The PNIE and TNIE are estimated conditional on the control and intervention group level of the direct effect, respectively, since these are the only two exposure values of substantive interest. The TE estimator on the log-odds scale is algebraically equivalent to the sum of the PNDE and TNIE estimators, and the sum of the TNDE and PNIE estimators (Pearl, 2001; Vanderweele & Vansteelandt, 2010). The effect estimates based on the causal estimators in Table 1 can be exponentiated to yield effect estimates on the OR scale.

When there is confounding, the d_2z term is included in the PNDE, TNDE, and TE estimators, where z represents a confounder value. Two types of confounder adjusted effects can be estimated, i.e., conditional on a specific confounder value and marginal over the confounder distribution. To estimate the PNDE, TNDE, and TE conditional on a specific confounder value, z is set to a value of substantive interest. To estimate the PNDE, TNDE, and TE marginal over the confounder distribution, z is set to the observed mean of the confounder. When there is no confounding, the d_2z term drops out of the PNDE, TNDE, and TE estimators.

Traditional mediation analysis

Traditional mediation analysis defines its effects in terms of linear path coefficients (Baron & Kenny, 1986; MacKinnon & Dwyer, 1993). If Equations 1, 2, and 3 are all estimated with linear regression, the total effect is defined as the c coefficient from Equation 1. The direct effect is defined as the c' coefficient from Equation 3. The indirect effect is defined as the product of the a and b coefficients, i.e., the product-of -coefficients estimator, and as the difference between the c and c' coefficients, i.e., the difference-between-coefficients estimator (MacKinnon, 2008, 2020; MacKinnon & Dwyer, 1993). These indirect effect estimators are algebraically equivalent when based on linear path coefficients (Mackinnon et al., 1995), but not when based on logistic path coefficients (MacKinnon & Dwyer, 1993; MacKinnon et al., 2007). The difference between these indirect effect estimators is caused by the non-collapsibility of the exposure-outcome effect across mediator values (MacKinnon et al., 2007; Rijnhart et al., 2020). The scale of a logistic path coefficient is dependent on the variables in the model (Greenland et al., 1999; Mood, 2010). As a result, logistic path coefficients change when variables are added to the model, even if the added variables are not associated with the other variables in the model. For mediation models this means that c and c' are estimated on different scales, causing the c and c' coefficients to differ in magnitude, even in the absence of mediation (MacKinnon et al., 2007; Rijnhart et al., 2020). As a result, effect estimates based on the difference-in-coefficients estimator do not only represent the indirect effect, but also non-collapsibility. The product-of-coefficients estimator is therefore the preferred indirect effect estimator for models with a continuous mediator and a binary outcome (MacKinnon et al., 2007; Rijnhart et al., 2019, 2020).

The traditional effect definitions do not automatically incorporate XM interaction. As early as 1981, Judd and Kenny advised the examination of XM interaction through the estimation of Equation 4 (Judd & Kenny, 1981). However, only limited guidance was available on the estimation of direct and indirect effects in the presence of a significant XM interaction effect (Judd & Kenny, 1981; MacKinnon, 2008, 2020). A recent study by MacKinnon et al. (2020) showed how recoding of the exposure variable and group-mean centering of the mediator variable in traditional mediation analysis can be used to estimate causal direct and indirect effects for models with a continuous mediator and a continuous outcome. That is, the CDE is estimated as the c' coefficient in Equation 4 when the mediator is centered at the substantive mediator value of interest, i.e., m. The PNDE is estimated as the c' coefficient in Equation 4 when the mediator is centered at the observed mean value in the control group. The TNDE is estimated as the c' coefficient in Equation 4 when the mediator is centered at the observed mean value in the intervention group. The PNIE is estimated as the product of the a and b coefficients when the exposure variable is coded in such a way that the zero value represents the control group. The TNIE is estimated as the product of the a and b coefficients, when the exposure variable is coded in such a way that the zero value represents the intervention group. When the exposure is a continuous variable, the PNIE and TNIE can be estimated by centering the exposure at the values that represent the contrast of substantive interest, i.e., x^* and x.

Analytical comparisons

To assess whether the methodology described by MacKinnon et al. (2020) is also successful for the estimation of causal effects for mediation models with a continuous mediator and binary outcome, we analytically compared the causal estimators for models with a continuous mediator and a continuous outcome with the causal estimators for models with a continuous mediator and a binary outcome. Table 2 shows the results of these analytical comparisons. There are no differences in the causal estimators of the CDE, PNIE, and TNIE for models with a continuous outcome and models with a binary outcome. This implies that traditional mediation analysis can also be used to estimate the CDE, PNIE, and TNIE for models with a continuous mediator and a binary outcome. However, there are differences in the causal estimators of the PNDE, TNDE, and TE for models with a continuous outcome and models with a binary outcome. This implies that traditional mediation analysis cannot be used to estimate the PNDE, TNDE, and TE for mediation models with a continuous mediator and a binary outcome.

The discrepancy in the causal and traditional direct effect estimators can be explained by the fact that traditional mediation analysis provides direct effect estimates that are conditional on a specific mediator value, while the PNDE and TNDE estimates from causal mediation analysis provide direct effect estimates that are averaged over the mediator distributions observed in the control and intervention group, respectively. For mediation models with a continuous mediator and a continuous outcome, the direct effects estimated conditional on group-mean centered mediator variables approximate averaging over the mediator distribution in the control and intervention group. However, this does not work for models with a continuous mediator and a binary outcome, as now the mediator value is assumed to follow a log-normal distribution, rather than a normal distribution. The differences in the traditional and causal direct effect estimates are also reflected in the total effect estimates.

Simulation study

We performed a simulation study to demonstrate when researchers should expect to observe differences between the causal and traditional direct and total effect estimates. The variables were generated from a normal distribution with a mean of 0 and a variance of 1 using the RANNOR function in SAS 9.4 programming language (SAS Institute, 2017). The values of the *a, b, c'*, and *h* coefficients corresponded to zero, small (i.e., 0.14, corresponding to 2% of the variance in the dependent variable), medium (i.e., 0.39, corresponding to 13% of the variance in the dependent variable), and large (i.e., 0.59, corresponding to 26% of the variance in the dependent variable) effect sizes for the continuous variables (Cohen, 1988). The d_2 and d_3 coefficients were set to either 0, representing conditions without confounding, or 0.39, representing conditions with medium confounder effects. The continuous exposure and outcome variables were split at the median to create binary variables. Three values were considered for the residual variance in Equation 2: 1, 4, and 9. Five sample sizes were considered: 50, 100, 200, 500, and 1,000. A total of 7,680 conditions were created with 1,000 replications per condition.

Figure 2 plots the difference between the PNDE estimates (solid lines) and the traditional control-group direct effect estimates (dashed lines) on the log-odds scale for the conditions with confounder effects of zero. The differences between the PNDE and traditional control-group direct effect estimates increased as the residual variance, the *h* coefficient, and the *b* coefficient increased in magnitude. These differences were more pronounced as the sample size decreased. Similar patterns were observed for conditions with medium confounder effects, and for the TNDE and TE (detailed plots can be found in the supplementary materials).

Real-life data example

Data were used from a randomized-controlled trial aiming to prevent unhealthy weight gain among school-aged children through an educational program and changes in the assortment of school canteens (Singh et al., 2006). The study protocol was approved by the Medical Ethical Committee of the VU University Medical Center and informed consent was obtained from all individual study participants. Of the 546 children in this trial, 285 were randomized to the intervention group and 261 to the control group. As displayed in Figure 3, we investigated screen behavior, measured as the average hours of TV watching and daily computer use (mean = 3.847), as a mediator of the relation between the intervention and being overweight (normal weight n = 465; overweight n = 81). Both screen behavior and overweight were measured 8 months after the intervention. The average amount of screen behavior was 4.242 hours per week in the control group and 3.485 hours per

Table 2. Comparison of causal estimators for models with a continuous mediator and a continuous outcome and for models with a continuous mediator and a binary outcome.[a.]

Effect estimate	Continuous outcome	Binary outcome	Difference
CDE	$c' + hm$	$c' + hm$	No difference
PNDE	$c' + h(i_2 + d_2z)$	$c' + h(i_2 + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2}$	$(b + 0.5\ h)h\sigma^2_{\varepsilon_2}$
TNDE	$c' + h(i_2 + a + d_2z)$	$c' + h(i_2 + a + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2}$	$(b + 0.5h)h\sigma^2_{\varepsilon_2}$
PNIE	ab	ab	No difference
TNIE	$ab + ha$	$ab + ha$	No difference
TE	$c' + h(i_2 + d_2z) + ab + ha$	$c' + h(i_2 + d_2z + b\sigma^2_{\varepsilon_2}) + 0.5h^2\sigma^2_{\varepsilon_2} + ab + ha$	$(b + 0.5h)h\sigma^2_{\varepsilon_2}$

Abbreviations: CDE, controlled direct effect; PNDE, pure natural direct effect; TNDE, total natural direct effect; PNIE, pure natural indirect effect; TNIE, total natural indirect effect; TE, total effect.

[a]Equations are derived from VanderWeele and Vansteelandt (2009, 2010)

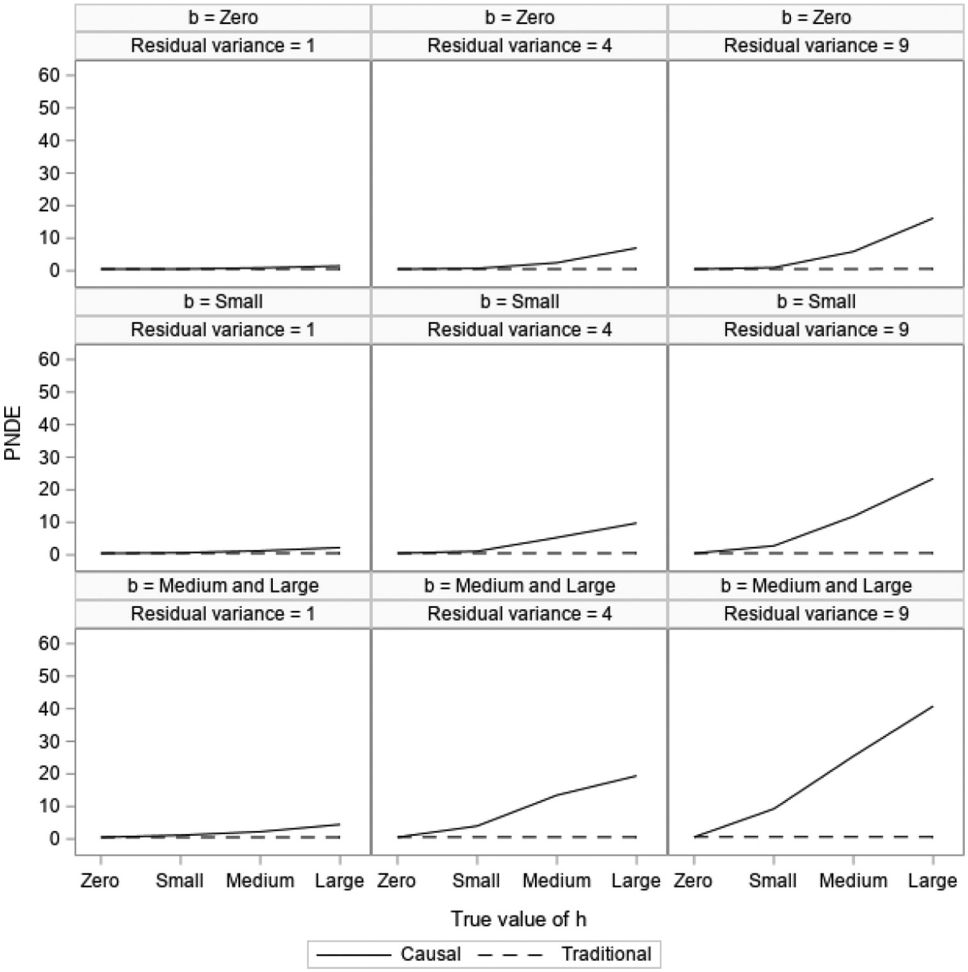

Figure 2. The PNDE estimates (solid lines) and traditional control-group direct effect estimates (dashed lines) as a function of the residual variance in the mediator model, the *h* coefficient, and the *b* coefficient, when the confounder effects are zero.

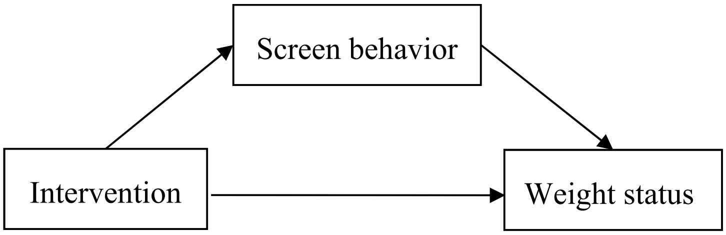

Figure 3. Path diagram of the data example in which screen behavior is hypothesized as a mediator of the relation between an intervention and weight status.

week in the intervention group. Equations 1, 2, and 4 were estimated using the 'gsem' command in Stata version 14.1 (StataCorp, 2016). The causal and traditional mediation effect estimates were accompanied by 95% percentile bootstrap confidence intervals based on 1,000 resamples (Cheung, 2007, MacKinnon et al. 2004). The data example in this paper does not include a confounder. A data example with a confounder can be found in the supplementary materials. Example Stata code for these analyses is also provided in the supplementary materials.

Table 3 shows the estimated path coefficients from Equations 1, 2, and 4. The coefficients in Table 3 were based on the non-centered mediator variable, as the causal estimators require the variables in Equations 2 and 4 to be non-centered. This means that the *c'* coefficient in Table 3 was estimated conditional on a mediator value of zero. To derive the traditional estimates of the CDE, control-group direct effect, and intervention-group direct effect, we also estimated the *c'* coefficient conditional on the grand-mean centered mediator variable (i.e., $c' = -0.341$), on the control-group mean-centered mediator variable (i.e., $c' = -0.276$), and on the intervention-group mean-centered mediator variable (i.e., $c' = -0.340$), respectively. The *b* coefficient in Table 3 was estimated conditional on an exposure value of zero, i.e., the control group. To estimate the traditional intervention-group indirect effect, we also estimated the *b* coefficient based on a model in which the exposure was recoded, yielding the mediator-outcome effect estimate for children in the intervention group (i.e., $b = 0.218$). The *h* coefficient in this example was not statistically

Table 3. Estimated path coefficients for the data example.

Coefficient	Estimate	Standard error	p-value	95% confidence interval
Equation (1)[a]: $logit(Pr(Y = 1\|x)) = i_1 + cX$				
Intercept i_1	−1.569			
c coefficient	−0.365	0.242	0.131	−0.840 to 0.109
Equation (2)[b]: $M = i_2 + aX + \varepsilon_2$[c]				
Intercept i_2	4.242			
a coefficient	−0.758	0.191	<0.001	−1.131 to −0.384
Equation (4)[a] $logit(Pr(Y = 1\|x, m)) = i_4 + c'X + bM + hXM$				
Intercept i_4	−1.804			
c' coefficient	−0.970	0.501	0.053	−1.951 to 0.011
b coefficient	0.054	0.067	0.422	−0.078 to 0.187
h coefficient	0.164	0.101	0.106	−0.035 to 0.362

[a] Equations (1) and (4) were estimated using logistic regression, yielding coefficients on the log-odds scale.
[b] Equation (2) was estimated using linear regression.
[c] The residual term of the mediator model, i.e., $\sigma^2_{\varepsilon_2}$, equaled 4.946.

Table 4. The causal and traditional effect estimates for the real-life data example.

Causal mediation analysis			Traditional mediation analysis		
	Estimate (OR)	95% Confidence Interval		Estimate (OR)	95% Confidence Interval
CDE at M = 3.847	0.711	0.422 to 1.140	Direct at M = 3.847	0.711	0.422 to 1.140
CDE at M = 4.242	0.759	0.447 to 1.227			
CDE at M = 3.485	0.670	0.390 to 1.091			
PNDE	0.847	0.493 to 1.412	Direct (control)	0.759	0.448 to 1.235
TNDE	0.749	0.442 to 1.238	Direct (intervention)	0.670	0.393 to 1.090
PNIE	0.960	0.870 to 1.054	Indirect (control)	0.960	0.870 to 1.054
TNIE	0.848	0.714 to 0.966	Indirect (intervention)	0.848	0.714 to 0.966
TE	0.718	0.426 to 1.161	Total	0.694	0.419 to 1.092

Abbreviations: OR, odds ratio; M, mediator; CDE, controlled direct effect; PNDE, pure natural direct effect; TNDE, total natural direct effect; PNIE, pure natural indirect effect; TNIE, total natural indirect effect; TE, total effect.

significant, but based on its magnitude relative to the b coefficient, the h coefficient can be considered relevant and was therefore included in the estimation of the direct and indirect effects.

Table 4 shows the causal and traditional effect estimates on the odds ratio scale. As expected based on the analytical comparisons, the causal and traditional estimators provided identical CDE, PNIE, and TNIE estimates, but different PNDE, TNDE, and TE estimates.

Since causal and traditional mediation analysis provide identical CDE, PNIE, and TNIE estimates, the interpretations of these estimates are also identical across the two methods. The CDE estimate in Table 4 was estimated conditional on the grand mean mediator value, i.e., 3.847 hours of screen behavior daily. The CDE estimate therefore indicated that children in the intervention group, who spend 3.847 hours daily on screen behavior, on average had a 0.711 times lower odds of being overweight compared to children in the control group who spent a similar amount of time on screen behavior. The PNIE estimate indicated that children in the intervention group on average had a 0.960 times lower odds of being overweight compared to children in the control group, through a decrease in the daily time spent on screen behavior, when the exposure effect was held constant at the control group level. The TNIE estimate indicated that children in the intervention group on average had a 0.848 times lower odds of being compared to children in the control group, through a decrease in the daily time spent on screen behavior, when the exposure effect was held constant at the intervention group level.

Because causal and traditional mediation analysis provide different PNDE, TNDE, and TE estimates, the interpretations of these estimates also differ. The traditional control-group and intervention-group direct effect estimates have the same interpretation as the CDE when estimated conditional on 4.242 and 3.485 hours of screen time respectively. For example, the traditional control-group direct effect of 0.759 indicated that children in the control group, who spent 4.242 hours daily on screen behavior, on average had a 0.759 times lower odds of being overweight compared to children in the control group who spent a similar amount of time on screen behavior. The causal PNDE estimate indicated that children in the intervention group on average had a 0.847 times lower odds of being overweight compared to children in the control group, when for each child the hours of screen behavior was held constant at the value that would naturally have been observed for that child when in the control group. Both the causal and traditional total effect estimates were interpreted as the ratio of the odds of being overweight in the intervention group and the odds of being overweight in the control group.

Discussion

The aim of this paper was to demonstrate the similarities and differences between the causal and traditional estimators for mediation models with a continuous mediator, a binary outcome, and XM interaction. We showed that traditional mediation analysis can be used to estimate the CDE, PNIE, and TNIE from causal mediation analysis, but not the PNDE, TNDE, and TE. The differences between the causal and traditional direct effect estimates occur because of the different types of effects estimated by causal and traditional mediation analysis when based on non-linear models. Whereas the PNDE and TNDE are averaged over the distribution of mediator values in the control and intervention group, respectively, the traditional direct effects are estimated conditional on one specific mediator value. The traditional direct effect estimates are therefore comparable to the CDE estimate in causal mediation analysis. With our analytical comparisons and simulation study we showed that the differences between the causal and traditional direct and total effect estimates increase as a function of the residual variance from Equation 2, the *h* coefficient, and the *b* coefficient. We also showed that the inclusion of confounders in the equations does not affect the similarities and differences between the causal and traditional effect estimates.

Even though causal mediation analysis is gaining in popularity, the uptake of causal mediation analysis for models with binary variables is still relatively low (Vo et al., 2020). A reason for this is the high level of technical details in the literature on causal mediation analysis (Naimi et al., 2017; Vo et al., 2020). This paper aimed to inform researchers about the similarities and differences between the causal estimators and the, still widely applied, traditional estimators for models with a continuous mediator, binary outcome, and XM interaction. It is important that researchers are aware of these similarities and differences, as the differences between the causal and traditional estimators have important implications for the interpretations of their respective effect estimates. The traditional direct effect estimates and the CDE estimate will be of interest when the goal is to inform policy, such as intervention protocols, while the causal direct effects will be of interest when the goal is to unravel causal mechanisms (Pearl, 2001).

For the mediation models described in this paper (estimated based on Equations 1, 2, 3, and 4), causal and traditional mediation analysis pose similar confounding and temporal precedence assumptions. That is, both methods require that the estimated regression equations are adjusted for the confounders of all paths in the mediation model (MacKinnon, 2008, 2020; Vanderweele & Vansteelandt, 2010). Both methods therefore also assume that, conditional on the confounders included in Equations 1, 2, 3, and 4, there are no unmeasured confounders of the paths in the mediation model. In the causal mediation literature, this is typically referred to as the sequential ignorability assumption (Imai et al., 2010), which is a non-parametric assumption, while in traditional mediation analysis the no-unmeasured confounders assumptions are specified within the context of specific parametric models. It is important to note that confounders of the mediator-outcome path always need to be considered, as this path is even observational when the participants are randomized to the exposure (Holland, 1988; MacKinnon, 2008, 2020). Both causal and traditional mediation analysis assume the temporal precedence of the exposure, mediator, and outcome in the mediation model (MacKinnon, 2008, 2020; Pearl, 2012). In other words, changes in the exposure are assumed to precede changes in the mediator, and changes in the mediator are assumed to precede changes in the outcome.

Compared to traditional mediation analysis, causal mediation analysis poses an additional rare outcome assumption for mediation models with a binary outcome. This means that the effect estimates on the OR scale only have a causal interpretation when the outcome prevalence is low, i.e., 10% or lower across both exposure groups, because then the estimated ORs approximate risk ratios (RRs) (Greenland, 1987; Nguyen et al., 2016; Vanderweele & Vansteelandt, 2010). Effect estimates on the RR scale have a population average interpretation, while effect estimates on the OR scale need to be interpreted with reference to the underlying study population (Greenland, 1987). When the outcome prevalence is higher than 10% in both exposure groups, the effect estimates on the OR scale no longer approximate RRs, and can only be used to statistically test the presence of the causal effects (Vanderweele & Vansteelandt, 2010). When the outcome prevalence is higher than 10%, Equation 4 may be estimated with log-linear regression, yielding effect estimates on the RR scale (Valeri & Vanderweele, 2013). The causal estimators in Table 1 can also be applied based on log-linear path coefficients, yielding causal effect estimates on the RR scale.

The implications of XM interaction for mediation analysis are not new (Judd & Kenny, 1981), but the rising popularity of causal mediation analysis raises awareness about the importance of investigating this interaction (Pearl, 2001; Vanderweele & Vansteelandt, 2010). It seems rare to observe a statistically significant XM interaction in practice (Anthopolos et al., 2014; Bauer & Scheim, 2019). An explanation might be that researchers choose mediators that are unlikely to have differential effects across the exposure groups (MacKinnon et al., 2020). Furthermore, significance tests of the *h* coefficient and the mediated interaction, i.e., *ha*, are underpowered in small samples and for small effect sizes (MacKinnon et al., 2020). However, even when there is no statistically significant XM interaction, a substantively relevant XM interaction effect can still substantially affect the direct and indirect effect estimates (Vanderweele & Vansteelandt, 2010). In such a situation it is recommended to include the XM-interaction term in the mediation model.

In this paper, we demonstrated that causal and traditional mediation analysis provide different direct and total effect estimates for models with a continuous mediator, a binary outcome, and XM interaction. We also demonstrated that causal and traditional mediation analysis provide identical indirect effect estimates. Based on the comparisons in this paper we conclude that, for models with a continuous mediator and binary outcome, traditional mediation analysis may only be used to estimate the direct effects when the aim is to determine the direct effect conditional on specific mediator values, and to estimate the indirect effects. Causal mediation analysis is the generally preferred method for mediation analysis, as its average causal direct and indirect effect estimates can be used to unravel causal mechanisms.

Acknowledgments

We want to thank the board and participants of the Dutch Obesity Intervention in Teenagers Study for providing us their data for the real-life data example in this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Institute on Drug Abuse under Grant R37DA09757 and under Grant F31DA043317; by the Netherlands America Commission for Educational Exchange; and by the Prins Bernhard Cultuurfonds.

ORCID

Judith J.M. Rijnhart http://orcid.org/0000-0002-1046-3741
Matthew J. Valente http://orcid.org/0000-0001-9130-2255
David P. MacKinnon http://orcid.org/0000-0003-0866-6010
Jos W.R. Twisk http://orcid.org/0000-0001-9617-1020
Martijn W. Heymans http://orcid.org/0000-0002-3889-0921

Data availability statement

The data that support the findings of this study are available from the corresponding author, JJMR, upon reasonable request.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*, 37–47. https://doi.org/10.2307/2094445

Anthopolos, R., Kaufman, J. S., Messer, L. C., & Miranda, M. L. (2014). Racial residential segregation and preterm birth: Built environment as a mediator. *Epidemiology*, *25*, 397–405. https://doi.org/10.1097/EDE.0000000000000079

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*, 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bauer, G. R., & Scheim, A. I. (2019). The intersectional discrimination index: Development and validation of measures of self-reported enacted and anticipated discrimination for intercategorical analysis. *Social Science & Medicine*. *226*, 225–235. https://doi.org/10.1016/j.socscimed.2018.12.016

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In Stephen L. Morgan, (Ed.) *Handbook of causal analysis for social research* (pp. 301–328). Springer.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140. https://doi.org/10.2307/271084

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Cheung, M. W. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling*, *14*, 227–246. https://doi.org/10.1080/10705510709336745

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, *125*, 761–768. https://doi.org/10.1093/oxfordjournals.aje.a114593

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46. https://www.jstor.org/stable/2676645

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960. https://doi.org/10.1080/01621459.1986.10478354

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, *1988*, i–50. https://doi.org/10.1002/j.2330-8516.1988.tb00270.x

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

Institute, S. A. S. (2017). *Base SAS 9.4*. procedures guide: Statistical procedures.

Judd, C. M., & Kenny, D. A. (1981). Process analysis - estimating mediation in treatment evaluations. *Evaluation Review*, *5*, 602–619. https://doi.org/10.1177/0193841X8100500502

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

MacKinnon, D. P. (2008, 2020). *Introduction to statistical mediation analysis*. Erlbaum.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144–158. https://doi.org/10.1177/0193841X9301700202

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, *4*, 499–513. https://doi.org/10.1177/1740774507083434

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, 99–128. https://doi.org/10.1207/s15327906mbr3901_4 1

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

Mackinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41–62. https://doi.org/10.1207/s15327906mbr3001_3

Miočević, M., Gonzalez, O., Valente, M. J., & MacKinnon, D. P. (2018). A tutorial in Bayesian potential outcomes mediation analysis. *Structural Equation Modeling*, *25*, 121–136. https://doi.org/10.1080/10705511.2017.1342541

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*, 67–82. https://doi.org/10.1093/esr/jcp006

Muthén, B. O., & Asparouhov, T. (2015). Causal effects in mediation modeling: an introduction with applications to latent variables. *Structural Equation Modeling*, *22*, 12–23. https://doi.org/10.1080/10705511.2014.935843

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

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2017). An introduction to g methods. *International Journal of Epidemiology*, *46*, 756–762. https://doi.org/10.1093/ije/dyw323

Nguyen, T. Q., Webb-Vargas, Y., Koning, I. M., & Stuart, E. A. (2016). Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. *Structural Equation Modeling*, *23*, 368–383. https://doi.org/10.1080/10705511.2015.1062730

Pearl, J. (2001). *Direct and indirect effects* [Paper presentation]. Proceedings of the seventeenth conference on uncertainty in artifical intelligence, Seattle, WA, United States of America.

Pearl, J. (2012). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*, 426–436. https://doi.org/10.1007/s11121-011-0270-1

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

Rijnhart, J. J. M., Valente, M. J., & MacKinnon, D. P. (2020). Total effect decomposition in mediation analysis in the presence of non-collapsibility. *Submitted manuscript*.

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

Singh, A. S., Chinapaw, M. J. M., Kremers, S. P. J., Visscher, T. L. S., Brug, J., & van Mechelen, W. (2006). Design of the Dutch Obesity Intervention in Teenagers (NRG-DOiT): Systematic development, implementation and evaluation of a school-based intervention aimed at the prevention of excessive weight gain in adolescents. *Bmc Public Health*, *6*, 304. https://doi.org/10.1186/1471-2458-6-304

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. https://doi.org/10.2307/270723

StataCorp, L. (2016). *STATA software (version 14.1)*.

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*, 137–150. https://doi.org/10.1037/a0031034

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, *20*, 18–26. https://doi.org/10.1097/EDE.0b013e31818f69ce

Vanderweele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*, 1339–1348. https://doi.org/10.1093/aje/kwq332

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*, 457–468. https://doi.org/10.1093/aje/kwq332

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

Wright, S. (1923). The theory of path coefficients a reply to Niles's criticism. *genetics*, *8*, 239–255.