Psychological Methods
2013, Vol. 18, No. 2, 137–150

1082-989X/13/$12.00 DOI: 10.1037/a0031034

Mediation Analysis Allowing for Exposure–Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros

Linda Valeri and Tyler J. VanderWeele
Harvard University

Mediation analysis is a useful and widely employed approach to studies in the field of psychology and in the social and biomedical sciences. The contributions of this article are several-fold. First we seek to bring the developments in mediation analysis for nonlinear models within the counterfactual framework to the psychology audience in an accessible format and compare the sorts of inferences about mediation that are possible in the presence of exposure–mediator interaction when using a counterfactual versus the standard statistical approach. Second, the work by VanderWeele and Vansteelandt (2009, 2010) is extended here to allow for dichotomous mediators and count outcomes. Third, we provide SAS and SPSS macros to implement all of these mediation analysis techniques automatically, and we compare the types of inferences about mediation that are allowed by a variety of software macros.

Keywords: causal inference, direct and indirect effects, mediation analysis, interaction, software macro

Supplemental materials: http://dx.doi.org/10.1037/a0031034.supp

Mediation analysis investigates the mechanisms that underlie an observed relationship between an exposure variable and an outcome variable and examines how they relate to a third intermediate variable, the mediator. Rather than hypothesizing only a direct causal relationship between the independent variable and the dependent variable, a mediational model hypothesizes that the exposure variable causes the mediator variable, which in turn causes the outcome variable. The mediator variable then serves to clarify the nature of the relationship between the exposure and outcome variable (MacKinnon, 2008). For example, it might be of interest to understand whether a rehabilitation program for drug-addicted individuals, with methadone as treatment, leads to increased work activity and whether drug use may mediate some of this effect. In this example, drug use may be a potential mediator of the relationship between the methadone treatment and the work activity outcome since the level of methadone may affect drug use, which may in turn affect work activity.

The use of mediation analysis in psychology and in the social sciences is widespread and has been strongly influenced by the article of Baron and Kenny (1986). More recently, new advances in mediation analysis have been made by using the counterfactual framework (Imai, Keele, & Tingley, 2010; Imai, Keele, Tingley, & Yamamoto, 2010; Pearl, 2001; Robins & Greenland, 1992; VanderWeele & Vansteelandt, 2009, 2010). Using the counterfactual framework has allowed for definitions of direct and indirect effects and for decomposition of a total effect into direct and indirect effects, even in models with interactions and nonlinearities. In many contexts investigators are interested in assessing whether most of the effect is mediated through a particular intermediate or the extent to which it is through other pathways. Decomposition of a total effect into direct and indirect effects accomplishes this goal.

It is then possible to use this counterfactual framework to extend formulae from Baron and Kenny (1986) to allow for mediation analysis even in the presence of exposure mediator interactions. Special cases for mediated effects in the presence of interaction have appeared previously in the literature (e.g., Preacher, Rucker, & Hayes, 2007) but do not give definitions of direct effects such that the total effect decomposes into a direct and indirect effect. However, VanderWeele and Vansteelandt (2009, 2010) derived results for direct and indirect effects for linear and logistic regressions when exposure–mediator interaction is present. In many studies it is unrealistic to assume that the exposure and mediator do not interact in their effects on the outcome. Carrying out mediation analysis incorrectly assuming no interaction may result in invalid inferences. The present article makes a number of important contributions to mediation analysis from both methodological and implementation perspectives. First, we extend work on causal mediation analysis for parametric models with interactions (VanderWeele & Vansteelandt, 2009, 2010) to allow for dichotomous mediators, and not simply continuous mediators as were previously considered. This is done using Pearl's mediation formula (Pearl, 2001), also described outside the context of counterfactuals elsewhere (Huang, Sivaganesan, Succop, & Goodman, 2004). Second, we extend the results to count data. Third, we

This article was published Online First February 4, 2013.

Linda Valeri, Department of Biostatistics, School of Public Health, Harvard University; Tyler J. VanderWeele, Departments of Biostatistics and Epidemiology, School of Public Health, Harvard University.

The research was supported by National Institutes of Health Grants ES017876 and HD060696.

Correspondence concerning this article should be addressed to Linda Valeri, Department of Biostatistics, 677 Huntington Avenue, Boston, MA 20115. E-mail: lvaleri@hsph.harvard.edu

provide SAS and SPSS macros, which give estimates and confidence intervals for direct and indirect effects when interactions between the mediator of interest and the exposure are present, and we compare the types of inference about mediation that are available in a variety of software packages. Finally, we compare and contrast the inferences that are possible about direct and indirect effects in the presence of exposure–mediator interaction, when using the counterfactual framework versus the traditional statistical approach. We consider both continuous and dichotomous variables as outcomes and mediators and allow for general treatment variables. The approach here enriches the contributions of Baron and Kenny and expands the previous software developed by Preacher and Hayes (2004) and Preacher et al. (2007) to allow for effect decomposition of a total effect into direct and indirect effects in the presence of exposure–mediator interaction and other nonlinearities.

The article is organized as follows. The first section discusses the approach to mediation analysis sometimes referred to as the "product method" and made popular by Baron and Kenny (1986). The second section introduces the reader to the counterfactual approach which gives rise to broader definitions of direct and indirect effects and allows one to carry out mediation analysis and effect decomposition when interaction between exposure and mediator is present. In the following section, conditions are given for the identifiability of direct and indirect effects in mediation analysis; these are the conditions needed for the results of statistical procedures to have a causal interpretation. The next section clarifies the relationship between the results on mediation analysis that arise within the counterfactual framework with other popular approaches to mediation analysis. The article continues with instructions for using the software developed (SAS and SPSS) and a description of the output is provided. We conclude by providing an example of mediation analysis performed using the mediation macros.

Classic Regression Approach to Mediation Analysis

The practice of mediation analysis in the field of psychology has been highly influenced by the work of Baron and Kenny (1986). The causal diagram in Figure 1 captures how they conceptualized the role of a mediator variable.

According to Baron and Kenny (1986), the following criteria need to be satisfied for a variable to be considered a mediator: (a) a change in levels of the exposure variable significantly affects the changes in the mediator (i.e., Path from *A* to *M*); (b) there is a significant relationship between the mediator and the outcome (i.e., Path from *M* to *Y*); (c) a change in levels of the exposure variable significantly affects the changes in the outcome (i.e., total effect of *A* on *Y* is significant); and (d) when the previously defined paths are controlled, a previously significant relation between the exposure and outcome is no longer significant, with the strongest demonstration of mediation occurring when the path from the independent variable to the outcome variable is zero.

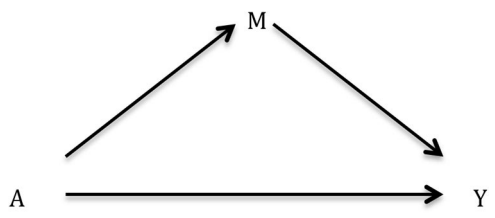

Figure 1. Mediation model in Baron and Kenny (1986) article.

While requirements (a) and (b) have been accepted as correct criteria to identify a potential mediator, requirement (c) has been critiqued by many scholars (MacKinnon, 2008). Consensus has now been reached that the relationship between *A* and *Y* need not be statistically significant for *M* to be a mediator. The reason is that the effect of *A* on *Y* may not be significant when direct and mediated effects have opposite sign. This phenomenon is commonly known as *inconsistent mediation*. Requirement (d) is also not necessary because mediation can be partial or complete. When mediation is complete, after controlling for *M*, the direct path from *A* to *Y* would be zero. When mediation is partial, the path from *A* to *Y* can still be significant, but the effect should be reduced if mediation is indeed present. In the present work we allow for both partial and complete mediation.

In 1986, Baron and Kenny also proposed a parametric approach to estimate and test for mediation. The approach is often simply referred to as the "Baron and Kenny approach"; however, others had proposed it previously (Alwin & Hausen, 1975; Hyman, 1955; Judd & Kenny, 1981; Sobel, 1982), and it is also more generally referred to as the "product method." Let *A* be the treatment, *Y* the outcome, *M* the mediator and *C* additional covariates. For the case of continuous mediator and outcome, consider the following regression models:

$$E[M|a, c] = \beta_0 + \beta_1 a + \beta'_2 c \quad (1)$$

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta'_4 c \quad (2)$$

The original Baron and Kenny approach did not have covariates, but the same general approach applies with covariates (i.e., $\beta'_2 c$ and $\theta'_4 c$ were not included in the original models by the authors; here c is considered a vector and may contain multiple confounders). In particular, Baron and Kenny proposed that the direct effect be assessed by estimating θ_1 and that the indirect effect be assessed by estimating $\theta_2\beta_1$. The direct effect can be conceived of as the treatment effect on the outcome at a fixed level of the mediator variable, which is different from the total effect, which represents simply the overall effect of exposure or treatment on the outcome. The indirect effect can be conceived of as the effect on the outcome of changes of the exposure that operate through mediator levels.

Counterfactual Approach to Mediation Analysis

While the concept of mediation, as defined within psychology and the social sciences, is theoretically appealing, the methods traditionally used to study mediation empirically have important limitations concerning their applicability in models with interactions or nonlinearities (Pearl, 2001; Robins & Greenland, 1992).

Recent contributions in mediation analysis have emphasized the importance of articulating identifiability conditions for a causal interpretation and have extended definitions and results on effect decomposition for direct and indirect effect to settings in which nonlinearities and interactions are present (Pearl, 2001; Robins & Greenland, 1992). This is relevant especially when mediation analysis is implemented in social science contexts where, for

example, the exposure of interest might interact in its effect on the outcome with the mediator.

The approach advocated by Baron and Kenny (1986) is widely applied for mediation analysis and software is available to implement it (Preacher & Hayes, 2004, 2008). However, this method does not fully accommodate settings in which the exposure and the mediator interact in their effects on the outcome. Although special cases for mediated effects in the presence of interaction are available (e.g., Preacher et al., 2007), these do not give definitions of direct effects such that the total effect decomposes into a direct and indirect effect. VanderWeele and Vansteelandt (2009, 2010) showed how the notions of direct and indirect causal effects from causal inference in the counterfactual framework (Robins & Greenland, 1992; Pearl, 2001) can extend the Baron and Kenny formulae for direct and indirect effects to settings in which there is an interaction term between exposure and mediator in the outcome regression.

Suppose we have a continuous outcome and mediator and the mediator regression remains as in Model 1 while the outcome regression is reformulated as

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c. \quad (3)$$

The use of the causal inference approach to mediation analysis gives rise to counterfactual definitions of direct and indirect effects, which were formulated by Pearl (2001) and Greenland and Robins (1992). These effects can be estimated from the regression parameters in Models 1 and 3, provided certain identifiability assumptions (no confounding), described below, hold and models are correctly specified (VanderWeele & Vansteelandt, 2009, 2010). In particular, from Models 1 and 3 what can be defined as the controlled direct effect (CDE), natural direct effect (NDE) and natural indirect effect (NIE) for change in exposure from level a^* to level a, are given by

$$CDE = (\theta_1 + \theta_3 m)(a - a^*)$$
$$NDE = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta' 2c)\}(a - a^*)$$
$$NIE = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*).$$

These expressions generalize those of Baron and Kenny (1986) to allow for interactions between the exposure and the mediator. We describe these effects below. Note that if interaction is not present, so that $\theta_3 = 0$, the controlled direct effect and the natural direct effect are equal to the direct effect obtained using Baron and Kenny approach θ_1 times $(a - a^*)$ and the natural indirect effect is equal to the indirect effect of the Baron and Kenny approach $\theta_2\beta_1$ times $(a - a^*)$.

For a binary exposure, the two exposure levels being compared would be $a^* = 0$ and $a = 1$. The controlled direct effect (*CDE*) expresses how much the outcome would change on average if the mediator were controlled at level m uniformly in the population, but the treatment were changed from level $a^* = 0$ to level $a = 1$. The natural direct effect (*NDE*) expresses how much the outcome would change if the exposure were set at level $a = 1$ versus level $a^* = 0$ but for each individual the mediator were kept at the level it would have taken in the absence of the exposure. The natural indirect effect (*NIE*) expresses how much the outcome would change on average if the exposure were controlled at level $a = 1$, but the mediator were changed from the level it would take if $a^* = 0$ to the level it would take if $a = 1$. The total effect (*TE*) can be defined as how much the outcome would change overall for a change in the exposure from level $a^* = 0$ to level $a = 1$. More formal definitions of these effects explicitly in terms of counterfactuals are given in the Appendix. An important property of the natural indirect effect and the natural direct effect is that the total effect decomposes into the sum of these two effects; this holds even in models with interactions or nonlinearities (Pearl, 2001). The expressions given above involving the coefficients of Models 1 and 3 will be equal to the effects we have just discussed under certain identifiability assumptions given in the next section. These identifiability assumptions allow for a causal interpretation of the direct and indirect effects. These effects are conditional on the level of the covariates C. For continuous outcomes, if C were set at its average level we would obtain marginal effects on the entire population.

While controlled direct effects are often of greater interest in policy evaluation (Pearl, 2001; Robins, 2003), natural direct and indirect effects may be of greater interest in evaluating the action of various mechanisms (Joffe, Small, & Hsu, 2007; Robins, 2003).

Identification

The conditions for a causal interpretation of the direct and indirect effects defined in the previous section can be usefully characterized via causal diagrams. Consider the relation between the variables in Figure 2, which might encompass a wide range of scenarios in mediation analysis. A careful study of this graph is useful in clearly formulating the identifiability assumptions for the direct and indirect causal effects of interest. The variables in the graph are as follows: exposure (A), mediator (M), outcome (Y), covariates ($C = [C_1, C_2]$), which include exposure-outcome confounders (C_1) and mediator–outcome confounders (C_2). All the following comments will still hold if C_1 affects C_2 or if C_2 affects C_1.

Consider the example in which working activity of a drug addicted individual is the outcome of interest (Y). Let the treatment be methadone (A), and the potential mediator be the level of drug use (M). Under this scenario, the investigator may be interested in studying how the effect of the treatment A on the outcome Y is mediated by the level of drug use of an individual (M). In addressing this question of interest, the investigator must think carefully about and try to control for variables that may be exposure-outcome confounders (C_1) or mediator–outcome confounders (C_2). For example, there might be social and biological factors, such as income and hypertension status (C_1), that affect the deci-

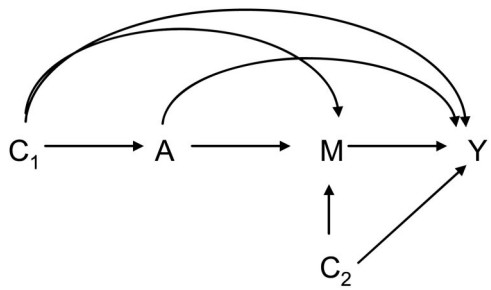

Figure 2. Causal diagram for mediation and confounding. A = exposure; M = mediator; Y = outcome; C_1 and C_2 = covariates.

sion of the level of treatment (A) and the working activity outcome (Y), or other factors, such as neighborhood of residence or alcohol consumption (C_2), which affect both the level of drug use (M) and the working activity outcome (Y).

In order for the effects to have a causal interpretation, control must be made for the confounding variables. In order to ensure identifiability of controlled direct effect, two assumptions are needed: namely, those of (a) no unmeasured confounding of the treatment–outcome relationship and (b) no unmeasured confounding of mediator–outcome relationship. The first of these assumptions would be automatically satisfied if treatment were randomized, but even with randomized treatment the second assumption might not be satisfied. If we refer to the example above, to control for (a) confounding of the treatment–outcome relationship the investigator must adjust for common causes of the treatment and the outcome (e.g., information on income and hypertension status and any other treatment–outcome confounding variable [C_1] in the analysis). To control for (b) mediator–outcome confounding the investigator must adjust for common causes of the mediator and the outcome (e.g., alcohol consumption and neighborhood of residence or any other mediator–outcome confounding variable [C_2]). In practice, both sets of covariates would simply be included in the overall set C for which adjustment is made; the investigator does not need to distinguish in this regression approach the treatment–outcome and the mediator–outcome confounding variables, but the collection of covariates must include both sets for estimates to have a causal interpretation.

The assumptions we have described are for controlled direct effects; the identification of natural direct and indirect effects uses these two assumptions above along with two additional assumptions. In particular, for natural direct and indirect effects there must also be (c) no unmeasured confounding of the treatment-mediator relationship. Control must be made for variables that cause both the level of treatment and the level of the mediator. In the context of our example, hypertension may be a factor which influences the use of treatment as well as the level of drug addiction, and it would need to be controlled for in the analysis. This third assumption, like the first, would also be satisfied automatically if the treatment were randomized. Finally, for the natural direct effect and indirect effects to be identified it also needs to be the case that (d) there is no mediator–outcome confounder that is affected by the treatment (i.e., no arrow from A to C_2 in Figure 2).

It should be noted that assumptions (a), (b), and (c) also require an assumption of temporal ordering. This assumption of temporal ordering is implicitly or explicitly present in various approaches to mediation analysis (Cole & Maxwell, 2003). In particular, the assumption of no unmeasured confounding of the treatment–outcome relationship implicitly assumes that the treatment temporally precedes the outcome. The assumption of no unmeasured confounding of the mediator–outcome relationship implicitly assumes that the mediator precedes temporally the outcome. Finally, the assumption of no unmeasured treatment-mediator confounding implicitly assumes that the exposure must precede the mediator. Formally the no unmeasured confounding assumptions require that associations reflect causal effects; if the temporal ordering assumptions were not satisfied then neither, in general, would the no unmeasured confounding assumptions be satisfied, since associations would not represent causal effects.

In summary, controlled direct effects require (a) no unmeasured treatment–outcome confounding and (b) no unmeasured mediator–outcome confounding. Natural direct and indirect effects require these assumptions and also (c) no unmeasured treatment-mediator confounding and (d) no mediator–outcome confounder affected by treatment. It is important to note that randomizing the treatment is not enough to rule out confounding issues in mediation analysis. This is because randomization of the treatment rules out the problem of treatment–outcome and treatment-mediator confounding but does not guarantee that the assumption of no confounding of mediator–outcome relationship holds. This is because even if the treatment is randomized, the mediator generally will not be. This was pointed out by Judd and Kenny (1981), James and Brett (1984), and MacKinnon (2008) but unfortunately was not mentioned in the popular article by Baron and Kenny (1986). If there are confounders of the mediator–outcome relationship for which control has not been made, then direct and indirect effect estimates will not have a causal interpretation; they will be biased. This is true for the controlled direct effect and natural direct and indirect effects described above and also for the effects described by Baron and Kenny. Investigators should think more carefully about and collect data on and control for such mediator–outcome confounding variables when mediation analysis is of interest. If the investigator is aware that unmeasured confounding may be an issue in his or her study, sensitivity analyses (Imai, Keele, & Tingley, 2010; VanderWeele, 2010) should be implemented to assess the extent to which violations in the assumptions may alter the results.

Binary Outcomes and Binary Mediators

We have thus far considered only the case in which both outcome and mediator are continuous. The results can be extended to cases in which one or both of the mediator and outcome variables are binary.

For example, when the outcome is binary and mediator is continuous the model for the mediator is represented by Model 1, and the outcome can be modeled via a logistic regression:

$$logit\{P(Y=1|a,m,c,)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c. \quad (4)$$

For this case, provided the outcome is relatively rare and assumptions (a)–(d) hold, we can derive controlled direct effects, and natural direct and indirect effects on the odds ratio scale (VanderWeele & Vansteelandt, 2010) as

$$\log\{OR^{CDE}\} = (\theta_1 + \theta_3 m)(a - a^*)$$
$$\log\{OR^{NDE}\} \cong \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2}),$$
$$\log\{OR^{NIE}\} \cong (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*),$$

where σ^2 is the variance of the error term in the regression for the mediator, M, and where the approximations hold to the extent that the outcome Y is rare. With these odds ratios, the total effect is equal to the product of the natural direct and indirect effects (rather than the sum).

When the outcome is not rare, the odds ratio does not approximate the risk ratio anymore. Therefore, the causal effects previously defined will be biased if logistic regression is used to model the outcome. In this case the investigator can estimate the causal effect by running a generalized linear model regression with a

binomial distribution and a log link and the causal effects will have a risk ratio interpretation and the formulas hold exactly.

When the outcome is rare then the direct and indirect effects can be estimated even in case-control designs. The formulas for the effects remain the same; however, the mediator regression is run only for controls, to take into account the case-control design (VanderWeele & Vansteelandt, 2010). This approach works because with a rare outcome Y, the distribution of M among the controls will approximate the distribution in the population.

We also extend the previous results to the cases in which the mediator is a dichotomous variable. The identifiability assumptions do not change but now we would use a logistic model for the mediator:

$$logit\{P(M = 1|a, c)\} = \beta_0 + \beta_1 a + \beta' 2c. \quad (5)$$

Formulas for controlled direct effects and natural direct and indirect effects when the mediator is dichotomous are given in the Appendix. Finally, in the online supplemental materials we show that these formulas for causal effects for binary outcome extend to count variables when modeled with a log link.

The total effect is equal to the sum of the natural direct effect and the natural indirect effect when the outcome is continuous and to the product of the natural direct and indirect effect odds ratios when the outcome is binary. Another measure that has been popular in mediation analysis is the proportion mediated. The proportion mediated can be defined as the ratio of the natural indirect effect to the total effect when the outcome is continuous; the proportion mediated on risk difference scale can also be calculated when the outcome is binary using a transformation of the odds ratios (VanderWeele & Vansteelandt, 2010). Several authors have, however, issued cautions on its use. MacKinnon, Warsi, and Dwyer (1995) warned about the instability of such measure, especially when the association between the exposure and the outcome is weak. Consequently, we have not implemented this measure in the macro; however, investigators can certainly calculate these measures from the output that is provided.

Estimates of the direct and indirect effects of interest are obtained by plugging in the estimated coefficient values into the formulas above, while the standard errors can be obtained using the delta method or by bootstrapping techniques. The reader can refer to the online supplement for derivations of the direct and indirect effects and delta method standard errors. The macro we provide calculates these automatically.

Mediation Analysis for Models With Nonlinearities: A Comparison of Approaches

The counterfactual approach to mediation analysis displays all its power and flexibility when the causal relationships under study are complex and the investigator needs to depart from simple linear models and allow for nonlinearities and interactions. In this section we describe some of the advantages of employing the counterfactual framework to causal mediation that we presented in the previous sections by comparing it to other popular methods to address mediation questions. In this comparison we focus on the so-called product method, the difference method, and the MacArthur approach and address also some developments with regard to "moderated mediation." We first describe traditional statistical approaches, and we then discuss what the counterfactual approach contributes over and above them and comment on the relation between the two.

Traditional Approaches to Mediation Analysis

Modern approaches to mediation have been inspired by the pioneering work of the geneticist Sewall Wright (1920), who developed the path analysis method. Path analysis is now viewed as a special case of structural equation modeling (SEM). Structural equations methods allow for the estimation of direct and indirect effects by modeling covariance and correlation matrices. Most mediation analyses in psychological studies have been conducted using the SEM approach (Baron & Kenny, 1986; Judd & Kenny, 1981; MacKinnon, 2008). Methods to improve estimation and inferential procedures for SEM-based mediation analyses have continued to develop (e.g., MacKinnon, 2008; Sobel, 1982). Structural equation models are often criticized for not adequately addressing issues of confounding/endogeneity in inferring causal relationships. However, if such issues of confounding are adequately addressed by including all relevant confounders (as described in detail above) in the structural equation model then the SEM approach can be a useful tool. The counterfactual approach has placed strong emphasis on identifiability assumptions and conceptual definitions of causal effects, and recently, a number of authors have been using the counterfactual framework to translate the SEM approach within the counterfactual framework[1] (e.g., Imai, Keele, & Tingley, 2010; Jo, 2008; Pearl, 2011; Sobel, 2008; VanderWeele & Vansteelandt, 2009). Among traditional SEM methods, we describe the product method and the difference method. Assume a simple mediation model with no exposure–mediator interaction. The rationale behind the product method is that mediation depends on the extent to which the exposure A changes the mediator M, β_1 from Equation 1, and the extent to which the mediator affects the outcome Y, θ_2 from Equation 2. The product method estimator of the indirect effect is then simply $\theta_2\beta_1$. Sobel (1982) proposed a test for a mediated effect from the product method estimator.

The difference method approach is implemented by fitting an outcome model with the mediator as in Equation 2 and also an outcome model with no mediator:

$$E(Y|a,c = \theta_0^* + \theta_1^* a + \theta_4^{*\prime} c. \quad (6)$$

The value of the mediated or indirect effect is then estimated by taking the difference in the coefficients from Equations 6 and 2, $\theta_1^* - \theta_1$ this corresponds to the reduction in the independent variable effect on the dependent variable when adjustment is made for the mediator. The algebraic equivalence of the indirect effect using the product method, $\theta_2\beta_1$, and the difference method, $\theta_1^* - \theta_1$ was shown by MacKinnon et al. (1995) for ordinary least squares in linear models with continuous outcomes and discussed also in Alwin and Hauser (1975). The product method and difference method diverge, however, when using a binary outcome and

[1] Note that a different way to think about inference with regard to an intermediate within the counterfactual approach framework is to use the concept of "principal strata" (Chiba, 2010; Frangakis & Rubin, 2002; Jo, 2008; Rubin, 2004; VanderWeele, 2008). For a discussion on the use of principal stratification in mediation analysis the interested reader can refer to the commentaries in the *International Journal of Biostatistics* (2011).

logistic regression (MacKinnon & Dwyer, 1993), a point to which we return below. When mediation models include an exposure–mediator interaction term in the outcome regression, this is a particular case or a variant of what is sometimes referred to as "moderated mediation" (James & Brett, 1984; Preacher et al., 2007). Moderated mediation considers the case in which a covariate moderates the mediated effect (cf. MacKinnon, Fairchild, & Fritz, 2007), that is, when the mediated effect varies by the level of a covariate. Such moderated mediation by a covariate was also analyzed by Yzerbyt, Muller, and Judd (2004) and Muller, Yzerbyt, and Judd (2008). When the treatment itself is the moderator for the mediator (as considered in Preacher et al., 2007), the effect of the mediator is allowed to vary by treatment status; or, conceived of another way, the effect of treatment is allowed to vary with (i.e., it interacts with) the mediator. In this setting, Preacher et al. (2007) derived an indirect effect estimator in the context of moderated mediation using the product method.

The MacArthur approach (Kraemer, Kiernan, Essex, & Kupfer, 2008) gives criteria somewhat different than that of Baron and Kenny (1986) in assessing mediation and allows also for assessing exposure–mediator interactions. This approach to mediation analysis is based on the assumption that temporal antecedence and association are necessary (but not sufficient) for a causal relationship. The approach allows for nonlinear relations among variables to qualify as mediation as long as there is a relationship between the exposure A and the mediator M. In particular, it is proposed, first, that if there is no association between A and M, and if M precedes A, and if the $A \times M$ interaction is significant, then the variable M is to be considered as a moderator rather than a mediator. Second, for M to be a mediator for the effect of A on outcome Y, A should precede M and M should precede Y, the variables A and M should be correlated, and either the main effect of M on the outcome or the $A \times M$ interaction should be significant.

Comparison of Traditional Approaches With the Counterfactual Approach When There Are Interactions and Nonlinearities

One of the chief advantages of the counterfactual approach to mediation analysis is that it allows for the decomposition of a total effect into a direct effect and an indirect effect even when there are interactions and nonlinearities. As noted above, some of the statistical approaches, such as that of Preacher et al. (2007) or Kraemer et al. (2008) allow one to assess mediation even when there is exposure–mediator interaction. In fact, the indirect effect of Preacher et al. (2007) for continuous outcome when there is an exposure–mediator interaction is equivalent to the one given here. However, neither Preacher et al. (2007) nor Kraemer et al. (2008) gave a definition of a direct effect in the presence of exposure–mediator interaction such that the sum of the direct and indirect effects equals a total effect. The counterfactual approach provides a general approach to do effect decomposition irrespective of the statistical model and irrespective of possible interactions. The counterfactual approach coincides with the criteria for mediation of the MacArthur approach (Kraemer et al., 2008) but provides actual direct and indirect effect estimates that combine to a total effect and makes clear the no-unmeasured-confounding assumptions needed for a causal interpretation. The counterfactual approach also helps in understanding mediation with binary outcomes and binary mediators. As noted previously, with a binary outcome and logistic regression, the product method and difference method give different results (MacKinnon & Dwyer, 1993). In fact, neither in general will be equal to an estimate of an indirect effect with a causal interpretation (VanderWeele & Vansteelandt, 2010). VanderWeele and Vansteelandt (2010) did, however, show that when there is no exposure–mediator interaction, the product method and difference method will be approximately equivalent when the outcome is rare; and both will then be approximately equal to the natural indirect effect when all the no confounding assumptions hold. The problem with dichotomous outcomes arises when the outcome is common and has to do with the fact that logistic regression uses the odds ratio, which is a measure that is "noncollapsible." Viewed intuitively, the problem occurs because when the outcome is common, the odds ratio does not approximate the risk ratio, and the extent of this lack of approximation can vary with the other covariates in the models. With a common outcome, the odds ratios with the mediator in the model versus without the mediator in the model are thus not directly comparable, and so the difference method essentially breaks down. The risk ratio does not suffer this problem, and it is for this reason that we propose using a log-linear model in this article when the outcome is common. Moreover, this approach also allows us to define and estimate direct and indirect effects when the outcome is binary and an exposure–mediator interaction is present. We have, moreover, using the counterfactual approach in this article, derived analytic expressions for cases when the mediator itself is binary. The counterfactual approach provides a versatile framework to derive direct and indirect effects and to do effect decomposition even with binary variables and nonlinear models.

As is perhaps now clear from this discussion, the traditional statistical approach and the counterfactual approach to mediation will in some settings coincide. For linear models and log-linear models, they will coincide when there is no exposure–mediator interaction; for logistic models, they will coincide when there is no exposure–mediator interaction and when the outcome is rare (VanderWeele & Vansteelandt, 2009, 2010). Thus, before an investigator proceeds with one of the traditional approaches (the product method or difference method) he or she should (a) consider whether control has been made for exposure–outcome confounders, mediator–outcome confounders, and exposure–mediator confounders; (b) check whether there is exposure–mediator interaction; and (c) if the outcome is binary and logistic regression is used, check whether the outcome is rare. If the no-unmeasured-confounding conditions are satisfied, there is no interaction, and the outcome is rare if logistic regression is used, then proceeding with the traditional statistical approaches is fine. If there are exposure–mediator interactions then the approach described in this article, or another counterfactual-based approach, should be used. If the outcome is common, a log-linear model can be used. If there are confounders of the exposure–outcome, mediator–outcome, or exposure–mediator relationship then, to the extent possible, these should be controlled for in the models; otherwise sensitivity analysis techniques (VanderWeele, 2010; Imai, Keele, & Tingley, 2010) can be used.

As a final point of discussion, we note that even in the presence of interaction and nonlinearities, the product method may be useful to test for mediation even if the estimates are not themselves interpretable as estimates of an indirect effect. In other words, to

test for mediation we can test for whether the product of the coefficients is nonzero even if this product is not equal to a causal indirect effect measure. For example, with logistic model with common outcome, the product method estimates will not in general have a causal interpretation as a natural indirect effect. It is nonetheless the case that although the product-method estimator is not itself a measure of an indirect effect, the product method still gives a valid test for the presence of a mediated effect, provided that the identification assumptions hold and that the models are correctly specified (a formal proof of this is given in the online appendix of VanderWeele, 2011). The intuition is that even if the product of the coefficients is not equal to a causal indirect effect, if the product is nonzero then there must be an effect of the exposure on the mediator and an effect of the mediator on the outcome, and under the identification assumptions, this would also imply the presence of a natural indirect effect. Thus, the product-method approach can still be useful in *testing* for mediation even when there are interactions and nonlinearities. For *estimation* and for decomposing a total effect into a direct and indirect effect (arguably the chief advantages of the counterfactual approach), rather than just testing, methods such as those described in this article can be employed.

Description of the SAS Macro

The present macro is designed to enable the investigator to easily implement mediation analysis in the presence of exposure–mediator interaction accounting for different types of outcomes (normal, dichotomous-logistic or dichotomous log-linear, Poisson, negative binomial) and mediators of interest (normal or dichotomous with logit link). The logit link for dichotomous outcomes should only be used if the outcome is rare. If the outcome is not rare the log link can be used (although the outcome model may not always converge). In the case of using the log link the direct and indirect effects are on the risk ratio scale. In particular, these macros for SAS and SPSS provide estimates, and confidence intervals for the direct and indirect effects previously defined. The estimates assume the model assumptions are correct and the identifiability assumptions discussed in the previous section hold.

Basic SAS Macro

The macro has been developed using SAS Version 9.2. In order to implement mediation analysis via the *mediation macro* in SAS the investigator first opens a new SAS session and inputs the data, which has to include the outcome, treatment and mediator variables as well as the covariates to be adjusted for in the model. Macro activation requires then the investigator to save the macro script and input information in the statement

```
%mediation(data= ,yvar= ,avar= ,mvar= ,cvar= ,a0= ,a1= ,m=
,nc= ,yreg= ,mreg= ,interaction=)

run;
```

First one inputs the name of the data set (*data* =), then the name of the outcome variable (*yvar* =), the treatment variable (*avar* =), the mediator variable (*mvar* =), the other covariates, (*cvar* =). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro *dumvar* from MCHP SAS Macros, for example, can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* (*a0* =), the new exposure level *a* (*a1* =), the level of mediator *m* (*m* =) at which the controlled direct effect is to be estimated and the number of covariates to be used (*nc* =). When no covariates are entered, then the user still needs to write the commands *cvar* = and *nc* =, even though both are left blank. The user must also specify which types of regression have to be implemented. In particular, linear, logistic, loglinear, poisson or negbin can be specified (*yreg* =). For the mediator either linear or logistic regressions are allowed (*mreg* =). Finally, the analyst needs to specify whether an exposure–mediator interaction is present (*interaction = true or false*).

The macro provides the following output: first the regression output for outcome and mediator models is provided. The output in the SAS macro is derived from the procedures of proc reg when the variable is continuous, and from proc logistic when the variable is binary. When the outcome is specified as Poisson, negative binomial or log-linear the procedure proc genmod is employed. If the data set contains missing data the macro implements a complete case only analysis. A table with direct and indirect effects together with total effects follows. The effects are reported for the mean level of the covariates *C*. The table contains standard errors, and confidence intervals for each effect.

Other Options in the SAS Macro

The reduced output is the default option. The table will just display controlled direct effect, natural direct effect, natural indirect effect and total effect described above. When the option *output = full* is used, both conditional effects and effects evaluated at the mean covariate levels are shown. When the *output = full* option is chosen, the investigator must enter fixed values for the covariates *C* at which compute conditional effects. The macro statement is as follows:

```
%mediation(data= ,yvar= ,avar= ,mvar= ,cvar= ,a0= ,a1= ,m=
,nc= ,yreg= ,mreg= ,interaction= ,output= ,c=)

run;
```

When *output = full* is added, then, in addition to the controlled direct effect, and the natural direct and indirect effects described above, other effects are also displayed. The natural direct and indirect effects we have been considering are sometimes called the "pure" natural direct effect and the "total" natural indirect effect (Robins & Greenland, 1992). We can also instead consider the "total" natural direct effect and the "pure" natural indirect effect. For binary exposure the total natural direct effect expresses how much the outcome would change on average if the exposure changed from level $a^* = 0$ to level $a = 1$, but the mediator for each individual was fixed at the natural level that would have taken at exposure level $a = 1$. The pure natural indirect effect expresses how much the outcome would change on average if the exposure were controlled at level $a^* = 0$ but the mediator were changed from the natural level it would take if $a^* = 0$ to the level that would have taken at exposure level $a = 1$. These effects are also reported if the user selects *output = full*. If there is no exposure–mediator interaction, the "pure" and "total" natural direct effects will coincide and the "pure" and "total" natural indirect effects will coin-

cide. These different types of effects are essentially different ways of accounting for the exposure–mediator interaction (Robins, 2003; VanderWeele, in press).

The investigator also has the option of implementing mediation analysis when data arise from a case-control design, provided the outcome in the population is rare. To do so the option *casecontrol = true* can be used. In this case the macro statement changes to

```
%mediation(data= ,yvar= ,avar= ,mvar= ,cvar= ,a0= ,a1= ,m=
,nc= ,yreg= ,mreg= ,interaction= ,casecontrol=)

run;
```

Finally, the investigator can choose whether to obtain standard errors and confidence intervals via the delta method or a bootstrapping technique. The default is the delta method. To use bootstrapping the option *boot = true* can be given. In this case the macro will compute 1,000 bootstrap samples from which causal effects are obtained along with their standard errors (*SE*) and percentile confidence intervals ($p\_95\_CI_{lower}$,$p\_95\_CI_{upper}$). If the investigator wishes to use a higher number of bootstrap samples, instead of "true" he or she inputs the number of bootstrap samples desired (e.g., *boot = 5,000* would estimate standard errors and confidence intervals using 5,000 bootstrap samples). The use of bootstrap for standard errors is generally to be preferred if the sample size of the original sample is small as it will lead to more accurate inferences than the delta method (MacKinnon, 2008). However, these issues are less important if the original sample is large and if this is the case the use of delta method standard errors may be preferred because of computational efficiency. (For example, Ananth & VanderWeele, 2011, conducted a mediation analysis using a sample of 26,000,000 individuals and bootstrapping would have been computationally infeasible.) Bootstrapped standard errors may also be preferred when fitting a log-linear model for the outcome, due to convergence issues. When using the bootstrap the macro statement changes to

```
%mediation(data= ,yvar= ,avar= ,mvar= ,cvar= ,a0= ,a1= ,m=
,nc= ,yreg= ,mreg= ,interaction= ,boot=)

run;
```

As noted previously, if the investigator wants to add a categorical variable as covariate, this must be recoded as a series of indicator variables. For example, if a covariate, named catvar, takes four levels (1, 2, 3, 4) we could construct three "dummy" or "indicator" variables, named, for example, ivar2, ivar3, and ivar4, leaving the first value as the reference. The variable ivar2 would take the value 1 for all observations that had catvar = 2, and 0 for all other observations. The variable ivar3 would take the value 1 for all observations that had catvar = 3 and 0 for all other observations, etc. The macro *dumvar* mentioned previously requires the user to list the data set (data=), the categorical variable (e.g., catvar) that needs to be transformed in the input (dvar=). The user needs also to input the prefix of the name of the dummy variables (e.g., ivar) that will be generated (prefix=) and the reference category (drop=). Categorical variables can be both character and numerical using *dumvar*. For example we can run the following:

```
dumvar data = dat dvar="catvar" prefix="ivar" drop="ivar1"
```

Running this command will generate three indicator variables: "ivar2," "ivar3," "ivar4." (The macro can be found at http://mchp-appserv.cpe.umanitoba.ca/concept/_dumvar.sas)

Comparison With Other Macros

Before concluding the section we would like the reader to be aware that a rich set of alternative programs is also available to implement mediation analyses in certain settings. We believe that our macro provides unique features that may be useful to investigators. At the end of this section Table 1 compares our macro to some of the existing and popular software tools. Preacher and Hayes (2004) developed several macros for mediation mainly implementable in SAS, SPSS and Mplus (*indirect, mediate, modmed, medcurve*); Imai et al. (2009) and Imai, Keele, Tingley, & Yamamoto (2010) also developed a macro in R (*mediate*). We also compare the macros to recent procedures that have been developed in Mplus (Muthén, 2012) in part based on the work we present in this article. We compare the macros on the basis of certain features. We check whether they provide both direct and indirect effects and if they allow for nonlinearities such as interactions, and binary or count variables. We also consider whether they accommodate case-control designs and in which software packages they can be implemented.

Our macro, in contrast with that of Preacher and Hayes (2004), (a) allows for effect decomposition into direct and indirect effects even in the presence of exposure–mediator interaction, (b) allows for dichotomous mediators and count outcomes, (c) allows for case-control designs, and (d) gives estimates with a clear interpretation within the counterfactual framework. In contrast with that of Imai, Keele, et al. (2010), our macro (a) provides direct and indirect effects on a ratio scale for dichotomous outcomes, (b) allows for case-control sampling designs, (c) is implemented in SAS and SPSS, which are more commonly employed in the social sciences. Our macro provides similar features to Mplus, which is in part because recent developments in Mplus (Muthén, 2012) were implemented following the results of our article. Our macro, in contrast to Mplus, allows for case-control designs; Mplus, in contrast to our macro, allows for the flexibility to handle ordinal outcomes.

Description of the SPSS Macro

The SPSS macro that we provide, which was developed under the Version 19.0, performs exactly the same tasks described in the previous section for the SAS macro. However, we point out some small differences that the investigator has to take into account when running mediation analysis using SPSS software.

Before invoking the mediation macro the user has to open a new SPSS session and needs to specify the path in which he or she wants to save relevant estimates from the mediator and outcome regressions. This is simply done by running this command:

```
DEFINE !path()"C: "!ENDDEFINE.
```

In between the quotation marks the path is defined, here for example the path "C:\" has been entered. For SPSS users, macro activation requires that the macro script is then saved as a syntax file (the syntax file should be called from the session that has just been opened) and information is input in the following statement:

Table 1
Macro Comparison

Variable	mediation[c]	mediation[b]	modmed[a]	mediate[a]	Sobel[a]	Indirect[a]	medcurve[a]	Mplus[d]
Causal effects								
direct effects	✓	✓	X	✓	X	✓	X	✓
indirect effects	✓	✓	✓	✓	✓	✓	✓	✓
Interaction								
M-A	✓	✓	✓	X	X	X	X	✓
M-C	X	X	✓	X	X	X	X	✓
Type of variables								
continuous M	✓	✓	✓	✓ (+ M & A)	✓	✓ (+ M)	✓	✓
binary M	✓	✓	X	X	X	X	X	✓
continuous Y	✓	✓	✓	✓	✓	✓	✓	✓
binary Y	✓	✓	X	✓	✓	✓	X	✓
count Y	✓	✓	X	X	X	X	X	✓
ordinal Y	X	✓	X	X	X	X	X	✓
Additional covariate	✓	✓	✓	✓	X	✓	✓	✓
Design								
Cross-Sectional	✓	✓	✓	✓	✓	✓	✓	✓
Cohort	✓	✓	✓	✓	✓	✓	✓	✓
Case-Control	✓	✓	X	X	✓	✓	X	X
Standard Errors								
delta method	✓	X	X	X	X	X	X	✓
bootstrap	✓	✓	✓	✓	✓	✓	✓	✓
Software								
SAS	✓	X	X	✓	✓	✓	✓	X
SPSS	✓	X	✓	✓	✓	✓	✓	X
R	X	✓	X	X	X	X	X	X
MPLUS	X	X	✓	X	X	X	X	✓

Note. Check means option is available; X means option is not available. M = mediator; A = exposure; C = covariates; Y = outcome.
[a] Preacher and Hayes (2004). [b] Imai et al. (Imai, Keele, Tingley, & Yamamoto, 2010; Imai, Keele, & Yamamoto, 2009). The Imai et al. macros contain a sensitivity analysis option. Mplus is adding these features in keeping up with the literature, and our macros will eventually have these features as well. [c] Valeri and VanderWeele (2012) (current article). [d] Muthén (2012). A number of the recent developments in Mplus were motivated by the results of the present article.


```
mediation data= / yvar= /avar= /mvar= /cvar= /NC= /a0= /a1=
/m= /yreg= /mreg= /interaction=

[/casecontrol= /Output= /c=]
```

First one inputs the name of the data set (including the path, e.g., data="C:\mydata.sav"), then the name of the outcome variable (*yvar*=), the treatment variable (*avar*=), the mediator variable (*mvar*=), and the other covariates (*cvar*=). Categorical variables need to be coded as a series of dummy variables before being entered as covariates. The macro *dummit* can be used for this purpose. Then the investigator needs to specify the baseline level of the exposure a^* (*a0*=), the new exposure level *a* (*a1*=), the level of mediator *m* at which the controlled direct effect is to be estimated and the number of covariates to be used (*nc*=). When no covariates are entered, then the user still needs to write the command *cvar*= and needs to specify that *nc = 0*. The user must also specify which types of regression have to be implemented. In particular, LINEAR, LOGISTIC, LOGLINEAR, POISSON or NEGBIN can be specified in the option yreg. Logistic links for yreg can be used for rare dichotomous outcomes; otherwise for dichotomous outcomes that are not rare, log links should be used for the outcome regression, and the effects are then given on the risk ratio scale. For the option mreg either LINEAR or LOGISTIC regressions are allowed. If the data set contains missing data the macro implements a complete case only analysis.

Finally, the analyst needs to specify whether an exposure–mediator interaction is present (TRUE or FALSE). As optional inputs, the investigator can use the option *casecontrol = TRUE*, when the data arise from a case-control study, and the outcome is rare. More complete output (described in the previous section) can be obtained using the option *Output = FULL* and entering the values for the covariates at which to compute causal effects conditional on those covariate values (*c*=). In order to enter the covariate values the investigator needs to create a separate data set that contains those values. For example, if two covariates *C* are present in the model and the value at which the investigator wants to fix the first is 4 and the value at which the investigator wants to fix the second is 10, at the beginning of the script the following commands need to be run:

```
Matrix.

compute c = make(1,2,0).

compute c(1,1) = 4.

compute c(1,2) = 10.

SAVE {c(1,:)} /OUTFILE="C:\\c.sav".

end matrix.
```

After having created data set for the covariate values, the user can specify the option *Output = FULL/c="C:\c.sav"* to obtain the more complete output.

If the investigator wishes to obtain bootstrap standard errors, he or she can use the option *boot = true* followed by the number

Table 2
Example: Output of Mediator and Outcome Regressions Ignoring Exposure–Mediator Interaction

Variable	*df*	Estimate	*SE*	*t*	Pr > |t|
		Dependent variable: Satis, parameter standard			
Intercept	1	−0.71479	0.20449	−3.50	0.0017
Therapy	1	0.66788	0.30147	2.22	0.0354
Attrib	1	0.67186	0.16923	3.97	0.0005
		Dependent variable: Attrib, parameter standard			
Intercept	1	−0.35357	0.21837	−1.62	0.1166
Therapy	1	0.81857	0.29902	2.74	0.0106

Note. Therapy = exposure; Attrib = mediator; Satis = outcome.

of observations in the data set (*nobs=*) to compute causal effects and standard errors with 1,000 bootstrap replications (or "boot=n," where n is the desired number of bootstrap samples). Otherwise delta method standard errors is the default option.

As we mentioned in the previous section, if the investigator needs to add a categorical variable as covariate, a series of indicator variables needs to be generated. The SPSS macro *dummit* works very similarly to the SAS macro. In particular, the investigator needs to call the macro followed by three parentheses. In the first parenthesis the number of levels is entered, in the second parenthesis the name of the variable needs to be specified. Finally, in the third parenthesis, the prefix for the new variables is entered. For example if the variable we need to recode is "smoking," which takes levels "never," "past," "current." Then we can run the following macro:

```
dummit (3) (smoking) (smoke)
```

This macro would generate the following variables: "smokedum2," "smokedum3." The category "never" is automatically taken as a reference. More examples can be found at http://www.glennlthompson.com/?p=92

Example

We present in this section an example of using the mediation macro. We implement the analyses on a modified version of the fictitious data set used by Preacher and Hayes (2004) to explain their Sobel macro. The interest lies in the effects of a new cognitive therapy intervention on life satisfaction after retirement. Residents of a retirement home diagnosed as clinically depressed are randomly assigned to receive 10 sessions of a new cognitive therapy (*A = 1*) or 10 sessions of an alternative therapeutic method (*A = 0*). After Session 8, the positivity of the evaluation the residents make for a recent failure experience is assessed (*M*). Finally, at the end of Session 10, the residents are given a questionnaire to measure life satisfaction (*Y*). We can then investigate whether the cognitive therapy's effect on life satisfaction is mediated by the positivity of their attributions of negative experiences.

The new data set that we employ differs with respect to that of Preacher and Hayes (2004) only in the way in which the outcome is simulated. In particular, the exposure and mediator variables are the same but now the outcome is simulated as a normally distributed variable with mean equal to the linear regression estimated with the original data (the coefficients given in the outcome regression in Preacher & Hayes, 2004) plus a new term, the exposure–mediator interaction term, with coefficient equal to $\theta_3 = 0.5$ indicating a positive interaction, and standard deviation equal to the standard error of the residuals obtained from the outcome regression using Preacher and Hayes data (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html).

We first consider the case in which the interaction between the therapy and the attributions of negative experiences is omitted by the investigator.

After having saved the data set and inserted macro script we run the following command:

```
%mediation(data = dat ,yvar = satis ,avar = therapy ,mvar = attrib
,cvar= ,a0 = 0 ,a1 = 1 ,m = 0, nc= ,yreg = linear ,mreg = linear
,interaction = false)

run;
```

The first output provided is the results of the outcome and mediator regressions (see Table 2).

Then the direct effects and indirect effects follow. We give the reduced output, which provides estimates for the controlled direct effect, the natural indirect effect, and the total effect (see Table 3).

We then run the mediation macro with the correctly specified outcome regression model that includes the exposure–mediator interaction term. We type the following command:

Table 3
Example: Direct and Indirect Effects Ignoring Exposure–Mediator Interaction

Obs	Effect	Estimate	*SE*	*p*	CI_95_{lower}	CI_95_{upper}
1	cde = nde	0.66788	0.30147	0.026733	0.07700	1.25877
2	nie	0.54997	0.24403	0.024215	0.07167	1.02827
3	total effect	1.21785	0.33475	0.000275	0.56174	1.87396

Note. CI_95 = 95% confidence interval; cde = controlled direct effect; nde = natural direct effect; nie = natural indirect effect.

Table 4
Example: Output of Outcome Regression Allowing for Exposure–Mediator Interaction

Variable	*df*	Estimate	*SE*	*t*	Pr > |*t*|
Intercept	1	−0.84424	0.1964	−4.30	0.0002
Therapy	1	0.62132	0.27901	2.23	0.0348
Attrib	1	0.30575	0.21913	1.40	0.1747
Int	1	0.74464	0.31251	2.38	0.0248

Note. Therapy = exposure; Attrib = mediator; Satis = outcome; Int = exposure-mediator interaction.

```
%mediation(data = dat ,yvar = satis ,avar = therapy ,mvar = attrib
,cvar= ,a0 = 0 ,a1 = 1 ,m = 0 ,nc= ,yreg = linear ,mreg = linear,
interaction = true)

run;
```

The output from the outcome regression is the following (the mediator regression will be the same; see Table 4). Table 5 contains our estimates for the effects.

We can see how the estimate of the indirect effect is downward biased and is less significant if the interaction term is omitted. Moreover, when the interaction term is correctly added in the model, controlled direct effects and natural direct effects differ. The hypothetical example here has been included to illustrate the software. In an actual application of these methods, one would want to control for variables confounding the relationship between assessment of a negative life experience and overall life satisfaction (i.e., of the mediator–outcome relationship).

Discussion

With the present work we have provided several contributions that will likely be important for research in psychology and in the social and biomedical sciences. First, by using a counterfactual approach for the definition of the causal effects of interest, along with their identifiability conditions, we give the reader some intuitive rules allowing for causal interpretation in mediation analysis. Issues of identification and causal interpretation have often been neglected when using the Baron and Kenny (1986) approach and other traditional approaches; the overview here will hopefully guide researchers in thinking about these questions. Second, we have described how progress in mediation analysis can be made in the case in which exposure–mediator interaction is present, and we have derived new formulas in the Appendix for settings with a binary mediator allowing for exposure–mediator interactions. We have also extended this approach to count outcomes. Third, the investigator who wishes to pursue mediation analysis using regression models will find useful resources in the SAS and SPSS macro that we developed. These macros implement mediation analysis allowing for the presence of exposure–mediator interaction. The macro was created by applying and extending the work on estimation of direct and indirect causal effects of VanderWeele and Vansteelandt (2009, 2010). The current macro also allows for binary and count data as outcomes and provides valid estimation under case-control designs provided the outcome is rare.

Mediation analysis from a counterfactual perspective with exposure–mediator interaction can also be performed in R and STATA using the macro provided by Imai et al. (Imai, Keele, & Tingley, 2010; Imai, Keele, Tingley, & Yamamoto, 2010). Their approach to mediation analysis relies on Monte Carlo methods. However, the connections to product method and other popular methods in mediation analysis are clearer with the regression-based approach we have presented in that we have provided analytic formulae for the direct and indirect effects, and these formulae coincide with the product method when there are no interactions. The approach of Imai et al. (Imai, Keele, & Tingley, 2010; Imai, Keele, Tingley, & Yamamoto, 2010) has the advantage of not needing separate formulas for each combination of the mediator and outcome models (since the calculations are done by simulation). It has the disadvantage of being much more computationally intensive, which may prohibit use in large data sets.

The reader should note that if interactions between exposure or mediator and additional covariates (C) are present, these might need to be included in order to have a correctly specified model. However, the identifiability conditions that we described above under the counterfactual framework are applicable also to these more complex models. An investigator can still pursue mediation analysis with these different models, but new formulas for the direct and indirect effects defined above would have to be derived. The derivations in the online supplemental materials provide a template that could be used to derive these new formulas for the direct and indirect effects and their standard errors in other types of models that may include interactions between covariates and the treatment or the mediator or that include quadratic terms.

Finally we emphasize that the investigator needs to take particular care in controlling for mediator–outcome confounding. The estimates from the product method or difference method or our approach will be

Table 5
Example: Direct and Indirect Effects Allowing for Exposure–Mediator Interaction

Obs	Effect	Estimate	*SE*	*p*	CI_95_{lower}	CI_95_{upper}
1	cde	0.62132	0.27901	0.02596	0.07446	1.16818
2	nde	0.35804	0.34759	0.30298	−0.32323	1.03931
3	nie	0.85981	0.28782	0.00281	0.29568	1.42395
4	total effect	1.21785	0.33407	0.00027	0.56307	1.87263

Note. CI_95 = 95% confidence interval; cde = controlled direct effect; nde = natural direct effect; nie = natural indirect effect.

biased if control is not made for these variables. Mediator–outcome confounding can be present even if the exposure is randomized (since the mediator is not randomized). Unfortunately, this point was not made in the popular Baron and Kenny (1986) article, although it was made by Judd and Kenny (1981) 5 years earlier, and it has now been emphasized and clarified in the causal inference literature and is being emphasized again in psychology. Psychologists, social scientists, and biomedical researchers need to take this assumption seriously if they hope to obtain valid conclusions about direct and indirect effects. If the investigator thinks that unmeasured confounding may be present, sensitivity analysis should be used (Imai, Keele, & Tingley, 2010; VanderWeele, 2010). We hope to automate sensitivity analysis in the macro in future work.

References

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

Ananth, C. V., & VanderWeele, T. J. (2011). Placental abruption and perinatal mortality with preterm delivery as a mediator: Disentangling direct and indirect effects. *American Journal of Epidemiology, 174,* 99–108. doi:10.1093/aje/kwr045

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. doi:10.1037/0022-3514.51.6.1173

Chiba, Y. (2010). Bounds on controlled direct effects under monotonic assumptions about mediators and confounders. *Biomedical Journal, 52,* 628–637. doi:10.1002/bimj.201000051

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112,* 558–577. doi:10.1037/0021-843X.112.4.558

Frangakis, C., & Rubin, D. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21–29.

Hafeman, D. M., & VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology, 22,* 753–764. doi:10.1097/EDE.0b013e3181c311b2

Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine, 23,* 2713–2728. doi:10.1002/sim.1847

Hyman, H. H. (1955). *Survey design and analysis: Principles, cases and procedures*. Glencoe, IL: Free Press.

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

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. In H. D. Vinod (Ed.), *Advances in social science research using R (pp*. 129–154). New York, NY: Springer. doi:10.1007/978-1-4419-1764-5_8

Imai, K., Keele, L., & Yamamoto, T. (2009). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science, 25,* .51–71. doi:10.1214/10-STS321.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69,* 307–321. doi:10.1037/0021-9010.69.2.307

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods, 13,* 314–336. doi:10.1037/a0014207

Joffe, M., Small, D., & Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science, 22,* 74–97. doi:10.1214/088342306000000655

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

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why the criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27, S101–S108*. doi:10.1037/0278-6133.27.2(Suppl.).S101

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.

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

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 5,* 593–614.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of effect measures. *Multivariate Behavioral Research, 30,* 41–62.

Muller, D., Yzerbyt, V., & Judd, C. M. (2008). Adjusting for a mediator in models with two crossed treatment variables. *Organizational Research Methods, 11,* 224–240. doi:10.1177/1094428106296639

Muthén, B. (2012). *Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus*. Manuscript submitted for publication.

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). San Francisco, CA: Morgan Kaufmann.

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,* 717–731. doi:10.3758/BF03206553

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42,* 185–227. doi:10.1080/00273170701341316

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.

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

Robins, J. M., & Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In P. Shrout (Ed.), *Causality and psychopathology: Finding the determinants of disorders and their cures* (pp. 104–158). New York, NY: Oxford University Press.

Rubin, D. (2004). Direct and indirect effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161–160.

Shpitser, I., & VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *International Journal of Biostatistics, 7,* 1–24. doi:10.2202/1557-4679.1297

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290–312). San Francisco, CA: Jossey-Bass. doi:10.2307/270723

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics, 33,* 230–251. doi:10.3102/1076998607307239

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters, 78,* 2957–2962.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology, 21,* 540–551. doi:10.1097/EDE.0b013e3181df191c

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology, 22,* 575–581. doi:10.1097/EDE.0b013e31821db37e

VanderWeele, T. J. (in press). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*.

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. doi:10.1093/aje/kwq332

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences, 6,* 320–332.

Yzerbyt, V., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40,* 424–431. doi:10.1016/j.jesp.2003.10.001

Appendix

Definitions Under the Counterfactual Framework

We let Y_a and M_a denote, respectively, the values of the outcome and mediator that would have been observed had the exposure A been set to level a. We let Y_{am} denote the value of the outcome that would have been observed had the exposure A, and mediator, M, been set to levels a, and m, respectively. The average controlled direct effect comparing exposure level a to a^* and fixing the mediator to level m is defined by $CDE_{a,a^*}(m) = E[Y_{am} - Y_{a^*m}]$. The average natural direct effect is then defined by $NDE_{a,a^*}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}]$. The average natural indirect effect can be defined as $NIE_{a,a^*}(a^*) = E[Y_{aM_a} - Y_{aM_{a^*}}]$, which compares the effect of the mediator at levels M_a and M_{a^*} on the outcome when exposure A is set to a. Controlled direct effects and natural direct and indirect effects within strata of $C = c$ are then defined by $CDE_{a,a^*|c}(m) = E[Y_{am} - Y_{a^*m}|c]$, $NDE_{a,a^*|c}(a^*) = E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c]$, and $NIE_{a,a^*|c}(a^*) = E[Y_{aM_a} - Y_{aM_{a^*}}|c]$ respectively.

For a dichotomous outcome the total effect on the odds ratio scale conditional on $C = c$ is given by $OR^{TE}_{a,a^*|c} = \frac{P(Y_a = 1|c)\{1-P(Y_a = 1|c)\}}{P(Y_{a^*} = 1|c)\{1-P(Y_{a^*} = 1|c)\}}$. The controlled direct effect on the odds ratio scale is given by $OR^{CDE}_{a,a^*|c}(m) = \frac{P(Y_{am} = 1|c)\{1-P(Y_{am} = 1|c)\}}{P(Y_{a^*m} = 1|c)\{1-P(Y_{a^*m} = 1|c)\}}$. The natural direct effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}} = 1|c)\{1-P(Y_{aM_{a^*}} = 1|c)\}}{P(Y_{a^*M_{a^*}} = 1|c)\{1-P(Y_{a^*M_{a^*}} = 1|c)\}}$. The natural indirect effect on the odds ratio scale conditional on $C = c$ is given by $OR^{NIE}_{a,a^*|c}(a) = \frac{P(Y_{aM_a} = 1|c)\{1-P(Y_{aM_a} = 1|c)\}}{P(Y_{aM_{a^*}} = 1|c)\{1-P(Y_{aM_{a^*}} = 1|c)\}}$.

As discussed in the text, identification assumptions (a)–(d) will suffice to identify these direct and indirect effects. If we let $X \perp Y|Z$ denote that X is independent of Y conditional on Z then these four identification assumptions can be expressed formally in terms of counterfactual independence as (a) $Y_{am} \perp A|C$, (b) $Y_{am} \perp M|\{A,C\}$, (c) $M_a \perp A|C$, and (d) $Y_{am} \perp M_{a^*}|C$. As discussed in the text, the intuitive interpretation of these assumptions is that conditional on C there is (a) no unmeasured exposure–outcome confounding, (b) no unmeasured mediator–outcome confounding, (c) no unmeasured exposure–mediator confounding and (d) no mediator–outcome confounder affected by the exposure. Assumptions (a) and (b) suffice to identify controlled direct effects; assumptions (a)–(d) suffice to identify natural direct and indirect effects (Pearl, 2001; VanderWeele & Vansteelandt, 2009). The intuitive interpretation of these assumptions as described in the text follows from the theory of causal diagrams interpreted as nonparametric structural equations (Pearl, 2001). Alternative identification assumptions have also been proposed (Hafeman & VanderWeele, 2011; Imai, Keele, & Tingley, 2010). However, it has been shown that the intuitive graphical interpretations of these alternative assumptions are in fact equivalent (Shpitser & VanderWeele, 2011). Technical examples can be constructed where one set of identification assumptions holds and another does not (see also Robins & Richardson, 2010), but on a causal diagram corresponding to a set of nonparametric structural equations, whenever one set of the assumptions among those in VanderWeele and Vansteelandt (2009); Imai, Keele, and Tingley (2010); and Hafeman and VanderWeele (2011) holds, the others will also.

(*Appendix continues*)

Continuous Outcome and Continuous Mediator

Suppose that both the mediator and the outcome are continuous and that the following models fit the observed data:

$$E[M|a,c] = \beta_0 + \beta_1 a + \beta'_2 c$$
$$E[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c$$

If the covariates C satisfied the no-unmeasured confounding assumptions (a)–(d) above, then the average controlled effect and the average natural direct and indirect effects would be given by (VanderWeele & Vansteelandt, 2009):

$$CDE = (\theta_1 + \theta_3 m)(a - a^*)$$
$$NDE = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta' 2c)\}(a - a^*)$$
$$NIE = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

Continuous Outcome and Binary Mediator

Suppose that the mediator is binary, and the outcome is continuous and that the following models fit the observed data:

$$E[Y|a,m,c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c$$
$$logit\{P(M = 1|a,c)\} = \beta_0 + \beta_1 a + \beta' 2c$$

If the covariates C satisfied the no-unmeasured confounding assumptions (a)–(d) above, then the average controlled effect and the average natural direct and indirect effects would be given by

$$CDE = (\theta_1 + \theta_3 m)(a - a^*)$$
$$NDE = \theta_1(a - a^*) + \{\theta_3(a - a^*)\}\frac{\exp(\beta_0 + \beta_1 a^* + \beta' 2c)}{1 + \exp(\beta_0 + \beta_1 a^* + \beta' 2c)}$$
$$NIE = (\theta_2 + \theta_3 a)\left\{\frac{\exp(\beta_0 + \beta_1 a + \beta' 2c)}{1 + \exp(\beta_0 + \beta_1 a + \beta' 2c)} - \frac{\exp(\beta_0 + \beta_1 a^* + \beta' 2c)}{1 + \exp(\beta_0 + \beta_1 a^* + \beta' 2c)}\right\}$$

Binary Outcome and Continuous Mediator

Suppose that the mediator is continuous, and the outcome is binary and rare and that the following models fit the observed data:

$$logit\{P(Y = 1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c$$
$$E[M|a,c] = \beta_0 + \beta_1 a + \beta' 2c$$

If the covariates C satisfied the no-unmeasured confounding assumptions (a)–(d) above, then the average controlled effect and the average natural direct and indirect effects would be given approximately by

$$\log\{OR^{CDE}\} = (\theta_1 + \theta_3 m)(a - a^*)$$
$$\log\{OR^{NDE}\} \cong \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta'_2 c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})$$
$$\log\{OR^{NIE}\} \cong (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*)$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the data; the expressions are then for direct and indirect effect risk ratios rather than odds ratios.

Binary Outcome and Binary Mediator

Suppose that both the mediator and the outcome are binary and that the following models fit the observed data:

$$logit\{P(Y = 1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta' 4c$$
$$logit\{P(M = 1|a,c)\} = \beta_0 + \beta_1 a + \beta' 2c$$

If the covariates C satisfied the no-unmeasured confounding assumptions (a)–(d) above, then the average controlled effect and the average natural direct and indirect effects would be given by

$$OR^{CDE} = \exp\{(\theta_1 + \theta_3 m)(a - a^*)\}$$
$$OR^{NDE} \cong \frac{\exp(\theta_1 a)\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta'_2 c)\}}{\exp(\theta_1 a^*)\{1 + \exp(\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta'_2 c)\}}$$
$$OR^{NIE} \cong \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta'_2 c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta'_2 c)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta'_2 c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta'_2 c)\}}$$

These expressions apply also if the outcome is not rare and log-linear rather than logistic models are fit to the data; the expressions are then for direct and indirect effect risk ratios rather than odds ratios. As discussed in the online supplement, the expressions for binary outcomes also apply to count outcomes using models with log links. Derivations and standard errors are also given in the online supplement.

Received May 20, 2011
Revision received August 29, 2012
Accepted September 27, 2012 ■