Published in final edited form as:
Psychol Methods. 2008 December ; 13(4): 314–336. doi:10.1037/a0014207.

Causal Inference in Randomized Experiments With Mediational Processes

Booil Jo
Stanford University

Abstract

This article links the structural equation modeling (SEM) approach with the principal stratification (PS) approach, both of which have been widely used to study the role of intermediate posttreatment outcomes in randomized experiments. Despite the potential benefit of such integration, the 2 approaches have been developed in parallel with little interaction. This article proposes the cross-model translation (CMT) approach, in which parameter estimates are translated back and forth between the PS and SEM models. First, without involving any particular identifying assumptions, translation between PS and SEM parameters is carried out on the basis of their close conceptual connection. Monte Carlo simulations are used to further clarify the relation between the 2 approaches under particular identifying assumptions. The study concludes that, under the common goal of causal inference, what makes a practical difference is the choice of identifying assumptions, not the modeling framework itself. The CMT approach provides a common ground in which the PS and SEM approaches can be jointly considered, focusing on their common inferential problems.

Keywords

cross-model translation; mediational process; principal stratification; randomized experiment; structural equation modeling

Embedding intended mediational processes is popular in randomized experiments for several reasons, including its usefulness in testing theories and improving future programs (Brown, 1991; MacKinnon & Dwyer, 1993). In addition, information about mediators might be used for screening individuals at risk into the intervention to make the intervention more effective (Emery, 1991; Pillow, Sandler, Braver, Wolchik, & Gersten, 1991). The common theory behind randomized experiments with mediational processes is that treatments change the status of mediators and that these changes in mediators improve the condition of final outcomes of interest. Naturally, this theory implies two different effects of treatment assignment—the effect of treatment assignment on the outcome through mediators (indirect effect) and the effect of treatment assignment on the outcome without going through (or, conditioning on) mediators (direct effect). In this article, I treat treatment assignment and the treatment actually received as being identical (i.e., full treatment compliance). It is also assumed that outcome information is available for every individual at posttreatment measurement.

Correspondence concerning this article should be addressed to Booil Jo, Department of Psychiatry and Behavioral Sciences, Stanford University, SMC 5795, 401 Quarry Road, Stanford, CA 94305. booil@stanford.edu.

Supplemental materials: http://dx.doi.org/10.1037/a0014207.supp

This article was particularly motivated by the Job Search Intervention Study (JOBS II; Vinokur, Price, & Schul, 1995; Vinokur & Schul, 1997), which is a theory-driven randomized field experiment intended to prevent poor mental health and to promote high-quality reemployment among unemployed workers. One of the hypothesized mediators in JOBS II was sense of mastery that might be triggered by the intervention and have an influence on key outcomes, such as reemployment and depression. However, beyond randomized intervention/prevention studies, the idea of mediation has been widely utilized in randomized experiments to test and generate theories. For example, DeSteno, Valdesolo, and Bartlett (2006) performed a randomized experiment in support of a theory that threatened self-esteem functions as a principal mediator of jealousy. In Milling, Reardon, and Carosella (2006), the mediating role of response expectancies was evaluated in the treatment of pain through a randomized experiment. Meyer and Gellatly (1988) conducted randomized experiments to show that assigned goals affect performance expectancy and performance valence, which, in turn, would affect personal goal and task performance.

To explore the plausibility of a mediational process theory in randomized experiments, we randomly assign individuals to different experimental or treatment conditions and then observe how individuals' mediator status changes given assigned treatment conditions and how these changes affect the ultimate outcome. From this design, strong causal inferences are possible regarding the overall effect of treatment, in which groups randomized to different treatment conditions are compared regardless of differences in mediator status. However, because individuals are not randomly assigned to different mediator values, strong causal inferences are not warranted for the effect of treatment assignment on the outcome through mediators and for the effect of treatment assignment on the outcome conditioning on mediators. Therefore, mediation analyses are regarded as exploratory analyses to generate hypotheses that ideally can be tested in future trials (Kraemer, Wilson, Fairburn, & Agras, 2002). However, the distinction between showing association and causal inferences is often ignored in practice.

While methods to improve estimation and inferential procedures for mediation analyses have continued to develop (e.g., Kraemer, Kiernan, Essex, & Kupfer, 2008; MacKinnon, 2008; MacKinnon & Dwyer, 1993; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Shrout & Bolger, 2002; Sobel, 1982, 1988), various efforts have been made to clarify the underlying assumptions necessary to make causal inferences in the mediation analysis context (e.g., Holland, 1988; Mealli & Rubin, 2003; Pearl, 2001; Robins & Greenland, 1992, [1994]; Rubin, 2004; Sobel, 2008; Ten Have, Elliott, Joffe, Zanutto, & Datto, 2004; Ten Have et al., 2007). The latter movement, led particularly by the potential outcomes approach (e.g., Angrist, Imbens, & Rubin, 1996; Frangakis & Rubin, 2002; Holland, 1986; Neyman, 1923/1990; Robins, 1986; Rosenbaum & Rubin, 1983; Rubin, 1974, 1978, 1980, 2005), improved our general understanding of causal modeling. In the potential outcomes approach, causal effects are defined on the basis of individual-level treatment assignment effect. In general, one individual can be assigned to only one of the all treatment conditions considered, and therefore his or her potential outcome is observed in one treatment condition and unobserved in the other conditions. In the potential outcomes approach, we consider both the observed and unobserved (potential) outcome values in defining causal effects. The advantage of this approach is the high level of clarity in underlying assumptions, which also makes model comparisons and sensitivity analyses more feasible. Unfortunately, however, there has been little communication between the structural equation modeling (SEM) and the potential outcomes approaches, hindering collaboration and understanding between the two approaches

Mediation analyses in the SEM approach have been largely motivated by theory-based experiments. The focus of this approach has been given to testing the fit of the data to analysis models that resemble the mediation theory, with less emphasis being given to clarification of assumptions that support causal interpretations of the analysis results. The key drawback of this approach is that a well-fitting model may or may not prove causal relations because alternative models may show equally good fit. Given that, unless individuals are randomized to both treatment assignment status and mediator status (Spencer, Zanna, & Fong, 2005), mediation analyses conducted in this approach might be misleading when the results are used to make causal claims. Among the potential outcomes community, the SEM approach is generally not regarded as a tool for causal inferences, partly because necessary assumptions to support causal interpretations are not clarified, partly because known assumptions are considered unrealistic, and partly because the SEM approach itself is not well understood. When it comes to the topic of mediating variables, development of the potential outcomes approach methods has been mainly driven by complications in randomized clinical trials, such as noncompliance. These mediating variables commonly dealt with in the potential outcomes approach are somewhat controlled by study designs. For example, treatment receipt behavior can be somewhat controlled by preventing individuals assigned to the control (or placebo) condition from taking the active treatments. In the potential outcomes approach, assumptions to identify causal effects have been developed to reflect these specific situations in clinical trials. Among the SEM community, the potential outcomes approach is, first of all, not well known. Further, frequently used assumptions to identify causal effects in the potential outcomes approach seem very unrealistic in the context of theory-based intervention trials. Given that, there is currently little motivation for the researchers in the SEM community to utilize the potential outcomes approach in their mediation analyses.

The main goal of this article is to improve communication between the SEM and potential outcomes approaches in the context of mediation analysis. In particular, the article focuses on bridging the SEM approach with the principal stratification (PS) approach (Frangakis & Rubin, 2002), which has been developed to specifically deal with causal inference problems associated with mediating variables in the potential outcomes framework. PS refers to classification of individuals on the basis of potential values of mediating variables under all treatment conditions compared. Because the resulting categories (principal strata) are unaffected by treatment, treatment effects conditioning on these categories can be interpreted as causal effects (i.e., as evidence that the difference in the outcome is caused by treatment assignment). In line with the potential outcomes tradition, the primary emphasis is given to clarification of assumptions to identify causal effects in the PS framework. Although the PS and SEM approaches have been developed in different contexts, both approaches explicitly deal with mediators, and therefore the SEM approach is likely to benefit from the methods used in the PS approach to identify causal effects. Clarifying the assumptions that permit causal inferences and checking sensitivity of the results to deviation from these assumptions are likely to lead to better use of data at hand and to improve the design of future trials. Similarly, from the PS perspective, mediation processes dealt with in the SEM approach tend to be more general than mediation processes commonly dealt with in the PS approach, so that the PS approach is likely to benefit from broadening its modeling practice and facilitating development of alternative models and sensitivity analysis methods.

Table 1 summarizes some differences between the SEM and PS approaches that will be further discussed in the following sections. In this article, these differences are viewed as different preferences, not as relative strengths and weaknesses of one approach compared with the other. It is not surprising that the two approaches have different emphases and preferences given their separate development in different contexts with different motivations. However, when we deal with the same context (i.e., same data) with the common motivation of causal inference, the difference between the two approaches becomes minimal, which is in fact the main message of this article.

SEM Approach

Data from intervention trials with embedded mediational processes have been largely analyzed using the SEM approach (Baron & Kenny, 1986; Bollen, 1987; Judd & Kenny, 1981; MacKinnon, 2008; MacKinnon & Dwyer 1993; MacKinnon, Fairchild, & Fritz, 2007), which naturally reflects the mediational process theory. A simple mediational process shown in Figure 1 depicts the key idea common to the conceptual mediation model and the analytical model in the SEM approach. Figure 1 illustrates that treatment assignment (*X*) changes the status of a mediator (*M*), and that this change in the mediator improves the condition of final outcome (*Y*). In this framework, the mediated effect of treatment on the outcome is evaluated in terms of whether the intervention targeted mediators are highly associated with the outcome and whether the intervention successfully improved the mediator status of individuals.

Identification of Causal Effects in the SEM$_{BK}$ Model

In the conventional mediation analysis model, widely known as Baron and Kenny's (1986) model (SEM$_{BK}$ will be used to refer to this model), the total effect of treatment assignment is seen as the sum of the direct and indirect effect. The three key parameters in this model are (1) the effect of treatment assignment on the mediator (*a*), (2) the effect of mediator on the outcome conditional on treatment assignment status (*b*), and (3) the effect of treatment assignment on the outcome conditional on the mediator status (*c*). Ultimately, the indirect effect is obtained as a product of the two effects *a* and *b*. Although the indirect effect is of main interest, this approach does not exclude the possibility that the intervention may directly influence outcomes without going through mediators (direct effect *c*). In many randomized experiments, treatments are likely to include various elements that do not necessarily target the change of mediators and measures of all mediators may not be available. Therefore, in the SEM approach, the total effect of treatment assignment is thought to be the combination of direct and indirect effects of treatment assignment ($T = ab + c$). This relation can be formally expressed by two linear equations.

For individual *i*, the mediator *M* is regressed on treatment assignment *X*:

$$M_i = \alpha_m + aX_i + \varepsilon_{mi}, \tag{1}$$

where the residual ε_{mi} is usually assumed to be normally distributed. The intercept parameter α_m is the mean of *M* in the control condition, and *a* is the effect of *X* on *M* (i.e., mean shift because of the treatment). We compare the SEM and PS approaches focusing on a setting where both *X* and *M* are binary. In that case, α_m can be directly interpreted as one of PS parameters. Centering of *X* does not affect the *a* estimate in this model, although it affects α_m.

For individual *i*, the outcome *Y* is regressed on treatment assignment *X* and the mediator *M*:

$$Y_i = \alpha_y + bM_i + cX_i + \varepsilon_{yi}, \tag{2}$$

where the residual ε_{yi} is usually assumed to be normally distributed. We can interpret *b* as the effect of *M* on *Y* conditioning on *X*, and *c* as the effect of *X* on *Y* conditioning on *M*. The parameter α_y is not of particular interest either in the SEM approach or in the PS approach.

The effect of treatment assignment on mediator status (*a*) can be interpreted as causal on the basis of random assignment of treatment conditions. However, causal interpretation is not readily warranted for the other two components of mediation effects defined in the SEM approach. These effects include the effect of mediator status on the outcome (*b*) and the direct effect of treatment assignment on the outcome that is not mediated by the intended mediator (*c*). Further, direct and indirect effects, which are defined in a unique way in the SEM approach, are not readily identifiable solely relying on observed data. To identify these effects and to interpret them as causal mediation effects in the SEM approach, a particular set of assumptions is necessary.

1. Ignorability of observed mediator status—This is the most critical assumption in the SEM approach. Under this assumption, individuals have the same observed and unobserved individual characteristics regardless of their mediator status *as if* they were randomly assigned to different mediator values. On the basis of this assumption, comparisons in the outcome across individuals with different mediator values lead to the interpretation that differences in mediator status caused differences in the outcome (i.e., *b* effect is interpreted as causal), conditioning on treatment assignment status (*X*). Also, we can interpret the difference in treatment assignment status as having caused the difference in the outcome (i.e., *c* effect is interpreted as causal), conditioning on mediator status (*M*). Without ignorability, the *b* effect is not necessarily causal: Individuals are not randomly assigned to different mediator values, and therefore, when we compare the outcome across individuals with different mediator values, we cannot be sure whether the difference is due to the difference in the mediator or is due to differences in characteristics other than mediator status. For example, in JOBS II, among individuals assigned to the intervention condition, individuals who substantially improved their sense of mastery are likely to have different characteristics from those of individuals who showed little change in their sense of mastery. Therefore, when we compare their outcomes, we cannot exclude the possibility that the difference may be due to characteristics of the individuals other than their mediator status. Causal interpretation of the effect of *X* on the outcome (*c*) conditional on *M* is also problematic because *M* is a posttreatment variable affected by treatment assignment. For example, in JOBS II, individuals who improved their sense of mastery by one point in the intervention program may have different observed and unobserved characteristics from those of individuals who equally improved their sense of mastery in the control condition. In other words, these two groups of individuals may not be comparable even though they have the same mediator status. Randomization of individuals to a second set of treatment conditions designed to affect the mediator will meet this assumptions (Spencer et al., 2005). For further discussions regarding inferential problems associated with analyses conditioning on measured posttreatment variables, see, for example, Rosenbaum (1983).

The ignorability assumption can be somewhat relaxed by allowing heterogeneity among individuals in terms of observed characteristics (i.e., ignorability conditional on observed variables). Under conditional ignorability, we assume that individuals have the same unobserved (unmeasured) characteristics regardless of their mediator status, but they may have different observed characteristics, in particular in terms of pretreatment covariates. Conditional ignorability is weaker than ignorability, although it is still an unverifiable assumption.

2. Constant effect—In conjunction with the ignorability assumption, this assumption plays a key role in identifying unique direct (*c*) and indirect (*ab*) effects of treatment assignment in the SEM approach. Under the assumption of constant effect, the direct effect of treatment assignment on the outcome (*c*) is the same across individuals with different mediator values, which simultaneously implies that the effect of mediator status on the outcome (*b*) is the same across individuals assigned to different treatment conditions. In other words, to uniquely identify the direct and indirect effects, this assumption implies that the interaction between *X* and *M* has no effect on the outcome. This assumption is also referred to as the additivity (Holland, 1988; Sobel, 2008) or the no-interaction (Ten Have et al., 2004) assumption. When both ignorability and constant effect assumptions are satisfied, unique *b* and *c* effects are identified and can be interpreted as causal.

3. Linearity—In conjunction with the ignorability and constant effect assumptions, this assumption makes it possible to uniquely identify direct and indirect effects in the presence of continuous mediators. Under this assumption, the outcome value linearly increases (or decreases) as the mediator value increases (or decreases).

Figure 2 illustrates these key identifying assumptions in the SEM approach. Let us assume a simplified illustration, in which six individuals are randomly assigned either to the treatment or to the control condition. Each individual has one of the three different mediator *M* values (0, 1, 2). Panel A in Figure 2 shows the assumption of ignorability, which is the backbone assumption in the SEM approach. Six individuals with different *M* values are displayed in the same graph. After conditioning on *X*, they are assumed to be comparable on all other measured and unmeasured covariates at baseline. In this case, comparisons between any pairs of these individuals will lead to causal interpretations. In Panel B, in addition to ignorability, a constant effect is also assumed. Now these six individuals follow a pattern that the treatment effect given *M* (distance in *Y* between the treatment and control conditions given *M*) is the same across different values of *M*. This also means that the relation between *Y* and *M* is the same across different treatment conditions. In Panel C, in addition to ignorability and constant effect, linearity is assumed. The relation between *Y* and *M* is linear in both the treatment and control conditions. The condition described in Panel C is necessary in the SEM approach to identify unique direct and indirect effects and to interpret them as causal effects. Panel D demonstrates, under the condition shown in Panel C, how the total effect can be partitioned into the direct (*c*) and indirect (*ab*) effects, which exactly corresponds to the mediation process described in Figure 1. In Panel D in Figure 2, it is assumed that the direct effect of treatment assignment *c* is 1.0 regardless of different *M* values. Also, it is assumed that the effect of the mediator on the outcome (*b*) is 1.0 regardless of different *X* values. If the *a* effect is 1.0, meaning that the value of mediator increases by one unit as individuals are assigned to the treatment condition instead of to the control condition, the total effect would be 2.0 (i.e., $T = c + ab = 1.0 + 1.0 \times 1.0$). If the *a* effect is 2.0, the total effect would be 3.0 (i.e., $T = c + ab = 1.0 + 2.0 \times 1.0$).

Identification of Causal Effects in the SEM$_{\mathrm{Mac}}$ Model

In the SEM$_{BK}$ model, direct and indirect effects are defined assuming that the direct effect of treatment assignment (*c*) is constant across different levels of the mediator. However, under the ignorability condition, a constant effect is not an essential assumption to identify and causally interpret mediation parameters. Without assuming the constant effect assumption, Equation 2 can be rewritten as

$$Y_i = \alpha_y + bM_i + cX_i + dX_iM_i + \varepsilon_{yi}, \tag{3}$$

where *d* represents how much the effect of *M* increases (or decreases) as the value of *X* increases by one unit (i.e., interaction effect). Given that, we can interpret *b* as the effect of *M* on *Y* in the $X = 0$ (control) condition, and $b + d$ in the $X = 1$ (treatment) condition. The parameter *d* can also represent how much the effect of *X* increases (or decreases) as the value of *M* increases by one unit. When *M* is binary, which is focused on in later sections of this article, we can interpret *c* as the effect of *X* on *Y* in the $M = 0$ condition, and $c + d$ as the effect of *X* on *Y* in the $M = 1$ condition. To be able to interpret *b* as the average effect of *M* on *Y*, centering is necessary (Aiken & West, 1991; Kraemer & Blasey, 2006). We do not center *X* and *M* in this study because using the original scores with binary variables permits

SEM parameters to be directly translated to PS parameters. PS parameters can also be directly translated to SEM parameters.

The model described in Equation 3 has been used to check deviation from the constant effect assumption (Judd & Kenny, 1981), and more recently, it has been formally introduced as the MacArthur model (Kraemer et al., 2008). In this article, SEM_{Mac} will be used to refer to this model. By including the interaction effect d, a constant direct effect is no longer assumed in the SEM_{Mac} model. In this framework, mediation effects are conceptualized as the main effect (reflected in b) and the interaction effect (reflected in d) under the condition that X and M are correlated. In both the SEM_{BK} and SEM_{Mac} models, Equation 1 is used to identify the effect of treatment on the mediator (a). However, there is no path analytic component in the SEM_{Mac} model because it does not aim to capture the indirect effect (ab). The conceptualization of mediation effects is quite different in the SEM_{BK} and SEM_{Mac} models. Technically, however, the SEM_{BK} model can be thought of as the SEM_{Mac} model with the restriction that the interaction parameter equals zero. In this article, the two models are categorized into the same category because (a) both models focus on observed mediator status of individuals only under the condition they are assigned to (as opposed to potential values of mediator status under all conditions that are compared), and (b) the two models share ignorability of the mediator status as the central assumption to identify causally interpretable mediation parameters given observed mediator status.

PS Approach

Potential Outcomes

Assume a randomized trial, where treatment assignment (X) has two values (1 = treatment, 0 = control). If an individual i is assigned to the treatment condition, his or her outcome under the treatment condition is observed, but outcome under the control condition is not observed. Let $Y_i(1)$ denote the potential outcome for individual i when assigned to the treatment condition, and $Y_i(0)$ when assigned to the control condition. In this setting, the effect of treatment assignment for individual i can be defined as $Y_i(1) - Y_i(0)$. This definition considers an idealized (potential) situation, in which each individual's outcome is simultaneously observed under all compared conditions in the same context. This way of defining treatment effects on the basis of potential outcomes is often referred to as Rubin's causal model, or more broadly as the potential outcomes approach (Holland, 1986; Neyman, 1923/1990; Rubin, 1978, 1980).

The individual-level treatment effect $Y_i(1) - Y_i(0)$ is interpreted as causal given that the only cause of the difference between $Y_i(1)$ and $Y_i(0)$ is the treatment assignment status. Similarly, the causal effect of treatment assignment can be defined at the average level. Let μ_1 denote the mean potential outcome for the population when assigned to the treatment condition, and μ_0 when assigned to the control condition. The average causal effect of treatment assignment can be defined at the population level as $\mu_1 - \mu_0$. The fundamental problem of causal inference (Holland, 1986) in practice is that we can observe only one of $Y_i(1)$ and $Y_i(0)$. Therefore, the individual level causal effect $Y_i(1) - Y_i(0)$ cannot be identified unless we assume that the effect of treatment assignment is constant across different individuals (Holland, 1986). However, the average causal effect $\mu_1 - \mu_0$ can be identified under the following conditions.

Randomization (ignorability of treatment assignment)—Treatment assignment is random (Holland, 1986; Rubin, 1974, 1978, 1980). Under this assumption, individuals assigned to different treatment conditions have similar pretreatment characteristics. Therefore, we can treat mean outcomes of individuals assigned to different treatment conditions as if they were obtained from the same individuals under different treatment conditions. Given a single randomization, it is not possible to jointly observe mean outcomes of the same individuals under different treatment conditions. However, it is possible to observe the mean outcomes of individuals assigned to different treatment conditions. In other words, under this assumption, the quantities that cannot be observed can be replaced by the quantities that can be observed. This is a key assumption that opens up possibilities for making causal inferences at the average level on the basis of observed data.

Stable unit treatment value assumption (SUTVA)—Potential outcomes for each person are unrelated to the treatment status of other individuals (Rubin, 1978, 1980, 1990). This is another critical assumption that makes identification of causal treatment effects possible. In randomized trials that deal with clusters of individuals who are highly likely to interact with one another, SUTVA is unlikely to hold. For example, let us assume a situation in which a mother and her daughter participate in the same intervention trial. This cluster of participants could be assigned to the same treatment condition or to different conditions. In either case, their outcomes are likely to be contaminated (influenced) by interaction between them. A less cited, but another important, implication of SUTVA is that potential outcomes for each person under a given treatment do not depend on the person who delivers the treatment, that is, there are no hidden variations of the treatment. For example, in a school intervention setting, students who received intervention treatments from their teachers may have better outcomes than students who received treatment from the research staff. For more recent discussions on SUTVA, see, for example, Hong and Raudenbush (2006), Rosenbaum (2007), and Sobel (2006).

Let $\bar{y}_1$ denote the observed sample mean outcome obtained from the treatment condition, and $\bar{y}_0$ from the control condition. If randomization and SUTVA hold, and if the sample is a good representation of the population, $\bar{y}_1$ and $\bar{y}_0$ are approximately unbiased estimates of μ_1 and μ_0. Then, an approximately unbiased estimate of the average causal effect of treatment assignment is identified as

$$\widehat{\gamma}_T = \widehat{\mu}_1 - \widehat{\mu}_0 = \bar{y}_1 - \bar{y}_0. \tag{4}$$

The advantage of approaching treatment effect estimation (although the results are the same) using potential outcomes is that underlying assumptions necessary for causal interpretation are explicitly clarified. The usefulness of the potential outcomes approach becomes more evident as we deal with more complex problems in randomized trials. Mediation analysis with data from randomized trials is a good example in which clarification of underlying assumptions that underlie causal interpretation may benefit from the potential outcomes approach. The disadvantage of the potential outcomes approach is having to deal with quantities that cannot be observed at the same time, which may seem awkward and may require cumbersome notations in expressing causal effects and their underlying assumptions as we deal with more complex problems. Despite its usefulness and decades of history, the potential outcomes approach is still not routinely practiced, particularly in the social, psychological, behavioral, and educational research fields. Some introductory discussions on the potential outcomes approach are provided in Gelman and Hill (2007); West, Biesanz, and Pitts (2000); Morgan and Winship (2007); and Winship and Morgan (1999).

Potential Outcomes and Potential Mediators

The need for handling mediators in the potential outcomes approach arises when we are interested in estimating causal treatment effects that vary across different mediator values. One way to reflect this causal relation among treatment assignment, mediator, and outcome in the potential outcomes approach is to use PS (Frangakis & Rubin, 2002), although other methods are also possible (e.g., Holland, 1988; Sobel, 2008; Ten Have et al., 2007). PS refers to classification of individuals on the basis of potential values of intermediate posttreatment outcomes under every treatment condition compared. The resulting categories (principal strata) provide an informative map for exploring possible sets of assumptions that will support causal inference taking into account the effect of mediators.

Let us extend the setting discussed above and assume that we are now interested in estimating causal treatment effects considering an intermediate outcome M. Assume that M can take only two values (1 = if the mediator status improves after treatment, 0 = otherwise). In this setting, the individual-level causal effect involves potential values of M in addition to potential values of Y. Let $M_i(1)$ denote the potential mediator for individual i when assigned to the treatment condition, and $M_i(0)$ when assigned to the control condition. Because $M_i(0)$ and $M_i(1)$ are potential values, each individual has both of them. However, only one of them is observed under the condition individual i is actually assigned to. The combination of potential values $M_i(0)$ and $M_i(1)$ represents the type of individual i. Types of individuals defined this way are called principal strata. The reason that both values are considered is that the PS approach makes a distinction between the mediator value obtained under the treatment condition and the mediator value obtained under the control condition, which is a key difference of the PS approach from the SEM approach. For example, individuals who improve their mediator status under the treatment condition might have different characteristics from individuals who improve their mediator status under the control condition (e.g., individuals who improve their mediator status despite being assigned to the control condition are likely to be more motivated people). Therefore, the difference in the outcome between the treatment and control condition conditioning on observed mediators (i.e., ignoring under which treatment condition mediator values are obtained) does not necessarily represent a causal effect.

Given two possible treatment assignment status (1 = treatment, 0 = control) and two possible mediator values (1 = improved, 0 = not improved), four possible principal strata are defined as shown in Table 2. A new notation C_i is used to represent which principal stratum each individual belongs to. For convenience, let us label these four types of individuals as never-improvers (n), forward-improvers (f), backward-improvers (b), and always-improvers (a). Never-improvers are individuals whose mediator status does not improve no matter which condition they are assigned to; forward-improvers are individuals whose mediator status improves only if they are assigned to the treatment condition; backward-improvers are individuals whose mediator status improves only if they are assigned to the control condition; and always-improvers are individuals whose mediator status always improves no matter which condition they are assigned to. As the numbers of values M and X can take increases, the number of principal strata also increases. For example, if M can take three values, and X can take two values, the number of possible strata is nine (i.e., 3 × 3 = 9).

The PS approach, illustrated in Table 2, is characterized by turning observed mediators that are affected by treatment assignment into variables that are not affected by treatment assignment. This is possible by using potential values of mediators. Whereas $M_i(0)$ and $M_i(1)$ are affected by treatment assignment (only one of them is observed), the variable that combines the two (i.e., C_i) is unaffected by treatment assignment. Given a mediator M, let $Y_i(M_i(0), 0)$ denote the potential outcome for individual i when assigned to the control condition and $Y_i(M_i(1), 1)$ when assigned to the treatment condition. Because $M_i(0)$ and $M_i(1)$ constitute C_i, we can rewrite these potential outcomes using C_i. That is, $Y_i(M_i(0), 0) = Y_i(C_i, 0)$ and $Y_i(M_i(1), 1) = Y_i(C_i, 1)$. The causal effect for individual i with the principal stratum membership C_i is then $Y_i(C_i, 1) - Y_i(C_i, 0)$. Individual-level potential outcomes and corresponding causal effects given four principal strata are shown in Table 2.

As with the overall average causal effect (γ_T), the causal effect of treatment assignment conditional on principal strata can be defined at the average level. Average causal effects for different strata and involved mean potential outcomes are shown in Table 3. Let μ_{n1} denote the mean potential outcome for never-improvers in the population when assigned to the treatment condition, and μ_{n0} when assigned to the control condition. Similarly, let μ_{f1} denote the mean potential outcome for forward-improvers when assigned to the treatment condition, and μ_{f0} when assigned to the control condition, μ_{b1} for backward-improvers when assigned to the treatment condition, and μ_{b0} when assigned to the control condition, μ_{a1} for always-improvers when assigned to the treatment condition, and μ_{a0} when assigned to the control condition. Because we can only observe one of $M_i(1)$ and $M_i(0)$, C_i is also unobservable, and therefore these mean potential outcomes conditional on principal strata are not directly estimable. The average causal effect of treatment assignment can be defined at the population level as $\mu_{n1} - \mu_{n0}$ (i.e., γ_n) for never-improvers, as $\mu_{f1} - \mu_{f0}$ (i.e., γ_f) for forward-improvers, as $\mu_{b1} - \mu_{b0}$ (i.e., γ_b) for backward-improvers, and as $\mu_{a1} - \mu_{a0}$ (i.e., γ_a) for always-improvers. Because C_i is unobserved, average causal effects conditional on principal strata are not readily identifiable. To identify these effects, more assumptions are necessary in addition to the earlier assumptions of randomization of treatment assignment and SUTVA.

In the PS framework, mediation effects are captured by looking at characteristics such as a high proportion of forward-improvers (π_f), a large treatment assignment effect for forward-improvers (γ_f), a low proportion of backward-improvers (π_b), and no harmful effects of treatment assignment on the outcome for any principal stratum.

Identification of Causal Effects in the PS$_{\text{AIR}}$ Model

In the SEM approach, in which inferences are made on the basis of observed values of mediators, the assumption of ignorability is always necessary for causal interpretation of mediation effects. In contrast, the definition of causal effects in the PS approach does not directly suggest which identifying assumptions should be imposed. Therefore, without knowing the specific assumptions utilized to identify causal effects in the PS approach, comparing the quality of mediation effects defined in the two approaches as causal effects would be unfair. In principle, there are no fixed rules to follow in identifying causal effects in the PS approach. However, the development of the PS approach has been mainly driven by complications in randomized clinical trials such as noncompliance, so that particular assumptions to reflect these specific situations have been more frequently used.

The set of assumptions that will be discussed here was developed in Angrist et al. (1996), in which the purpose was to identify average treatment assignment effects conditioning on individuals' compliance types. These assumptions are applicable in randomized experiments with mediational processes when it is reasonable to assume that (a) there are no harmful effects of treatment assignment on mediators for any principal strata and that (b) the effect of treatment assignment on the outcome is completely mediated through the intended mediators (i.e., there is no direct effect of treatment assignment). In Angrist et al.'s study, these assumptions are labeled as monotonicity and the exclusion restriction. Note that, in this article, these assumptions are applied to handle general mediators. Therefore, full treatment compliance is assumed unless compliance itself is the mediator of interest. It is also assumed that outcome information is available for every individual at posttreatment measurement.

1. Monotonicity—There are no backward-improvers. This implies that individuals' mediator status always shows more (or at least the same) positive change when assigned to the treatment condition than when assigned to the control condition, which excludes the possibility of having individuals whose mediator status improves less when assigned to the treatment condition (i.e., backward-improvers). Monotonicity is often a critical assumption that supports the identifiability of causal inference models in the PS approach. Admitting the existence of backward-improvers is the same as admitting that the treatment may have harmful (iatrogenic) effects on the mediator status of some individuals. That is, some individuals may do better if the treatment is not provided, which is not a desirable situation. Monotonicity is considered a benign assumption in randomized clinical trials where we can control the status of the posttreatment variables by study design. For example, compliance can be controlled to some extent by disallowing individuals from receiving treatments other than those they are assigned to receive. However, in more general situations, such as in JOBS II, in which sense of mastery was a targeted mediator, we cannot prevent people from improving (or not improving) their mastery status, no matter which condition they are assigned to. Therefore, depending on the research context, we might not be able to exclude the possibility of having backward-improvers.

2. Exclusion restriction—For never-improvers and always-improvers, there is no effect of treatment assignment on the outcome. This implies that the effect of treatment assignment on the outcome is completely mediated through the intended mediator (i.e., no direct effect of treatment assignment). The exclusion restriction assumption is more likely to hold in masked experiments because it is hard for study participants to tell which treatment condition they are in. However, in randomized field trials, such as JOBS II, masking is not a realistic option. Without successful masking, treatment assignment itself may have some effect on the outcome (e.g., being assigned to the treatment condition may have a positive or a negative effect on outcomes, and in particular, on psychological outcomes), resulting in some direct effect that is not mediated by the intended mediators. Treatment or intervention programs, such as JOBS II, also typically target multiple mediators. If all of the mediators are not included in the analysis, which is quite common in the potential outcomes approach because handling multiple mediators is extremely difficult, the effect of treatment through mediators that are excluded from the analysis will take the form of a direct effect. Further, treatment may affect unintended (or unobserved) mediators, creating another source of a direct effect. Given that, the exclusion restriction might be violated frequently in practice.

Under these two assumptions, identification of the average causal effect of treatment assignment defined in Table 3 is straightforward. Let us call this particular PS model the PS$_{AIR}$ (Angrist, Imbens, & Rubin) model. On the basis of the mean potential quantities in Table 3, these two assumptions can be unfolded into three specific restrictions. That is, as a result of monotonicity, $\pi_b = 0$ (no backward-improvers). As a result of the exclusion restriction, $\mu_{n1} - \mu_{n0} = 0$, and $\mu_{a1} - \mu_{a0} = 0$ (no effect of treatment on outcome for never-improvers and always-improvers).

Consider first the model in which the two identifying assumptions above are not imposed. First, individuals in the four principal strata form the population of interest. That is,

$$\pi_n+\pi_f+\pi_b+\pi_a=1. \tag{5}$$

Second, the observable mean potential outcome under each treatment condition can be expressed in terms of the mean potential outcomes and proportions of principal strata as

$$\mu_0=\pi_n\mu_{n0}+\pi_f\mu_{f0}+\pi_b\mu_{b0}+\pi_a\mu_{a0}, \tag{6}$$

$$\mu_1=\pi_n\mu_{n1}+\pi_f\mu_{f1}+\pi_b\mu_{b1}+\pi_a\mu_{a1}. \tag{7}$$

I now impose monotonicity, $\pi_b = 0$. Then, from Equation 5, π_f is derived as

$$\pi_f=1-\pi_n-\pi_a-\pi_b=1-\pi_n-\pi_a, \tag{8}$$

where π_f is identifiable because π_n and π_a have corresponding sample statistics. That is, if monotonicity holds, we can distinguish always-improvers from other types in the control condition, because always-improvers are the only type of individuals who would improve their mediator status when assigned to the control condition (i.e., $M_i(0) = 1$). In the treatment condition, we can distinguish never-improvers from other types, because never-improvers are the only type of individuals who would not improve their mediator status when assigned to the treatment condition (i.e., $M_i(1) = 0$). Therefore, at the average level, we observe π_n in the treatment condition and π_a in the control condition. Because treatment assignment is randomized, we assume that the proportions of three strata are the same regardless of treatment conditions.

From Equations 6 and 7 and imposing monotonicity, $\pi_b = 0$, γ_f can be derived as

$$\gamma_f=\mu_{f1}-\mu_{f0}=\frac{\mu_1-\mu_0+\pi_n\left(\mu_{n1}-\mu_{n0}\right)+\pi_a\left(\mu_{a1}-\mu_{a0}\right)}{\pi_f}. \tag{9}$$

I now impose the exclusion restriction, $\mu_{n1} = \mu_{n0}$, and $\mu_{a1} = \mu_{a0}$. Then, γ_f in Equation 9 can be rewritten as

$$\gamma_f=\frac{\mu_1-\mu_0}{\pi_f}, \tag{10}$$

where γ_f is identifiable, given that μ_1 and μ_0 have corresponding sample statistics, and π_f is identifiable as shown in Equation 8. Also, note that the numerator $\mu_1 - \mu_0$ is the overall (total) effect of treatment assignment (γ_T) as defined in Equation 4.

Identification of Causal Effects in the PS$_{Ig}$ Model

In the SEM approach, ignorability of mediator status is an essential assumption that cannot be separated from causally interpretable mediation parameters. One clear advantage of the PS approach is that the true (or potential) status of parameters of interest before applying any identifying assumptions can be explicitly presented, as shown in Tables 2 and 3. In other words, in the PS approach, there are no fixed identifying assumptions that must be present to define causal effect parameters. In the PS$_{AIR}$ model, for example, monotonicity and the exclusion restriction are used to identify causal effect parameters. However, these two assumptions can be replaced with other alternative assumptions (especially if they are more plausible). This open structure is possible due to the fact that the PS approach considers potential mediator status instead of observed values of the mediator. One interesting alternative identifying assumption that will be considered to identify PS parameters in this article is ignorability, which is the backbone assumption in the SEM approach.

Recall that ignorability in the SEM approach means that individuals have the same unmeasured (unobserved) characteristics regardless of their mediator status. On the basis of this assumption, comparisons in the outcome across individuals with different mediator values lead to the interpretation that differences in mediator status (*M*) caused differences in the outcome conditioning on treatment assignment status (*X*). From the PS perspective, this assumption can be interpreted that individuals have the same potential outcome conditional on the observed mediator (*M*) and the treatment assignment condition (*X*). In other words, individuals with the same *M* and *X* values have the same potential outcome values regardless of their principal stratum membership.

Let us apply ignorability to four principal strata defined in Table 3 to see which strata contribute to each possible observable outcome. As a direct result of ignorability, $\mu_{n0} = \mu_{f0} = \mu_{0,0}$, where $\mu_{0,0}$ is the conditional mean potential outcome of individuals whose mediator status does not improve under the control condition. Similarly, $\mu_{b0} = \mu_{a0} = \mu_{1,0}$, where $\mu_{1,0}$ is the conditional mean potential outcome of individuals whose mediator status improves under the control condition, and $\mu_{n1} = \mu_{b1} = \mu_{0,1}$, where $\mu_{0,1}$ is the conditional mean potential outcome of individuals whose mediator status does not improve under the treatment condition. Finally, $\mu_{f1} = \mu_{a1} = \mu_{1,1}$, where $\mu_{1,1}$ is the conditional mean potential outcome of individuals whose mediator status improves under the treatment condition. Table 4 summarizes the mean potential outcomes and average causal effects redefined under ignorability.

Under ignorability, unidentifiable mean potential outcomes and average causal effects conditional on principal stratum membership can be replaced by identifiable quantities. That is, whereas the eight mean potential outcomes μ_{n0}, μ_{n1}, μ_{f0}, μ_{f1}, μ_{b0}, $\mu_{b1,}$ μ_{a0}, and μ_{a1} are not identifiable, the four mean potential outcomes $\mu_{0,0,}$ $\mu_{1,0}$, $\mu_{0,1}$, and $\mu_{1,1}$ are. The observed means $\bar{y}_{0,0}$, $\bar{y}_{1,0}$, $\bar{y}_{0,1}$, and $\bar{y}_{1,1}$ are approximately unbiased estimates of $\mu_{0,0}$, $\mu_{1,0}$, $\mu_{0,1}$, and $\mu_{1,1}$, where $\bar{y}_{0,0}$ denotes the observed sample mean outcome of individuals whose mediator status did not improve when assigned to the control condition, $\bar{y}_{1,0}$ denotes the observed sample mean outcome of individuals whose mediator status improved when assigned to the control condition, $\bar{y}_{0,1}$ denotes the observed sample mean outcome of individuals whose mediator status did not improve when assigned to the treatment condition, and $\bar{y}_{1,1}$ denotes the observed sample mean outcome of individuals whose mediator status improved when assigned to the treatment condition. Of importance, because observed outcomes replace potential outcomes, the distinction between the SEM and PS approaches no longer exists under the assumption of ignorability.

Given that both treatment assignment *X* and the mediator *M* are binary, estimates of the four means $\bar{y}_{0,0}$, $\bar{y}_{1,0}$, $\bar{y}_{0,1}$, and $\bar{y}_{1,1}$ can be expressed in terms of the SEM parameter estimates from Equation 3, or from Equation 2 if the constant effect assumption holds, as

$$\widehat{y}_{0,0}=\widehat{\alpha}_y, \tag{11}$$

$$\widehat{y}_{1,0}=\widehat{\alpha}_y+\widehat{b}, \tag{12}$$

$$\widehat{y}_{0,1}=\widehat{\alpha}_y+\widehat{c}, \tag{13}$$

$$\widehat{y}_{1,1}=\widehat{\alpha}_y+\widehat{b}+\widehat{c}+\widehat{d}, \tag{14}$$

where $\widehat{\alpha}_y$, $\widehat{b}$, $\widehat{c}$, and $\widehat{d}$, can be obtained from a simple ordinary least squares regression analysis or from more efficient large sample estimation methods, such as a maximum likelihood (ML) estimation, although the results will be practically identical given the simple structure of the model.

Then, approximately unbiased estimates of the average causal effects defined in Table 4 are identified as

$$\widehat{\gamma}_n=\widehat{\mu}_{0,1}-\widehat{\mu}_{0,0}=\widehat{y}_{0,1}-\widehat{y}_{0,0}=\widehat{c}, \tag{15}$$

$$\widehat{\gamma}_f=\widehat{\mu}_{1,1}-\widehat{\mu}_{0,0}=\widehat{y}_{1,1}-\widehat{y}_{0,0}=\widehat{b}+\widehat{c}+\widehat{d}, \tag{16}$$

$$\widehat{\gamma}_b=\widehat{\mu}_{0,1}-\widehat{\mu}_{1,0}=\widehat{y}_{0,1}-\widehat{y}_{1,0}=\widehat{c}-\widehat{b}, \tag{17}$$

$$\widehat{\gamma}_a=\widehat{\mu}_{1,1}-\widehat{\mu}_{1,0}=\widehat{y}_{1,1}-\widehat{y}_{1,0}=\widehat{c}+\widehat{d}, \tag{18}$$

which shows that the PS$_{IG}$ model and the SEM$_{Mac}$ model (or the SEM$_{BK}$ model if the constant effect assumption holds) merges into the same model if the same ignorability assumption is used to identify causally interpretable parameters.

Whereas ignorability is sufficient to identify average causal treatment effects, a further restriction is necessary to identify principal strata proportions (π_n,π_f,π_b,π_a). Although the PS$_{Ig}$ model (or the SEM models) does not provide a complete picture regarding the principal strata proportions, it allows us to relax the exclusion restriction and allows us to learn whether treatment assignment had any harmful effects on any subpopulations under the assumption of ignorability. If the γ_b estimate is negligible or does not indicate any harmful effects, existence of some backward-improvers becomes less problematic. It is also possible to construct reasonable ranges of principal strata proportions. That is, we may estimate proportions assuming monotonicity, as explained with Equation 8, and then adjust these proportions given a reasonable range of deviation from monotonicity. In fact, the monotonicity assumption leads to the lower bound estimate of π_f, which provides us with a good rough estimate of the effect of treatment assignment on the mediator.

A Common Ground: Cross-Model Translation (CMT) Approach

The equality between the PS and SEM parameter estimates shown in Equations 15–18 exists because of an identifying assumption that is common to both models (i.e., ignorability). In contrast, this equality is unlikely to hold if the two compared models employ different assumptions to identify causal parameters. Regardless of identifying assumptions employed, translation between PS and SEM parameters is possible as we focus on the substantive meaning of these parameters. When the PS and SEM approaches rely on different identifying assumptions, it is unclear how different results from the two approaches should be compared. Translation between PS and SEM parameters is critical here to put these different findings into the same metric so that we can interpret and compare them. Further, translation between PS and SEM parameters provides multiple perspectives. The same phenomenon may look very different from different perspectives. Having different perspectives is useful because it is likely to lead to better sensitivity analyses, and therefore a better understanding of the phenomenon.

From the SEM approach perspective, translatability between PS and SEM parameters is a convenient property, which allows the SEM approach to borrow the open structure of the PS approach. In other words, it is possible to identify SEM parameters using various other assumptions than the ones typically used in the SEM approach. Translatability is also a useful property from the PS approach perspective because it allows the PS approach to borrow assumptions from the SEM approach. For example, there is no reason to completely exclude the possibility of employing the ignorability assumption when other alternative assumptions such as monotonicity and the exclusion restriction are likely to be violated. The CMT approach will be used to refer to the method of obtaining PS parameter estimates on the basis of the SEM approach and the method of obtaining SEM parameter estimates on the basis of the PS approach. These methods are possible on the basis of translatability between PS and SEM parameters.

A Single Binary Mediator

When making a causal inference with mediators, it is necessary to make assumptions about something we do not observe. For example, in identifying causal mediation effects in the SEM approach, we assume ignorability, which involves information we cannot collect from the observed data. In this case, linearity is also an unverifiable assumption because it has a meaning only under the condition that ignorability holds (and we do not know whether ignorability holds) as illustrated in Figure 2. However, in the conventional SEM approach, linearity has been viewed as an assumption about observed data. In this framework, when linearity approximately holds given observed data, there is little need to dichotomize (or stratify) continuous mediator values especially given negative consequences of dichotomizing, such as loss of information (MacCallum, Zhang, Preacher, & Rucker, 2002) and attenuated estimates, even when the hypothesized model is true (Kenny, 1979;Judd & Kenny, 1981). In the PS approach, in which causal inference has been the focus, mediators have not generally been handled as continuous variables. As shown in the PS_{AIR} and PS_{Ig} models, strong identifying assumptions are often necessary even with a single binary mediator. To accommodate continuous mediators, we need to impose more and stronger assumptions that are unverifiable, which makes causal inference modeling extremely difficult. Therefore, despite its possible pitfalls, dichotomizing or categorizing continuous mediators has been commonly practiced in the PS approach. Similarly, to make causal inferences given multiple mediators, we need to impose more and stronger assumptions that are typically unverifiable. Therefore, multiple mediators are not generally handled simultaneously in the PS approach.

In the current article, we will establish a close connection between the PS and SEM approaches focusing on a situation in which both treatment assignment and the mediator are binary. As briefly discussed above, in terms of handling continuous and/or multiple mediators, the two approaches have developed different preferences. However, disagreement between the two approaches becomes minimal under the same goal of making causal inferences. Therefore, making a close connection between the PS and SEM approaches given continuous and/or multiple mediators is a feasible task, which should be continued in future studies.

Translatability Between the SEM and PS Parameters

The PS and SEM approaches are similar in the sense that it considers change in mediator status due to treatment assignment. The mediational process theory that underlies the SEM approach implies both direct and indirect effects of treatment assignment. In the PS approach, what can be considered as direct effects are the effects of treatment assignment for never-improvers and always-improvers. For these two types of individuals, mediator status does not change regardless of treatment assignment. That is, if we know the true PS parameter values, the direct effect of treatment assignment in the SEM approach can be expressed in terms of PS parameters as

$$c=\gamma_n, \tag{19}$$

$$c+d=\gamma_a, \tag{20}$$

where d represents the difference between the two direct effects (i.e., interaction effect). If $\gamma_n = \gamma_a$, then $d = 0$ (i.e., constant effect holds). If γ_n does not equal γ_a (i.e., constant effect does not hold), according to the conventional SEM's definition, neither c nor $c + d$ can be interpreted as direct effect because it is not constant across different levels of M. From the PS perspective, individuals belonging to different principal strata may have heterogeneous characteristics (because they are not randomly assigned to different mediator status). Because never-improvers and always-improvers belong to qualitatively different subpopulations, direct effects may differ for these two types of individuals.

If causal effects of treatment assignment for forward- and backward-improvers (their potential mediator status changes depending on the treatment assignment status) are different from those for never-improvers and always-improvers, this suggests that there may be indirect effects of treatment assignment because, in that case, the differences in the outcome across assignment conditions may be partly due to the change in mediator status (for related discussions, see Rubin, 2004, and Mealli & Rubin, 2003). The effect of the mediator on the outcome conditional on treatment assignment can be expressed as

$$b+d=\gamma_f-\gamma_n, \tag{21}$$

$$b=\gamma_f-\gamma_a, \tag{22}$$

where b can be used to define the indirect effect if $\gamma_f - \gamma_n = \gamma_f - \gamma_a$. According to the conventional SEM definition, if $\gamma_f - \gamma_n \neq \gamma_f - \gamma_a$, neither b + d nor b can be used to define the indirect effect because it is not constant across different levels of X.

The effect of treatment assignment on the mediator in the SEM approach can be expressed in terms of PS parameters as

$$a=\pi_f-\pi_b. \tag{23}$$

Then, if $\gamma_n = \gamma_a$ (i.e., $d = 0$), the indirect effect in the SEM approach can be expressed in terms of PS parameters as

$$ab=\left(\pi_f-\pi_b\right)\left(\gamma_f-\gamma_n\right). \tag{24}$$

If we know the true SEM parameter values, we can also obtain PS parameter values on the basis of translatability between PS and SEM parameters. The effects of treatment assignment for never-improvers and always-improvers in the PS approach can be expressed in terms of SEM parameters as

$$\gamma_n=c, \tag{25}$$

$$\gamma_a=c+d. \tag{26}$$

The effects of treatment assignment for forward-improvers in the PS approach can be expressed in terms of SEM parameters as

$$\gamma_f=b+c+d. \tag{27}$$

From the PS perspective, whether the mediator status changes because of the treatment assignment can be inferred by looking at the proportions of forward- (π_f) and backward-improvers (π_b), whose potential mediator status changes depending on which treatment condition they are assigned to. In particular, a large proportion of forward-improvers can be considered as evidence of positive effect of treatment assignment on a mediator. Given the spirit of mediational processes, the existence of backward-improvers (i.e., negative effect of treatment assignment on a mediator) may not be welcomed. However, in the context of mediational processes, we cannot confidently exclude the possibility of having this category of individuals. For example, in comparing a standard treatment and a new treatment, some individuals' mediator status may improve under the standard treatment condition but not under the new treatment condition. If we know true SEM parameter values, principal strata proportions in the PS approach can be expressed in terms of SEM parameters as

$$\pi_f-\pi_b=a. \tag{28}$$

$$\pi_a+\pi_b=\alpha_m. \tag{29}$$

$$\pi_n+\pi_b=1-\alpha_m-a, \tag{30}$$

where the SEM parameter α_m is the mean of the binary mediator (i.e., proportion of individuals whose mediator status improved) in the control condition. The SEM parameter a is the difference between the proportion of individuals whose mediator status improved in the treatment condition and the proportion of individuals whose mediator status improved in the control condition.

In the possible presence of backward-improvers, the effect of treatment assignment for this stratum of individuals would be also of concern because this effect is more likely to be harmful than treatment assignment effects for other strata of individuals. In the PS approach, the treatment assignment effect for backward-improvers can be explicitly addressed with a designated parameter (γ_b). In the SEM approach, we do not make an explicit distinction between forward- and backward-improvers. In other words, if the effect of changing the mediator from 0 to 1 is b, the effect of changing the mediator from 1 to 0 is $-b$. Given that, γ_b can be expressed as

$$\gamma_b = -b + c. \tag{31}$$

Simulation Studies

On the basis of the connection between the SEM and PS parameters, as shown in Equations 19–31, comparing the SEM and PS_{AIR} models is quite straightforward if we know the true situation. However, what should be the most plausible and realistic situation is usually unclear, which is a fundamental problem in comparing models that operate under different identifying assumptions. In other words, which approach performs better depends on which situation we are looking at. Further, to properly examine the relative performance of the two approaches, we need to know detailed information on all involved parameters, which is not a likely situation in reality. Given that, in this section, we employ Monte Carlo simulations to further clarify the relation between the two approaches rather than to show superiority of one approach. In line with the previous sections, we focus on a randomized trial setting with a continuous outcome, a binary treatment assignment, and a binary mediator.

The PS model before assuming any identifying assumptions is used as the base (data generation) model for Monte Carlo simulations. This choice was made to fully describe the true setting before involving any SEM and PS model assumptions (i.e., ignorability is more difficult to address if we choose the SEM model as the base model as multiple sensitivity analyses would be required). Population mean and average causal effect parameters in the true PS model are presented in Table 3. On the basis of true PS parameter values, true SEM parameter values can be obtained as shown in Equations 19–24.

The PS_{AIR} model is used as an example of the PS approach in which a particular set of identifying assumptions are imposed. The key parameter γ_f is estimated in the PS_{AIR} model on the basis of Equation 10. The key parameters of interest (a, b, c) are estimated in the SEM_{BK} model on the basis of linear equations shown in Equations 1–2. Because the mediator is binary, it is also possible to use a logistic or probit regression model instead of a linear regression model in Equation 1. Whether we use a linear or a logistic (or probit) regression model to capture the effect of treatment assignment on the mediator does not affect conclusions of simulation studies presented in this article (i.e., estimates of b have the same coverage rates and power, although they are on different scales). The same model in Equation 1 is used to estimate π_f and π_a in the PS_{AIR} model, given that these parameters are equivalent to SEM parameters a and α_m under monotonicity, as shown in Equations 28 and 29. Once π_f and π_a (or, a and α_m) are identified, π_n can also be identified as shown in Equation 30. A computationally more demanding finite mixture approach (e.g., Jo, 2002b; Little & Yau, 1998) can also be employed for this purpose. However, in the absence of covariates, the two methods yield practically identical results (i.e., very close parameter and standard error estimates). In the presence of covariates, the latter approach is known to be generally more efficient. To focus on understanding of connection between the PS and SEM approaches, we limit our discussions in this article to unconditional models. Including covariates in the analyses does not change general conclusions of this study.

Once PS and SEM parameters are estimated using the PS_{AIR} and SEM_{BK} models, PS parameter estimates are translated to SEM parameter estimates, and SEM parameter estimates are translated to PS parameter estimates on the basis of the CMT approach represented in Equations 19–31. At the end of this section, we consider a setting in which constant effect does not hold. Because whether a constant effect is assumed matters in this case, the PS$_{Ig}$ (= SEM$_{Mac}$) model is additionally employed as an analysis model. The indirect effect (*ab*) estimates is not reported when the PS$_{Ig}$ (= SEM$_{Mac}$) model is employed.

The current study employs a ML estimation approach for all analyses, although simpler estimation methods (e.g., ordinary least squares) will also yield similar results. For ML estimation of the SEM and PS models, the Mplus program (Version 5.1; Muthén & Muthén, 1998–2008) was used. Parametric standard errors are computed from the information matrix of the ML estimator under the assumption of normally distributed outcomes. For parameter estimates that involve more than one parameter (e.g., *ab* and γ_T/π_f), standard errors were calculated using the delta method (e.g., Sobel, 1982), which is embedded in the Mplus program. However, it is also possible to employ more efficient ways of estimating standard errors that have been previously suggested (MacKinnon, Fritz, Williams, & Lockwood, 2007; MacKinnon et al., 2002; MacKinnon, Lockwood, & Williams, 2004; Shrout & Bolger, 2002). The delta method was chosen here because it is computationally less demanding, especially in the context of Monte Carlo simulations. For Mplus codes used for data generation and SEM and PS model analyses, see the supplemental materials.

As summarized in Table 3, the base model involves 12 populations values: four proportions (π_n, π_f, π_b, π_a) and eight means (μ_{n0},μ_{f0},μ_{b0},μ_{a0},μ_{n1},μ_{f1},μ_{b1},μ_{a1}). Depending on assumptions that we impose on these parameters, the resulting model will be differently identified, and the results will be sensitive to violation of these identifying assumptions. To causally interpret PS and SEM parameter estimates on the basis of the SEM$_{BK}$ model, ignorability and constant effect are assumed. To causally interpret PS and SEM parameter estimates on the basis of the PS$_{AIR}$ model, monotonicity and the exclusion restriction are assumed. To causally interpret PS and SEM parameter estimates on the basis of the PS$_{Ig}$ (= SEM$_{Mac}$) model, ignorability is assumed.

Ignorability of mediator status means that individuals with the same mediator value (*M*) and the same treatment assignment status (*X*) have the same potential outcome values (*Y*). This implies at the average level that

$$\mu_{n0}=\mu_{f0}, \tag{32}$$

$$\mu_{b0}=\mu_{a0}, \tag{33}$$

$$\mu_{n1}=\mu_{b1}, \tag{34}$$

$$\mu_{f1}=\mu_{a1}. \tag{35}$$

Constant effect means that the effect of treatment assignment conditional on the mediator value (direct effect) is the same for individuals with different mediator values. This implies at the average level that

$$\mu_{n1}-\mu_{n0}=\mu_{a1}-\mu_{a0}. \tag{36}$$

Monotonicity means that there are no backward-improvers. That is,

$$\pi_b=0. \tag{37}$$

Exclusion restriction means that there is no effect of treatment assignment for individuals who would not change their mediator status regardless of the treatment assignment status. This implies at the average level that

$$\mu_{n1}-\mu_{n0}=\mu_{a1}-\mu_{a0}=0, \tag{38}$$

which is a stronger form of the constant effect assumption shown in Equation 36.

When Both SEM$_{BK}$ and PS$_{AIR}$ Assumptions Hold

In the first simulation setting, all of the SEM and PS$_{AIR}$ assumptions shown in Equations 32–38 are satisfied. True values employed for principal stratum proportions are π_n = 0.4, π_f = 0.4, π_b = 0.0, and π_a = 0.2. The continuous outcome Y is normally distributed with variance of 1.0. True values employed for outcome means are μ_{n0} = μ_{f0} = 1.0, μ_{a0} = 1.5, μ_{n1} = 1.0, and μ_{f1} = μ_{a1} = 1.5. According to these true mean values, true average causal treatment effect values (see Table 3) in the PS$_{AIR}$ model are obtained as γ_n = μ_{n1} – μ_{n0} = 0.0, γ_f = μ_{f1} – μ_{f0} = 0.5, and γ_a = μ_{a1} – μ_{a0} = 0.0. Then, according to Equations 19–24, true SEM parameter values are obtained as a = 0.4, b = 0.5, and c = 0.0. According to Equation 29, the true value of the SEM parameter α_m = 0.2. This parameter is closely related to principal stratum proportions in the PS approach, although it is not considered a key parameter in the SEM approach.

Table 5 shows the results from the SEM$_{BK}$ and PS$_{AIR}$ analyses. The Monte Carlo simulation results presented in this article are based on 500 replications with a sample size of 500. Once SEM and PS parameters are estimated using the SEM$_{BK}$ and PS$_{AIR}$ models, SEM and PS parameter estimates can be translated back and forth according to Equations 19–31. For example, the γ_f estimate can be obtained using SEM parameter estimates on the basis of Equation 27 as $\widehat{\gamma}_f=\widehat{b}+\widehat{c}+\widehat{d}=0.496+0.001+0.000=0.497$, where d is fixed at zero according to the constant effect assumption (see the γ_f estimate in the SEM$_{BK}$ model in Table 5). The b estimate can be obtained using PS$_{AIR}$ parameter estimates on the basis of Equation 22 as $\widehat{b}=\widehat{\gamma}_f-\widehat{\gamma}_a=0.497-0.000=0.497$, where γ_a is fixed at zero according to the exclusion restriction (see the b estimate in the PS$_{AIR}$ model in Table 5). In summarizing analysis results with the simulated data, coverage is defined as the proportion of replications out of 500 replications in which the true parameter values are covered by the nominal 95% confidence interval of the parameter estimates. Power is defined as the proportion of replications out of 500 replications in which parameter estimates are significantly different from zero (significance level = .05, two-sided). In this article, power is reported if the coverage rate is at least 0.900 (i.e., Type I error rate is not greater than twice the nominal rate). Given the medium range sample size (i.e., 500), coverage rates may fluctuate somewhat around the nominal rate (i.e., coverage rate = 0.95, Type I error rate = 0.05). It was necessary to set the acceptable limit of this fluctuation when reporting power because some bias is expected in parameter estimation due to violation of underlying assumptions. Information on power is not useful and misleading when parameter estimates are seriously biased. However, depending on the research context, practically acceptable Type I error rates may vary in the presence of possible bias. Table 5 shows that the SEM$_{BK}$ and PS$_{AIR}$ models show very close key parameter estimates with good coverage rates, indicating compatibility of the two models. This also confirms that the SEM parameters are properly interpreted in the PS framework. Whereas coverage rates are very close between the two sets of parameters, power to detect an indirect effect is different in the two models. That is, the effect b in the SEM$_{BK}$ model (power = 1.000) is detected with greater power compared with the effect γ_f in the PS$_{AIR}$ model (power = 0.586). This implies an advantage of using the SEM model when its identifying assumptions hold, although both the SEM and PS$_{AIR}$ parameters (b and γ_f) are positioned to capture a mathematically equivalent quantity.

When SEM$_{BK}$ Assumptions Hold

When SEM$_{BK}$ assumptions (ignorability, constant effect) hold, PS$_{AIR}$ assumptions can be violated only in certain ways. That is, either monotonicity or the exclusion restriction can be violated, but both of them cannot be violated at the same time. If both are violated at the same time, at least one of the SEM$_{BK}$ assumptions is also violated.

When only monotonicity is violated (ignorability, constant effect, and exclusion restriction hold), the relation between γ_f and γ_b is straightforward. That is, $\gamma_b = -\gamma_f$. Under this condition, violation of monotonicity does not lead to bias in the estimation of the average causal treatment effect γ_f as pointed out in Angrist et al. (1996), although it leads to bias in the estimation of principal strata proportions. In the simulation setting that satisfies this condition, true values employed for principal stratum proportions are $\pi_n = 0.3, \pi_f = 0.4, \pi_b = 0.1$, and $\pi_a = 0.2$. True values employed for outcome means are $\mu_{n0} = \mu_{f0} = 1.0$, $\mu_{b0} = \mu_{a0} = 1.5$, $\mu_{n1} = \mu_{b1} = 1.0$, and $\mu_{f1} = \mu_{a1} = 1.5$. According to these true mean values, the true average causal treatment effect values in the PS$_{AIR}$ model are obtained as $\gamma_n = \mu_{n1} - \mu_{\mathrm{n}0} = 0.0$, $\gamma_f = \mu_{f1} - \mu_{f0} = 0.5$, $\gamma_b = \mu_{b1} - \mu_{b0} = -0.5$, and $\gamma_a = \mu_{a1} - \mu_{a0} = 0.0$. Then, according to Equations 19–24, true SEM parameter values are obtained as $a = 0.3$, $b = 0.5$, and $c = 0.0$. According to Equation 29, the true value of the SEM parameter $\alpha_m = 0.3$.

Table 6 shows that the SEM$_{BK}$ and PS$_{AIR}$ models generated very close SEM and PS parameter estimates and coverage rates, indicating that the two models are still compatible when faced with violation of monotonicity. In both models, principal strata proportions are estimated on the basis of Equations 28–30. These proportion estimates show low coverage rates, which is expected given that $\pi_b = 0.1$ (i.e., monotonicity is violated) in the simulation setting. Other than that, both SEM and PS parameters are well estimated with good coverage rates in the SEM$_{BK}$ and PS$_{AIR}$ models. As in Table 5, power to detect an indirect effect is different in the two models because the effect b in the SEM$_{BK}$ model (power = 1.0) is detected with much greater power compared with the effect γ_f in the PS$_{AIR}$ model (power = 0.382). This again implies that the SEM model will detect an indirect effect with greater power as long as its identifying assumptions hold.

Unlike violation of monotonicity, violation of the exclusion restriction actually results in biased parameter estimates in the PS$_{AIR}$ model even if SEM$_{BK}$ assumptions hold. How violation of the exclusion restriction affects PS parameter estimates has been discussed in several articles (e.g., Angrist et al., 1996; Hirano, Imbens, Rubin, & Zhou, 2000; Jo, 2002a). As shown in Equations 36 and 38, the exclusion restriction is a stronger form of the constant effect assumption. Therefore, if SEM$_{BK}$ assumptions hold and the exclusion restriction is violated, the violation only affects the PS$_{AIR}$ model. In the simulation setting that satisfies this condition, true values employed for principal stratum proportions are $\pi_n = 0.4$, $\pi_f = 0.4$, $\pi_b = 0.0$, and $\pi_a = 0.2$. True values employed for outcome means are $\mu_{n0} = \mu_{f0} = 1.0$, $\mu_{a0} = 1.3$, $\mu_{n1} = 1.2$, and $\mu_{f1} = \mu_{a1} = 1.5$. According to these true mean values, true average causal treatment effect values in the PS$_{AIR}$ model are obtained as $\gamma_n = \mu_{n1} - \mu_{n0} = 0.2$, $\gamma_f = \mu_{f1} - \mu_{f0} = 0.5$, and $\gamma_a = \mu_{a1} - \mu_{a0} = 0.2$. Then, according to Equations 19–24, true SEM parameter values are obtained as $a = 0.4$, $b = 0.3$, and $c = 0.2$. According to Equation 29, the true value of the SEM parameter $\alpha_m = 0.2$.

Table 7 shows that the PS$_{AIR}$ model overestimated the PS parameter γ_f and SEM parameters related to γ_f (i.e., *b* and *ab*) with coverage rates much lower than the nominal rate. The results imply that indirect effect estimates of the PS$_{AIR}$ model need to be interpreted with caution when plausibility of the exclusion restriction (i.e., no direct effect of treatment on the outcome) is questionable. For the same parameters, the SEM$_{BK}$ model generated estimates with good coverage rates. Power has little meaning for these parameters in the PS$_{AIR}$ model and therefore is not reported in Table 7. In both models, parameters π_f and *a* were estimated with good coverage rates.

When PS$_{AIR}$ Assumptions Hold

When PS$_{AIR}$ assumptions (monotonicity, exclusion restriction) hold, the only possible combination of violation of SEM$_{BK}$ assumptions is violation of ignorability. Because the exclusion restriction is a stronger form of the constant effect assumption, constant effect holds if exclusion restriction holds. In the simulation setting that satisfies this condition, true values employed for principal stratum proportions are $\pi_n = 0.4$, $\pi_f = 0.4$, $\pi_b = 0.0$, and $\pi_a = 0.2$. True values employed for outcome means are $\mu_{n0} = 1.0$, $\mu_{f0} = 1.2$, $\mu_{a0} = 1.9$, $\mu_{n1} = 1.0$, $\mu_{f1} = 1.7$, and $\mu_{a1} = 1.9$. According to these true mean values, true average causal treatment effect values in the PS$_{AIR}$ model are obtained as $\gamma_n = \mu_{n1} - \mu_{n0} = 0.0$, $\gamma_f = \mu_{f1} - \mu_{f0} = 0.5$, and $\gamma_a = \mu_{a1} - \mu_{a0} = 0.0$. Then, according to Equations 19–24, true SEM parameter values are obtained as $a = 0.4$, $b = 0.5$, and $c = 0.0$. According to Equation 29, the true value of the SEM parameter $\alpha_m = 0.2$.

Table 8 shows that, given violation of ignorability, the SEM$_{BK}$ model misestimated SEM parameters *b*, *c*, *ab*, and the PS parameter γ_f with coverage rates much lower than the nominal rate. As shown in Equations 32–35, ignorability is a strong assumption that imposes restrictions on multiple parameters. In the presence of pretreatment covariates, conditional ignorability is usually a preferred assumption. Conditional ignorability is still unverifiable but weaker than the ignorability assumption. As shown in Tables 5–7, having smaller (compared with those of the PS$_{AIR}$ model) standard errors is an advantage in the SEM$_{BK}$ model when its identifying assumptions hold. Table 8 shows that this property may work as a disadvantage when its identifying assumptions are violated because it will provide biased estimates with greater statistical significance. In particular, the *b* estimates in the SEM model show an extremely low coverage rate (0.198) because of the combination of overestimation of the parameter and small standard errors. The results imply that parameter estimates of the SEM$_{BK}$ model need to be interpreted with caution. As long as mediator values are not randomly assigned to individuals, the plausibility of ignorability is questionable. For the same parameters, the PS$_{AIR}$ model generated estimates with good coverage rates. In both models, parameters π_f and *a* were estimated with good coverage rates.

When Both SEM$_{BK}$ and PS$_{AIR}$ Assumptions Are Violated

In practice, SEM$_{BK}$ and PS$_{AIR}$ assumptions can be simultaneously violated (i.e., combinations of monotonicity, exclusion restriction, ignorability, and constant effect violations), which results in higher dimensional interactions among violated assumptions in creating bias. In this case, it is unlikely that consistent conclusions can be reached about the simultaneous impact of these violations. That is, relative performance of different models may vary depending not only on directions and sizes of deviations from employed assumptions but also on interactions among violated assumptions.

One example of multiple violations of assumptions is shown in Table 9. In this simulation setting, all of the SEM$_{BK}$ and PS$_{AIR}$ assumptions are violated (i.e., monotonicity, exclusion restriction, ignorability, constant effect). Because the constant effect assumption is violated in this setting, the SEM$_{BK}$ and SEM$_{Mac}$ (or PS$_{Ig}$) models may result in different parameter estimates. Therefore, both models are employed to analyze simulated data in this setting. True values employed for principal stratum proportions are $\pi_n = 0.3$, $\pi_f = 0.4$, $\pi_b = 0.1$, and $\pi_a = 0.2$. The existence of backward-improvers ($\pi_b = 0.1$) indicates that treatment assignment has some negative effect on the mediator. True values employed for outcome means are $\mu_{n0} = 1.0$, $\mu_{f0} = 1.2$, $\mu_{b0} = 1.0$, $\mu_{a0} = 1.6$, $\mu_{n1} = 0.9$, $\mu_{b1} = 0.6$, $\mu_{f1} = 1.7$, $\mu_{a1} = 1.8$. On the basis of these true mean values, $\gamma_n=\mu_{n1}-\mu_{n0}=-0.1$, $\gamma_f=\mu_{f1}-\mu_{f0} = 0.5$, $\gamma_b = \mu_{b1} -\mu_{b0} = -0.4$, and $\gamma_a = \mu_{a1} - \mu_{a0} = 0.2$. This set of parameter values indicates some negative effect of treatment assignment on never-improvers, some positive effect of treatment assignment on always-improvers. The absolute size of γ_b (effect size = 0.4) is nearly as big as that of γ_f (effect size = 0.5), indicating that treatment assignment has a substantially negative effect on the outcome for backward-improvers. By using Equations 19–24, true SEM parameter values are obtained as $a = 0.3$, $b = 0.3$, $c=-0.1$, and $d = 0.3$. According to Equation 29, the true value of the SEM parameter $\alpha_m = 0.3$.

Table 9 shows that all three models generated substantially biased estimates for SEM parameters, which was expected given that all three models violate some of the identifying assumptions. The average effect of treatment assignment conditional on the mediator (c) is estimated with bias both in the SEM$_{BK}$ and SEM$_{Mac}$ (or PS$_{Ig}$) models (this parameter is fixed in the PS$_{AIR}$ model). The interaction effect d is also estimated with bias in the SEM$_{Mac}$ (or PS$_{Ig}$) model (this parameter is fixed in the PS$_{AIR}$ and SEM$_{BK}$ models). This indicates that violation of constant effect cannot be properly tested on the basis of observed data given a possible violation of ignorability. The effect b is estimated with bias in the PS$_{AIR}$ and SEM$_{BK}$ models but not in the SEM$_{Mac}$ model. The good coverage of b estimates can be explained by cancellation of biases being combined in this specific example. However, biases due to multiple violations can also accumulate, and therefore b estimates in the SEM$_{Mac}$ can be more biased depending on the setting. In all three models, PS parameters are estimated with substantial bias except for the parameter γ_f in the PS$_{AIR}$ model. The reasonable coverage rate of γ_f estimates in the PS$_{AIR}$ model can be seen as another example of cancellation of biases when terms are combined. However, γ_f can be estimated with more bias in the PS$_{AIR}$ model depending on the setting.

Conclusions

To examine the relation between the SEM and PS approaches in identifying causal mediation effects, we employed the CMT approach in this article. In this approach, PS parameter estimates and SEM parameter estimates are translated back and forth across the PS and SEM models, enabling a complete comparison between the two modeling approaches. This approach is possible because of the close conceptual connection between SEM and PS parameters. From the CMT approach perspective, what affects our inference is the choice of identifying assumptions, not the modeling framework itself. The PS$_{Ig}$ and SEM$_{Mac}$ models show an ultimate example in which the two approaches completely converge into the same model when the same identifying assumptions are employed even though they start from different modeling approaches.

Despite the close connection between their parameters, the SEM and PS approaches differ significantly in terms of the choice of identifying assumptions. In the SEM approach, which focuses on observed mediator status, there is little room for flexibility because the assumption of ignorability always has to be present. In the PS approach, however, many different sets of identifying assumptions can be employed because there are no inherently embedded assumptions. From the CMT approach perspective, the relative inflexibility of the SEM approach is not necessarily a problem given translatability between SEM and PS parameters. That is, SEM parameter estimates can be restructured using PS parameter estimates. In this way, SEM modeling is possible using diverse identifying assumptions borrowing from the open structure of the PS approach. For example, there is no clear reason to exclude the possibility of employing monotonicity and the exclusion restriction assumptions when plausibility of the ignorability assumption is highly suspect. The PS parameters estimated using the PS_{AIR} model can be translated into SEM parameter estimates, which provide the same interpretation as that of the SEM_{BK} model estimates.

Translatability is also a useful property from the PS approach perspective because it facilitates communication between the PS framework and the more conventional modeling framework that does not involve potential outcomes. Adopting SEM's ignorability assumption, which supports comparisons across different principal strata, may seem like a self-contradictory choice in the PS approach, which has been developed emphasizing the fact that individuals in different principal strata are not comparable. However, in situations in which the plausibility of monotonicity and the exclusion restriction are questionable, ignorability or conditional ignorability can be a helpful alternative assumption. Further consideration of mediational processes dealt with in the SEM approach, which tend to be more general than mediational processes commonly dealt with in the PS approach, is also likely to broaden PS modeling practice and to facilitate development of alternative models and sensitivity analysis methods.

Monte Carlo simulations employed in this study demonstrated how the choice of identifying assumptions affects the quality of parameter estimates under known conditions (i.e., we know which assumptions are violated). However, when analyzing real data, making a choice among competing sets of assumptions and different results is not an easy task. Even if we know relative plausibility of these assumptions, successful decision making is still not guaranteed because assumptions with higher plausibility do not necessarily result in less biased estimates (Jo, 2008). Further, even if we choose the set of identifying assumptions that provide the least biased parameter estimates, it is possible that the chosen results will be closer, but not close enough, to the truth. These are fundamental difficulties that arise when identification of causal effects relies on assumptions about quantities we do not observe. Naturally, good sensitivity analysis methods are critical to effectively assess causal mediation effects.

Given that translation between the SEM and PS approaches is possible, the next logical step will be to develop improved sensitivity analysis methods that consider both approaches' perspectives. To properly estimate bias due to violations of identifying assumptions, we need to understand the bias mechanisms, which can be highly complex in practice. For example, ignorability can be violated in many different ways because it involves multiple restrictions as shown in Equations 32–35. Monotonicity and the exclusion restriction can be simultaneously violated. In these situations, the resulting bias mechanisms may involve high dimensional interactions between violated assumptions that produce different magnitudes of bias depending on the degree of the violation of each assumption. Therefore, without knowing the exact bias mechanisms, it is hard to estimate bias quantities and corresponding causal effects. Another important component in sensitivity analysis is science-based information (expert knowledge) on plausible ranges of deviations from identifying assumptions. If possible ranges of deviations and bias mechanisms are known, possible ranges of causal effects can also be estimated. Much research is needed to develop efficient sensitivity analysis methods that facilitate probing the multiple assumptions and perspectives of the PS and SEM approaches to mediation analysis.

Acknowledgments

The research was supported by National Institute of Mental Health Grants MH066319 and MH066247. I thank several referees for their encouragement and insightful comments. I also thank David MacKinnon, Amiram Vinokur, Michael Sobel, and Helena Kraemer for valuable advice. The study also benefitted from helpful feedback provided by the Prevention Science Methodology Group.

References

Aiken, LS.; West, SG. Multiple regression: Testing and interpreting interactions. Sage; Newbury Park, CA: 1991.

Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 1996;91:444–455.

Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 1986;51:1173–1182. [PubMed: 3806354]

Bollen, KA. Total, direct, and indirect effects in structural equation models. In: Clogg, CC., editor. Sociological methodology. American Sociological Association; Washington, DC: 1987. p. 37-69.

Brown CH. Comparison of mediational selected strategies and sequential designs for preventive trials: Comment on a proposal by Pillow et al. American Journal of Community Psychology 1991;19:837–846. [PubMed: 1793091]

DeSteno D, Valdesolo P, Bartlett MY. Jealously and the threatened self: Getting to the heart of the green-eyed monster. Journal of Personality and Social Psychology 2006;91(4):626–641. [PubMed: 17014289]

Emery RE. Mediational screening in theory and in practice. American Journal of Community Psychology 1991;19:853–857. [PubMed: 1793093]

Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58:21–29. [PubMed: 11890317]

Gelman, A.; Hill, J. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press; New York: 2007.

Hirano K, Imbens GW, Rubin DB, Zhou XH. Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics 2000;1:69–88. [PubMed: 12933526]

Holland PW. Statistics and causal inference. Journal of the American Statistical Association 1986;81:945–960.

Holland, PW. Causal inference, path analysis, and recursive structural equation models. In: Clogg, CC., editor. Sociological methodology. American Sociological Association; Washington, DC: 1988. p. 449-484.

Hong G, Raudenbush SW. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. Journal of the American Statistical Association 2006;101:901–910.

Jo B. Estimating intervention effects with noncompliance: Alternative model specifications. Journal of Educational and Behavioral Statistics 2002a;27:385–420.

Jo B. Statistical power in randomized intervention studies with noncompliance. Psychological Methods 2002b;7:178–193. [PubMed: 12090409]

Jo B. Bias mechanisms in intention-to-treat analysis with data subject to treatment noncompliance and missing outcomes. Journal of Educational and Behavioral Statistics 2008;33:158–185. [PubMed: 20689663]

Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. Evaluation Review 1981;5:602–619.

Kenny, DA. Correlation and causality. Wiley; New York: 1979.

Kraemer HC, Blasey CM. Centering in regression analyses: A strategy to prevent errors in statistical inference. International Journal of Methods in Psychiatric Research 2006;13:141–151. [PubMed: 15297898]

Kraemer HC, Kiernan M, Essex MJ, Kupfer DJ. How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. Health Psychology 2008;27:101–108.

Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trial. Archives of General Psychiatry 2002;59:877–883. [PubMed: 12365874]

Little RJA, Yau L. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. Psychological Methods 1998;3:147–159.

MacCallum RC, Zhang S, Preacher KJ, Rucker D. On the practice of dichotomization of quantitative variables. Psychological Methods 2002;7:19–40. [PubMed: 11928888]

MacKinnon, DP. Introduction to statistical mediation analysis. Erlbaum; New York: 2008.

MacKinnon DP, Dwyer JH. Estimating mediating effects in prevention studies. Evaluation Review 1993;17:144–158.

MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annual Review of Psychology 2007;58:593–614.

MacKinnon DP, Fritz MS, Williams J, Lockwood CM. Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. Behavior Research methods 2007;39:384–389. [PubMed: 17958149]

MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychological Methods 2002;7:83–104. [PubMed: 11928892]

MacKinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research 2004;39:99–118. [PubMed: 20157642]

Mealli F, Rubin DB. Comment on Adams et al. and assumptions allowing the estimation of direct causal effects. Journal of Econometrics 2003;112:79–87.

Meyer JP, Gellatly IR. Perceived performance norm as a mediator in the effect of assigned goal on personal goal and task performance. Journal of Applied Psychology 1988;73(3):410–420.

Milling LS, Reardon JM, Carosella GM. Mediation and moderation of psychological pain treatments: Response expectancies and hypnotic suggestibility. Journal of Consulting and Clinical Psychology 2006;74(2):253–262. [PubMed: 16649870]

Morgan, SL.; Winship, C. Counterfactuals and causal inference. Cambridge University Press; New York: 2007.

Muthén, LK.; Muthén, BO. Mplus user's guide. Author; Los Angeles: 1998–2008.

Neyman J. On the application of probability theory to agricultural experiments. Essay on principles Section 9 translated in Statistical Science 1990;5:465–480. Original work published 1923.

Pearl, J. Direct and indirect effects. In: Besnard, P.; Hanks, S., editors. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence; San Francisco: Morgan Kaufmann; 2001. p. 411-420.

Pillow DR, Sandler IN, Braver SL, Wolchik SA, Gersten JC. Theory based screening for prevention: Focusing on mediating processes in children of divorce. American Journal of Community Psychology 1991;19:809–836. [PubMed: 1793090]

Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect. Mathematical Modeling 1986;7:1393–1512.

Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992;3:142–155.

Robins JM, Greenland S. Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. Journal of the American Statistical Association 1994;89:737–749.

Rosenbaum PR. Inference between units in randomized experiments. Journal of the American Statistical Association 2007;102:191–200.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974;66:688–701.

Rubin DB. Bayesian inference for causal effects: The role of randomization. Annals of Statistics 1978;6:34–58.

Rubin DB. Discussion of “Randomization analysis of experimental data in the Fisher randomization test” by D. Basu. Journal of the American Statistical Association 1980;75:591–593.

Rubin DB. Comment on Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 1990;5:472–480.

Rubin DB. Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics 2004;31:161–170.

Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 2005;100:322–331.

Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods 2002;7:422–445. [PubMed: 12530702]

Sobel, ME. Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt, S., editor. Sociological methodology. American Sociological Association; Washington, DC: 1982. p. 290-312.

Sobel, ME. Direct and indirect effects in linear structural equation models. In: Long, JS., editor. Common problems/proper solutions. Sage; Beverly Hills, CA: 1988. p. 46-64.

Sobel ME. What do randomized studies of housing mobility demonstrate: Causal inference in the face of interference. Journal of the American Statistical Association 2006;101:1398–1407.

Sobel ME. Identification of causal parameters in randomized prevention studies with mediators. Journal of Educational and Behavioral Statistics 2008;33:230–251.

Spencer SJ, Zanna MP, Fong GT. Establishing a causal chain: Why experiments are often more effective in examining psychological process than mediational analyses. Journal of Personality and Social Psychology 2005;89:845–851. [PubMed: 16393019]

Ten Have TR, Elliott MR, Joffe M, Zanutto E, Datto C. Causal models for randomized physician encouragement trials in treating primary care depression. Journal of the American Statistical Association 2004;99:16–25.

Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. Causal mediation analysis with rank preserving models. Biometrics 2007;63:926–934. [PubMed: 17825022]

Vinokur AD, Price RH, Schul Y. Impact of the JOBS intervention on unemployed workers varying in risk for depression. American Journal of Community Psychology 1995;23:39–74. [PubMed: 7572826]

Vinokur AD, Schul Y. Mastery and inoculation against setbacks as active ingredients in intervention for the unemployed. Journal of Consulting and Clinical Psychology 1997;65:867–877. [PubMed: 9337505]

West, SG.; Biesanz, J.; Pitts, SC. Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In: Reis, HT.; Judd, CM., editors. Handbook of research methods in social psychology. Cambridge University Press; New York: 2000. p. 40-84.

Winship C, Morgan SL. The estimation of causal effects from observational data. Annual Review of Sociology 1999;25:659–707.

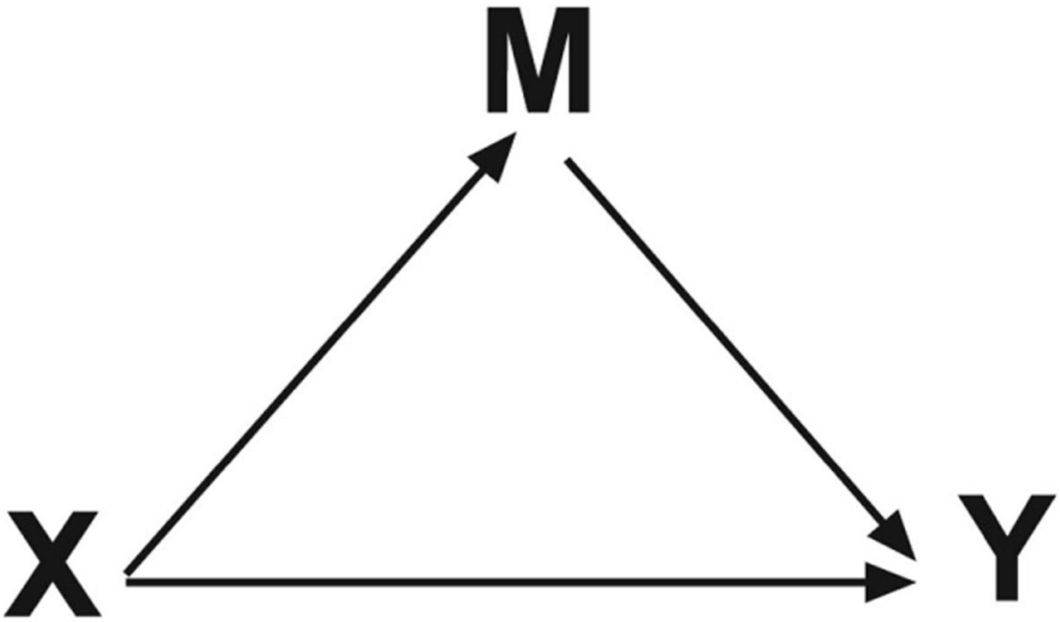

Figure 1.
Conceptual framework of mediational process. *X* = experimental (treatment) condition assignment status; *M* = mediator; *Y* = outcome.

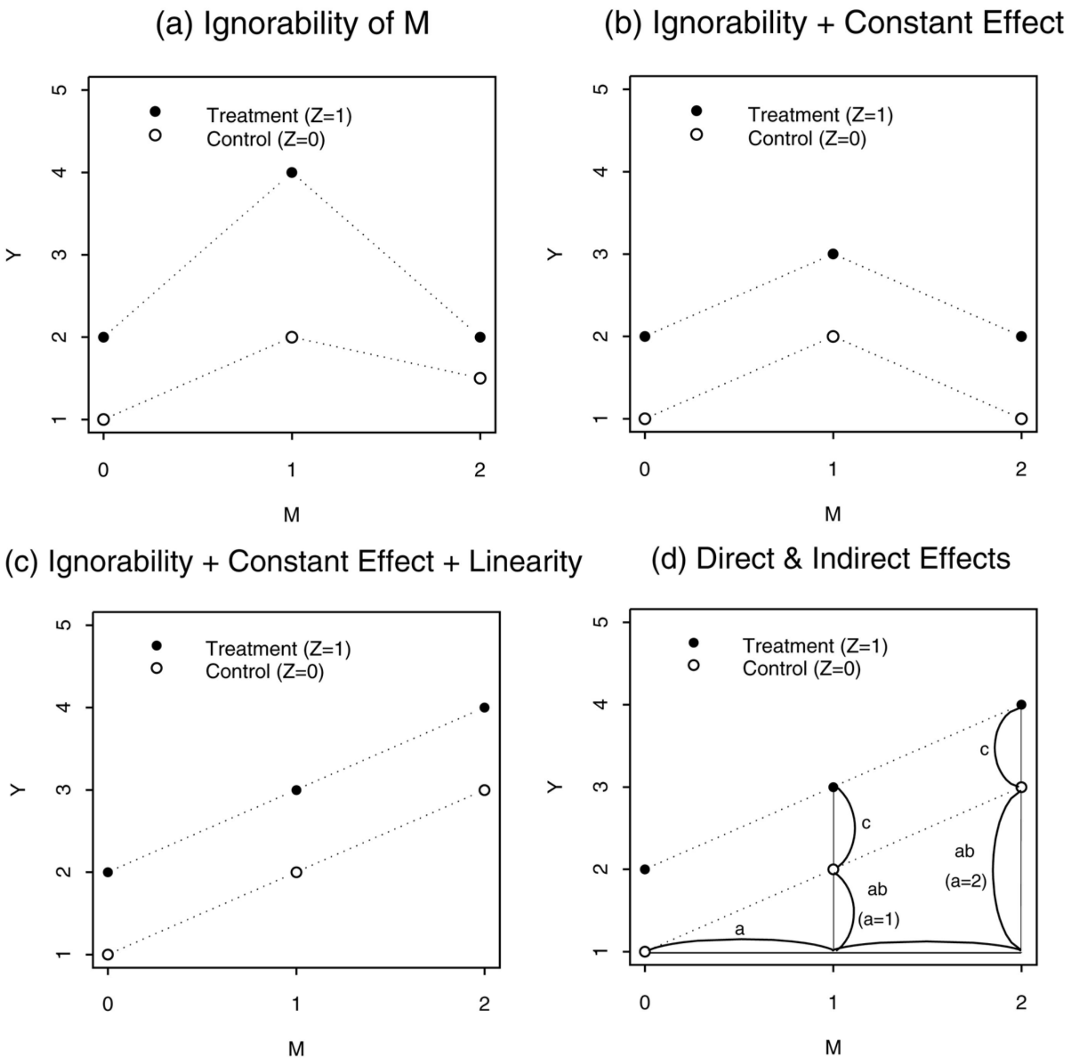

Figure 2.

Assumptions to identify causal effects in the structural equation modeling approach. In this hypothetical setting, 6 individuals are randomly assigned either to the treatment or to the control condition. *Y* = outcome; *M* = mediator; *X* = treatment assignment; *a* = the effect of treatment assignment on *M*; *b* = the effect of *M* on *Y* conditional on the effect of *X* on *Y*; *c* = the direct effect of *X* on *Y* conditional on the effect of *M* on *Y*; *ab* = *a* × *b*, which is the indirect effect of *X* on *Y*. In this illustrative example, *c* = 1.0, *b* = 1.0, and *a* takes one of the three values (0, 1, 2). Panel A shows the assumption of ignorability (i.e., 6 individuals with different *M* values are comparable on all measured and unmeasured covariates). In Panel B, constant effect implies that treatment assignment effect is the same across different *M* values. In Panel C, linearity implies that the relation between *Y* and *M* is linear. Panel D shows how the total effect is partitioned into the direct (*c*) and indirect (*ab*) effects.

Table 1

Comparison Between the SEM and PS Approaches

Aspect	SEM approach	PS approach
Key modeling difference	Focus on observed mediator status	Focus on potential mediator status
Key strength	Compatibility with the mediational process theory	Clarity in underlying assumptions necessary for causal interpretation
Frequent context	Very general, theory-based mediation	Treatment noncompliance
Fixed assumptions for causal interpretation	Ignorability of mediator status	None
Frequently used assumptions for causal interpretation	Ignorability, constant effect, linearity	Monotonicity, exclusion restriction
Key method of checking validity of causal effect estimates	Test the fit of the model to the data	Test sensitivity of estimates to violating underlying assumptions

Note. SEM = structural equation modeling; PS = principal stratification.

Table 2

Principal Stratification on the Basis of Potential Mediator Status

	Potential mediator		Potential outcome		
Principal stratum membership (C_i)	$\boldsymbol{M_i(0)}$	$\boldsymbol{M_i(1)}$	$\boldsymbol{Y_i(C_i, 0)}$	$\boldsymbol{Y_i(C_i, 1)}$	**Individual causal effect given C_i**
n (never-improver)	0	0	$Y_i(n, 0)$	$Y_i(n, 1)$	$Y_i(n, 1) - Y_i(n, 0)$
f (forward-improver)	0	1	$Y_i(f, 0)$	$Y_i(f, 1)$	$Y_i(f, 1) - Y_i(f, 0)$
b (backward-improver)	1	0	$Y_i(b, 0)$	$Y_i(b, 1)$	$Y_i(b, 1) - Y_i(b, 0)$
a (always-improver)	1	1	$Y_i(a, 0)$	$Y_i(a, 1)$	$Y_i(a, 1) - Y_i(a, 0)$

Note. C = principal stratum membership; M = mediator; Y = outcome.

Table 3

Mean Potential Values and Average Causal Effects Given Principal Strata

	Mean potential outcome		
Proportions of principal strata	***X* = 0**	***X* = 1**	**Average causal effect given *C***
π_n	μ_{n0}	μ_{n1}	$\gamma_n = \mu_{n1} - \mu_{n0}$
π_f	μ_{f0}	μ_{f1}	$\gamma_f = \mu_{f1} - \mu_{f0}$
π_b	μ_{b0}	μ_{b1}	$\gamma_b = \mu_{b1} - \mu_{b0}$
π_a	μ_{a0}	μ_{a1}	$\gamma_a = \mu_{a1} - \mu_{a0}$

Note. *X* = experimental (treatment) condition assignment status; *C* = principal stratum membership; *n* = never-improver; *f* = forward-improver; *b* = backward-improver; *a* = always-improver.

Table 4

Mean Potential Values and Average Causal Effects Under Ignorability

	Mean potential outcome		
Proportions of principal strata	***X* = 0**	***X* = 1**	**Average causal effect given *C***
π_n	$\mu_{n0} = \mu_{0,0}$	$\mu_{n1} = \mu_{0,1}$	$\gamma_n = \mu_{0,1} - \mu_{0,0}$
π_f	$\mu_{f0} = \mu_{0,0}$	$\mu_{f1} = \mu_{1,1}$	$\gamma_f = \mu_{1,1} - \mu_{0,0}$
π_b	$\mu_{b0} = \mu_{1,0}$	$\mu_{b1} = \mu_{0,1}$	$\gamma_b = \mu_{0,1} - \mu_{1,0}$
π_a	$\mu_{a0} = \mu_{1,0}$	$\mu_{a1} = \mu_{1,1}$	$\gamma_a = \mu_{1,1} - \mu_{1,0}$

Note. *X* = experimental (treatment) condition assignment status; *C* = principal stratum membership; *n* = never-improver; *f* = forward-improver; *b* = backward-improver; *a* = always-improver.

Table 5

When All of the SEM_{BK} and PS_{AIR} Model Assumptions Hold

		SEM_{BK} model				PS_{AIR} model			
	True value	**Estimate**	***SE***	**Coverage**	**Power**	**Estimate**	***SE***	**Coverage**	**Power**
SEM parameter									
a	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
b	0.50	0.496	0.100	0.956	1.000	0.497	0.229	0.942	0.586
c	0.00	0.001	0.098	0.938	0.062	0.000			
ab	0.20	0.197	0.045	0.938	1.000	0.198	0.092	0.938	0.566
d	0.00	0.000				0.000			
α_m	0.20	0.201	0.028	0.964	1.000	0.201	0.028	0.964	1.000
PS parameter									
π_n	0.40	0.402	0.028	0.930	1.000	0.402	0.028	0.930	1.000
π_f	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
π_b	0.00	0.000				0.000			
π_a	0.20	0.201	0.028	0.964	1.000	0.201	0.028	0.964	1.000
γ_n	0.00	0.001	0.098	0.938	0.062	0.000			
γ_f	0.50	0.497	0.108	0.944	0.996	0.497	0.229	0.942	0.586
γ_b		−0.495	0.166						
γ_a	0.00	0.001	0.098	0.938	0.062	0.000			

Note. SEM = structural equation modeling; *BK* model = Baron and Kenny's (1986) model; PS = principal stratification; *AIR* model = Angrist, Imbens, and Rubin's (1996) model; *a* = the effect of treatment assignment *X* on the mediator *M*; *b* = the effect of *M* on outcome *Y* conditioning on *X*; *c* = the effect of *X* on *Y* conditioning on *M*; *ab* = the indirect effect (= *a* × *b*); *d* = the interaction effect (how much the effect of *X* increases [or decreases] as the value of *M* increases by one unit); α_m = the mean of *M* in the control condition; π_n = the proportion of never-improvers; π_f = the proportion of forward-improvers; π_b = the proportion of backward-improvers; π_a = the proportion of always-improvers; γ_n = the average causal treatment effect for never-improvers; γ_f = the average causal treatment effect for forward-improvers; γ_b = the average causal treatment effect for backward-improvers; γ_a = the average causal treatment effect for always-improvers.

Table 6

When Only Monotonicity Is Violated

True value		SEM$_{BK}$ model				PS$_{AIR}$ model			
		Estimate	***SE***	**Coverage**	**Power**	**Estimate**	***SE***	**Coverage**	**Power**
SEM parameter									
a	0.30	0.297	0.042	0.942	1.000	0.297	0.042	0.942	1.000
b	0.50	0.500	0.094	0.960	1.000	0.497	0.312	0.948	0.382
c	0.00	0.000	0.094	0.938	0.062	0.000			
ab	0.15	0.148	0.035	0.938	1.000	0.148	0.092	0.938	0.368
d	0.00	0.000				0.000			
α_m	0.30	0.301	0.030	0.968	1.000	0.301	0.030	0.968	1.000
PS parameter									
π_n	0.30	0.402	0.030	0.066		0.402	0.030	0.066	
π_f	0.40	0.297	0.042	0.324		0.297	0.042	0.324	
π_b	0.10	0.000				0.000			
π_a	0.20	0.301	0.030	0.054		0.301	0.030	0.054	
γ_n	0.00	0.000	0.094	0.938	0.062	0.000			
γ_f	0.50	0.499	0.111	0.958	0.994	0.497	0.312	0.948	0.382
γ_b	−0.50	−0.500	0.152	0.938	0.882				
γ_a	0.00	0.000	0.094	0.938	0.062	0.000			

Note. SEM = structural equation modeling; *BK* model = Baron and Kenny's (1986) model; PS = principal stratification; *AIR* model = Angrist, Imbens, and Rubin's (1996) model; a = the effect of treatment assignment X on the mediator M; b = the effect of M on outcome Y conditioning on X; c = the effect of X on Y conditioning on M; ab = the indirect effect (= $a \times b$); d = the interaction effect (how much the effect of X increases [or decreases] as the value of M increases by one unit); α_m = the mean of M in the control condition; π_n = the proportion of never-improvers; π_f = the proportion of forward-improvers; π_b = the proportion of backward-improvers; π_a = the proportion of always-improvers; γ_n = the average causal treatment effect for never-improvers; γ_f = the average causal treatment effect for forward-improvers; γ_b = the average causal treatment effect for backward-improvers; γ_a = the average causal treatment effect for always-improvers.

Table 7

When Only Exclusion Restriction Is Violated

True value		SEM$_{BK}$ model				PS$_{AIR}$ model			
		Estimate	***SE***	**Coverage**	**Power**	**Estimate**	***SE***	**Coverage**	**Power**
SEM parameter									
a	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
b	0.30	0.296	0.100	0.956	0.826	0.805	0.235	0.396	
c	0.20	0.201	0.098	0.938	0.528	0.000			
ab	0.12	0.118	0.042	0.938	0.824	0.318	0.090	0.408	
d	0.00	0.000				0.000			
α_m	0.20	0.201	0.028	0.966	1.000	0.201	0.028	0.966	1.000
PS parameter									
π_n	0.40	0.402	0.028	0.930	1.000	0.402	0.028	0.930	1.000
π_f	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
π_b	0.00	0.000				0.000			
π_a	0.20	0.201	0.028	0.966	1.000	0.201	0.028	0.966	1.000
γ_n	0.20	0.201	0.098	0.938	0.528	0.000			
γ_f	0.50	0.497	0.108	0.944	0.996	0.805	0.235	0.766	
γ_b		−0.095	0.166						
γ_a	0.20	0.201	0.098	0.938	0.528	0.000			

Note. SEM = structural equation modeling; *BK* model = Baron and Kenny's (1986) model; PS = principal stratification; *AIR* model = Angrist, Imbens, and Rubin's (1996) model; *a* = the effect of treatment assignment *X* on the mediator *M*; *b* = the effect of *M* on outcome *Y* conditioning on *X*; *c* = the effect of *X* on *Y* conditioning on *M*; *ab* = the indirect effect (= *a* × *b*); *d* = the interaction effect (how much the effect of *X* increases [or decreases] as the value of *M* increases by one unit); α_m = the mean of *M* in the control condition; π_n = the proportion of never-improvers; π_f = the proportion of forward-improvers; π_b = the proportion of backward-improvers; π_a = the proportion of always-improvers; γ_n = the average causal treatment effect for never-improvers; γ_f = the average causal treatment effect for forward-improvers; γ_b = the average causal treatment effect for backward-improvers; γ_a = the average causal treatment effect for always-improvers.

Table 8

When Only Ignorability Is Violated

		SEM$_{BK}$ model				PS$_{AIR}$ model			
	True value	Estimate	*SE*	Coverage	Power	Estimate	*SE*	Coverage	Power
SEM parameter									
a	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
b	0.50	0.776	0.100	0.198		0.492	0.232	0.944	0.562
c	0.00	−0.111	0.098	0.788		0.000			
ab	0.20	0.308	0.051	0.424		0.197	0.095	0.952	0.528
d	0.00	0.000				0.000			
α_m	0.20	0.201	0.028	0.964	1.000	0.201	0.028	0.964	1.000
PS parameter									
π_n	0.40	0.402	0.028	0.930	1.000	0.402	0.028	0.930	1.000
π_f	0.40	0.397	0.040	0.936	1.000	0.397	0.040	0.936	1.000
π_b	0.00	0.000				0.000			
π_a	0.20	0.201	0.028	0.964	1.000	0.201	0.028	0.960	1.000
γ_n	0.00	−0.111	0.098	0.788		0.000			
γ_f	0.50	0.665	0.108	0.666		0.492	0.232	0.944	0.562
γ_b		−0.887	0.167						
γ_a	0.00	−0.111	0.098	0.788		0.000			

Note. SEM = structural equation modeling; *BK* model = Baron and Kenny's (1986) model; PS = principal stratification; *AIR* model = Angrist, Imbens, and Rubin's (1996) model; *a* = the effect of treatment assignment *X* on the mediator *M*; *b* = the effect of *M* on outcome *Y* conditioning on *X*; *c* = the effect of *X* on *Y* conditioning on *M*; *ab* = the indirect effect (= *a* × *b*); *d* = the interaction effect (how much the effect of *X* increases [or decreases] as the value of *M* increases by one unit); α_m = the mean of *M* in the control condition; π_n = the proportion of never-improvers; π_f = the proportion of forward-improvers; π_b = the proportion of backward-improvers; π_a = the proportion of always-improvers; γ_n = the average causal treatment effect for never-improvers; γ_f = the average causal treatment effect for forward-improvers; γ_b = the average causal treatment effect for backward-improvers; γ_a = the average causal treatment effect for always-improvers.

Table 9

When None of the SEM_{BK} and PS_{AIR} Model Assumptions Hold

		***SEM**$_{BK}$* **model**		***PS**$_{AIR}$* **model**		***SEM**$_{Mac}$(= **PS**$_{Ig}$)* **model**	
True value		**Estimate(*SE*)**	**Coverage[Power]**	**Estimate(*SE*)**	**Coverage[Power]**	**Estimate(*SE*)**	**Coverage[Power]**
SEM parameter							
a	0.30	0.297 (0.042)	0.942 [1.000]	0.297 (0.042)	0.942 [1.000]	0.297 (0.042)	0.942 [1.000]
b	0.30	0.617 (0.096)	0.096	0.562 (0.319)	0.864	0.294 (0.139)	0.950 [0.580]
c	−0.10	−0.016 (0.096)	0.840	0.000		−0.283 (0.126)	0.680
ab	0.09	0.183 (0.039)	0.304	0.167 (0.095)	0.860		
d	0.30	0.000		0.000		0.607 (0.191)	0.654
α_m	0.30	0.301 (0.030)	0.968 [1.000]	0.301 (0.030)	0.968 [1.000]	0.301 (0.030)	0.968 [1.000]
PS parameter							
π_n	0.30	0.402 (0.030)	0.066	0.402 (0.030)	0.066	0.402 (0.030)	0.066
π_f	0.40	0.297 (0.042)	0.324	0.297 (0.042)	0.324	0.297 (0.042)	0.324
π_b	0.10	0.000		0.000		0.000	
π_a	0.20	0.301 (0.030)	0.054	0.301 (0.030)	0.054	0.301 (0.030)	0.054
γ_n	−0.10	−0.016 (0.096)	0.840	0.000		−0.283 (0.126)	0.680
γ_f	0.50	0.602 (0.114)	0.874	0.562 (0.319)	0.944 [0.442]	0.618 (0.112)	0.810
γ_b	−0.40	−0.633 (0.155)	0.664			−0.577 (0.154)	0.752
γ_a	0.20	−0.016 (0.096)	0.392	0.000		0.324 (0.143)	0.864

Note. SEM = structural equation modeling; *BK* model = Baron and Kenny's (1986) model; *Mac* model = MacArthur model (Kraemer et al., 2008); PS = principal stratification; *AIR* model = Angrist, Imbens, and Rubin's (1996) model; *Ig* model = ignorability model; *a* = the effect of treatment assignment *X* on the mediator *M*; *b* = the effect of *M* on outcome *Y* conditioning on *X*; *c* = the effect of *X* on *Y* conditioning on *M*; *ab* = the indirect effect (= *a* × *b*); *d* = the interaction effect (how much the effect of *X* increases [or decreases] as the value of *M* increases by one unit); α_m = the mean of *M* in the control condition; π_n = the proportion of never-improvers; π_f = the proportion of forward-improvers; π_b = the proportion of backward-improvers; π_a = the proportion of always-improvers; γ_n = the average causal treatment effect for never-improvers; γ_f = the average causal treatment effect for forward-improvers; γ_b = the average causal treatment effect for backward-improvers; γ_a = the average causal treatment effect for always-improvers; Power is reported if the coverage rate is at least 0.9.