Published in final edited form as:

Multivariate Behav Res. 2011 May ; 46(3): 425–452. doi:10.1080/00273171.2011.576624.

The Use of Propensity Scores in Mediation Analysis

Booil Jo,

Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine

Elizabeth A. Stuart,

Department of Mental Health and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

David P. MacKinnon, and

Department of Psychology, Arizona State University

Amiram D. Vinokur

Institute for Social Research, University of Michigan

Abstract

Mediation analysis uses measures of hypothesized mediating variables to test theory for how a treatment achieves effects on outcomes and to improve subsequent treatments by identifying the most efficient treatment components. Most current mediation analysis methods rely on untested distributional and functional form assumptions for valid conclusions, especially regarding the relation between the mediator and outcome variables. Propensity score methods offer an alternative whereby the propensity score is used to compare individuals in the treatment and control groups who would have had the same value of the mediator had they been assigned to the same treatment condition. This article describes the use of propensity score weighting for mediation with a focus on explicating the underlying assumptions. Propensity scores have the potential to offer an alternative estimation procedure for mediation analysis with alternative assumptions from those of standard mediation analysis. The methods are illustrated investigating the mediational effects of an intervention to improve sense of mastery to reduce depression using data from the Job Search Intervention Study (JOBS II). We find significant treatment effects for those individuals who would have improved sense of mastery when in the treatment condition but no effects for those who would not have improved sense of mastery under treatment.

Researchers are increasingly interested in investigating not just whether a treatment or intervention has an overall effect on an outcome variable but also in how treatments achieve their effects. For example, in the Job Search Intervention Study (JOBS II) study described more later, the intervention was designed to increase sense of personal mastery, which was hypothesized to then lead to increased subsequent reemployment. In an anabolic steroid prevention program, the underlying theory was that changing a social norm to enhance dietary methods to increase strength would lead to improved diet (MacKinnon et al., 2001).

© Taylor & Francis Group, LLC

Correspondence concerning this article should be addressed to Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health, Department of Mental Health, 624 N. Broadway, 8th Floor, Baltimore, MD 21205. estuart@jhsph.edu.

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Mediation analysis (Jo, 2008; MacKinnon, 2008) is a key statistical tool in identifying the processes by which a treatment affects an outcome. However, most current mediation analysis methods rely on untested distributional and functional form assumptions for valid conclusions, especially regarding the relation between the mediator and outcome variables. In this article we propose the use of propensity scores as an alternative estimation technique for mediation, allowing the estimation of mediation effects under different assumptions, which may be more believable in some settings. By its nature, mediation analysis almost always involves untestable assumptions; what we propose here is an addition to the mediation toolbox.

A particular challenge of mediation analysis is that mediator values are only observed under the treatment condition to which each individual is assigned. Propensity scores help get around this issue by matching individuals in one treatment group, with an observed mediator value, to individuals in another treatment group who look as if they would have had that value of the mediator had they been in the other treatment group. For example, in the motivating example, we match individuals whose sense of mastery improved when in the treatment group to individuals in the control group who look as if their sense of mastery would have improved had they been in the treatment group. The proposed propensity score-based approach relies on the framework of principal stratification, which defines groups (strata) of individuals with the same values of a posttreatment variable, such as a mediator. Principal stratification methods highlight the danger in conditioning on an observed mediator because that destroys the comparability induced by randomization: individuals with a particular value of a mediator when in the treatment group are likely different from those individuals who would have that same value of the mediator when in the control group. Principal stratification, in contrast, estimates effects within subgroups defined by the pair of potential mediator values in the two conditions. Propensity scores have been proposed (Hill, Waldfogel, & Brooks-Gunn, 2003; Jo & Stuart, 2009) to estimate principal strata effects for one particular type of principal stratification—noncompliance in a randomized experiment—but their use in principal stratification more generally, or in mediation in particular, has not been examined.

MOTIVATING EXAMPLE: THE JOBS II STUDY

The Job Search Intervention Study (JOBS II; Vinokur, Price, & Schul, 1995) was a randomized field experiment intended to prevent poor mental health and to promote high-quality reemployment among unemployed workers. The JOBS II project was among recent efforts to develop a theory-driven intervention that could provide an empirical basis for identifying hypothesized mediators. The mediating role of sense of mastery (Vinokur & Schul, 1997), which was particularly emphasized in the study, can be a key to the success of a wide range of interventions that are intended to prevent poor mental health or to achieve difficult cognitive and behavioral changes (Meichenbaum, 1985; Ozer & Bandura, 1990; Wood & Bandura, 1989). Therefore, sense of mastery was targeted as an active ingredient that might be triggered by the intervention and have an influence on mental health outcomes such as depression.

The treatment condition consisted of five half-day training sessions that focused on the application of problem-solving and decision-making processes, inoculation against setbacks, provision of social support and positive regard from trainers, and learning and practicing job search skills. The control condition consisted of a booklet briefly describing job search methods and tips. We focus on the high-risk group based on previous studies (Price, van Ryn, & Vinokur, 1992; Vinokur et al., 1995) that indicated that the job search intervention had its primary impact on high-risk individuals. In this example, we focus on depression as the outcome, which was one of the primary outcomes of the JOBS II study. Specifically, the

outcome used here is the change score for depression (Y : range $D = -2.64$ to 1.91), which was calculated by subtracting the depression score measured 6 months after the intervention (range = 1.00 to 4.73) from the baseline depression score (range = 1.73 to 3.00). A larger value means a more beneficial change over time. A change score for sense of mastery (M : range = -1.21 to 1.55) was calculated by subtracting the baseline mastery score (range = 1.88 to 4.85) from the mastery score measured 2 months after the intervention (range = 2.11 to 4.84). Again, a larger value means a more beneficial change over time. To focus on modeling of mediational processes, only the cases with complete data were included in the analyses. A total sample of 422 cases (treatment = 278, control = 144) was analyzed after deleting cases with missing outcome and/or missing pretreatment data (293 cases out of 715 were excluded). Although this sort of complete case analysis is not a proper way of dealing with missingness, we use it here for simplicity and illustration purposes. The methods described in this article could be fairly easily implemented within a multiple imputation context as an alternative method for dealing with the missingness. Table 1 shows descriptive statistics for the sample and variables used in this article.

On average, the level of depression improved (i.e., decreased) over time in both the treatment and control groups. On the basis of an intention-to-treat (ITT) analysis (two-sample t test, two-tailed) without conditioning on covariates, there was no significant difference in the change in depression (Y) across the treatment and control groups (ITT effect estimate = 0.130 , $SE = 0.080$, p value = $.107$, effect size = 0.166 on the basis of the pooled SD of Y). When we condition on covariates (i.e., analysis of covariance [ANCOVA]), the difference in the change in depression (Y) across the treatment and control groups was significant (ITT effect estimate = 0.198 , $SE = 0.074$, p value = $.008$, effect size = 0.253), indicating a positive impact of the treatment on depression. Sense of mastery showed significant differences both in an unconditional analysis (ITT effect estimate = 0.151 , $SE = 0.047$, p value = $.001$, effect size = 0.331 on the basis of the pooled SD of M) and in an analysis conditional on covariates (ITT effect estimate = 0.130 , $SE = 0.041$, p value = $.001$, effect size = 0.285), indicating a positive impact of the treatment on sense of mastery. No significant differences were found in pretreatment covariates (W_1 – W_9) across the treatment and control groups.

STRUCTURAL EQUATION MODELING APPROACH TO MEDIATION

Data from randomized trials with embedded mediational processes have been largely analyzed using the structural equation modeling (SEM) approach (Baron & Kenny, 1986; Bollen, 1987; Judd & Kenny, 1981; MacKinnon & Dwyer, 1993), 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), which in turn changes the final outcome (Y).

In this framework, the mediated effect of treatment assignment on the outcome (indirect effect) is evaluated in terms of whether the treatment-targeted mediators are highly associated with the outcome (i.e., the effect b is sizable and significant) and whether the treatment successfully improved the mediator status of individuals (i.e., the effect a is sizable and significant). Ultimately, the indirect effect is calculated as a product of these two effects (ab). Although the indirect effect is of main interest, this approach does not exclude the possibility that the treatment may directly influence outcomes without going through mediators (direct effect c'). Direct effects are likely present in many situations because 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 regressions. For individual i , the observed mediator M is regressed on treatment assignment X :

$$M_i = \alpha_m + a X_i + \varepsilon_{mi}, \quad (1)$$

where α_m is the intercept, a is the effect of X on M , and ε_{mi} is the residual, which is assumed to be normally distributed. For individual i , the observed outcome Y is regressed on treatment assignment X and the observed mediator M :

$$Y_i = \alpha_y + b M_i + c' X_i + \varepsilon_{yi}, \quad (2)$$

where α_y is the intercept, b is the effect of M on Y conditioning on X , and c' is the effect of X on Y conditioning on M . The residual ε_{yi} is assumed to be normally distributed.

It is also a common practice in the SEM approach to include pretreatment covariates in the analyses. Let W be a vector of covariates. Conditioning on W , the two linear regressions in Equations (1) and (2) are simply modified as

$$M_i = \alpha_m + a X_i + \beta_m' W_i + \varepsilon_{mi}, \quad (3)$$

$$Y_i = \alpha_y + b M_i + c' X_i + \beta_y' W_i + \varepsilon_{yi}, \quad (4)$$

where β_m is the vector of regression coefficients that represents the association between M and W conditional on X . The vector of regression coefficients β_y represents the association between Y and W conditional on M and X . In this model, interpretations of the key effects (a , b , c') are modified as effects conditional on covariates.

ASSUMPTIONS OF THE STRUCTURAL EQUATION MODELING APPROACH

Although the effect of treatment assignment on mediator status (a) can be interpreted as causal on the basis of random assignment of treatment conditions, causal interpretation is not readily warranted for the other two components of mediation effects defined in the SEM approach (Holland, 1988). These effects include the effect of the mediator on the outcome (b) and the direct effect of treatment assignment on the outcome that is not mediated by the intended mediator (c'). Causal interpretation is not warranted for b because M is a posttreatment variable affected by treatment assignment X . Causal interpretation is not warranted for c' because this is the effect of X on Y conditional on M , and M is a posttreatment variable affected by treatment assignment X . For example, in JOBS II, individuals who improved their sense of mastery 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 (i.e., same M value). 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. A particular set of assumptions is necessary to identify these effects and to interpret them as causal mediation effects in the SEM approach. Here we briefly summarize these identifying assumptions. For further details, see Imai, Keele, and Tingley (2010); Jo (2008); and Sobel (2008).

Assumption SEM-1. Ignorability of Observed Mediator Status (M)

Within each treatment group, individuals with different M have the same characteristics and therefore can be compared. This makes causal interpretation of the b effect possible. It also means that across treatment groups, individuals with the same M have the same characteristics and thus treated and control individuals with the same M can be compared. This makes causal interpretation of the direct effect of X conditional on M (i.e., c' effect) possible. Ignorability of M is the most critical assumption in the SEM approach but unfortunately is generally untestable. This assumption can be somewhat relaxed by allowing heterogeneity among individuals in terms of observed characteristics (i.e., ignorability conditional on observed variables). Conditional ignorability is weaker than ignorability, although it is still an untestable assumption.

Assumption SEM-2. Constant Effect

The interaction between treatment assignment (X) and the mediator (M) has no effect on the outcome. Assuming that ignorability holds, this assumption plays a key role in identifying the unique direct (c') and indirect (ab) effects of treatment assignment in the SEM approach. This assumption is untestable as long as the ignorability assumption remains untestable.

Assumption SEM-3. Linearity

The outcome value linearly increases (or decreases) as the mediator value increases (or decreases). 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. This assumption is also untestable as long as the ignorability assumption remains untestable.

Application of the SEM Approach to JOBS II

We analyzed the JOBS II data according to the structural equation model described in Equations (3) and (4). First we analyzed the data treating the change score of sense of mastery as continuous and then reanalyzed the data after dichotomizing the change score of sense of mastery at its median. The latter analysis is not a necessary step in the SEM approach if the aforementioned three assumptions hold but was conducted to help our comparison of the results between the SEM approach and principal stratification approaches, which are discussed in the following sections. In addition, without knowing whether ignorability holds, linearity becomes an additional untestable assumption. In other words, to accommodate continuous mediators, we need to impose more and stronger assumptions that are unverifiable, which makes causal inference modeling more difficult (see Jo, 2008, for more details). By using a binary mediator, we focus on comparing the SEM and the principal stratification approach in a relatively simple setting with a reduced number of untestable assumptions. With the dichotomized mediator, we used the same linear model shown in Equation (3) for the purposes of comparison, although it is possible to use a logistic or probit regression model instead. We employed a maximum likelihood (ML) estimation method using Mplus (Muthén & Muthén, 1998–2010), although simpler estimation methods (e.g., ordinary least squares) will yield similar results. Standard errors of the ab estimates were calculated using the delta method implemented in Mplus (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Sobel, 1982). All nine covariates in Table 1 are included in each model in the article, but for each model we report only the covariates that are significant predictors of the outcome or the principal stratum membership. As an approximate way of maintaining the interpretation of intercepts in the outcome model as outcome means instead of as intercepts (so that they can be easily compared between different methods), we centered covariates at their observed means.

Table 2 summarizes the results of mediation analysis in JOBS II using the SEM framework. The results show that the a , b , and c' estimates are significantly different from zero whether M is continuous or binary. The indirect effect ab is significant when the mediator is continuous but not significant when it is dichotomized. Except for some loss of power with the dichotomized M , the two results are quite consistent and provide a good picture of how X , M , and Y are related. First of all, the treatment was successful at promoting sense of mastery, providing evidence of action theory for how the intervention affects the mediator. We can also conclude that the sense of mastery was a well-theorized mediator in the sense that it is a very good predictor of the outcome, providing evidence for conceptual theory. The significant c' estimate can be interpreted as that there is a remaining direct effect of treatment assignment on the outcome and/or that there may be additional mediators not included in our model. Table 2 also shows the effects of baseline covariates on the mediator and on the outcome. Baseline sense of mastery is significantly associated with the change in sense of mastery. That is, the higher the baseline mastery score, the smaller the change. Baseline depression and sense of mastery were good predictors of change in depression after controlling for the effects of X and M . Individuals with higher scores on these baseline variables improved more in terms of their depression.

Whether the mediation effect estimates reported in Table 2 can be considered good representations of causal effects depends on the plausibility of the assumptions used to identify these effects. Based on random assignment to treatment and control conditions, the a estimate can be interpreted as causal. Causal interpretation of b depends on the degree of deviation from the ignorability assumption. Because the direct effect of X on Y is estimated conditioning on M , causal interpretation of the c' estimate is also subject to bias depending on the degree of deviation from ignorability. The indirect effect ab additionally requires the constant effect assumption. Further, linearity is also assumed if M is continuous. Because all three assumptions (ignorability, constant effect, linearity) are closely connected in identifying the direct and indirect effects, finding consistent patterns of bias and conducting sensitivity analysis can be a challenging task.

PRINCIPAL STRATIFICATION APPROACH

The second approach we consider for mediation analysis is based on principal stratification (PS; Frangakis & Rubin, 2002). In defining and identifying treatment or mediation effects, the PS approach uses the concept of potential outcome values for any variables that may be affected by treatment assignment, including the mediator M and the outcome Y . Potential outcomes may be considered as latent or unobserved variables. However, unlike latent variables commonly conceptualized in SEM, the latent variables are defined considering potential status under a particular or all treatment conditions that are compared. The latent variables in SEM do not necessarily correspond to potential outcomes in this way.

To formalize these ideas, assume a randomized trial, where treatment assignment (X) has two values (1 = treatment, 0 = control). In this setting, each individual is assigned either to the treatment or to the control condition. 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 way of defining treatment effects at the individual level on the basis of potential outcomes is often referred to as the Rubin causal model, or more broadly as the potential outcomes approach (Holland, 1986; Neyman, 1923; Rubin, 1978). Although in practice we can observe only one of $Y_i(1)$ and $Y_i(0)$ for each individual, which makes it impossible to identify individual-level causal effects, average causal effects across individuals can be estimated under certain conditions.

Two assumptions underlie all of the methods discussed in this article: (1) ignorable treatment assignment and (2) stable unit treatment value assumption. These assumptions are usually more explicit in the principal stratification approach, but they are also necessary for causal interpretation of effect estimates in the SEM approach, even though explicit discussions of these assumptions are rare.

Assumption Common-1. Ignorable Treatment Assignment

Treatment assignment is independent of the potential outcomes, given the covariates. This assumption is automatically satisfied in randomized experiments such as JOBS II; its validity is thus known in experiments but is unknown (and untestable) in non-experimental studies.

Assumption Common-2. Stable Unit Treatment Value Assumption (SUTVA)

This assumption has two components: (a) there is only one version of each treatment condition and (b) there is no interference between units (individuals), that is, one individual's treatment assignment does not affect another individual's potential outcomes (Rubin, 1978, 1980). Assumption Common-2a can usually be dealt with by redefining the treatment to be the aspect of the treatment that is constant across individuals in the treatment group (e.g., the opportunity to participate in the JOBS II intervention rather than level of involvement in the intervention). Statistical adjustment for violation of assumption Common-2b is also possible when interaction among individuals happens within clusters assigned to the same treatment condition (Jo, Asparouhov, & Muthén, 2008). However, interaction among individuals across different treatment conditions is a very difficult problem to handle; it is usually dealt with as much as possible in the design stage, designing the study to limit interactions between participants (Sobel, 2006).

In JOBS II, we assume that Common-2a is satisfied, in that every individual in the treatment group received the same treatment. In addition, study participants were sampled from six major unemployment offices serving the entire Detroit area, and therefore substantial deviation from Common-2b is very unlikely. Although either Common-2a or Common-2b may have been violated in part given that the treatment was conducted in a group setting, we assume that this deviation has little impact on our effect estimates.

Mediation in the PS Framework

The most fundamental difference between the PS and SEM approaches in identifying mediation effects is that the PS approach focuses on the potential mediator status (we use S for this), whereas the SEM approach focuses on the observed mediator status (M). PS provides a way to reflect the causal relation among treatment assignment, mediator, and outcome in the potential outcomes approach, although other methods are also possible (e.g., Ten Have et al., 2007).

PS refers to classification of individuals based on potential values of post-treatment variables, such as mediators, under all treatment conditions that are compared. In general, as discussed earlier, a causal interpretation is not warranted for a simple comparison of treated and control individuals with the same observed mediator values because the mediator value obtained under the treatment is not comparable to the mediator value obtained under the control condition. Therefore, the difference in the outcome between the treatment and control conditions simply conditioning on observed mediators does not necessarily represent a causal effect.

The main goal in the PS approach is to achieve causal interpretability by considering the potential mediator values. Assuming two possible treatment assignment statuses ($1 =$

treatment, 0 = control) and two potential mediator values (1 = improved, 0 = not improved), four possible principal strata are defined, as discussed in Angrist, Imbens, and Rubin (1996). A new notation C_i is used to represent which principal stratum each individual belongs to. That is,

$$C_i = \begin{cases} n \text{ (never-improver)} & \text{if } S_i(0)=0, \text{ and } S_i(1)=0 \\ f \text{ (forward-improver)} & \text{if } S_i(0)=0, \text{ and } S_i(1)=1 \\ b \text{ (backward-improver)} & \text{if } S_i(0)=1, \text{ and } S_i(1)=0 \\ a \text{ (always-improver)} & \text{if } S_i(0)=1, \text{ and } S_i(1)=1. \end{cases}$$

In line with Jo (2008), we label these four types of individuals 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. Among these four types, forward- and backward-improvers potentially change their mediator status in response to treatment assignment, and therefore treatment effects for these two types may include indirect effects (Rubin, 2004; 2005). This way of defining indirect effects is technically different from that of Baron and Kenny's (1986) approach, although what the two definitions try to capture is conceptually consistent.

The aforementioned categorization can be simplified if we consider potential mediator values under only one assigned condition (Jin & Rubin, 2008; Jo, Wang, & Ialongo, 2009; Joffe, Small, & Hsu, 2007). In line with Jo et al. (2009), we use reference stratification to refer to this type of PS. In JOBS II, the treatment effect for those who improved their sense of mastery when receiving the intervention (and for those who did not) is of primary interest. In this case, the treatment condition serves as the reference condition. In other situations, it may be of more interest to identify causal effects for those who would and would not improve their mediator status under the control condition. In this case, the control condition serves as the reference condition.

Using the treatment condition as the reference condition, the four mediator improvement types (principal strata) can be combined as

$$C_{1i} = \begin{cases} 0 \text{ (treatment nonimprover: } n+b) & \text{if } S_i(1)=0, \text{ and } S_i(0)=0 \text{ or } 1 \\ 1 \text{ (treatment improver: } f+a) & \text{if } S_i(1)=1, \text{ and } S_i(0)=0 \text{ or } 1, \end{cases}$$

where C_{1i} is the stratum membership for individual i when stratification is based on potential values of the mediator status under the treatment condition only. The stratum membership C_{1i} is fully observed in the treatment condition.

Using the control condition as the reference condition, the four mediator improvement types (principal strata) can be combined as

$$C_{0i} = \begin{cases} 0 \text{ (control nonimprover: } n+f) & \text{if } S_i(0)=0, \text{ and } S_i(1)=0 \text{ or } 1 \\ 1 \text{ (control improver: } b+a) & \text{if } S_i(0)=1, \text{ and } S_i(1)=0 \text{ or } 1, \end{cases}$$

where C_{0i} is the stratum membership for individual i when stratification is based on potential values of the mediator status under the control condition only. The stratum membership C_{0i} is fully observed in the control condition.

Although reference stratification considering potential mediator values under only one condition (as opposed to under both treatment conditions) is not as detailed as the usual PS, we chose to take this route for two reasons. First, in the PS method that uses propensity scores, stratum membership needs to be observed for individuals in at least one of the assigned conditions (which is the case for C_{1i} and C_{0i}). Second, as the number of strata increases, more and stronger assumptions are usually necessary to identify causal effects conditioning on the strata. For example, in the two-strata setting described earlier, we can identify the causal treatment effect for one stratum of individuals by assuming no treatment assignment effect for the other, which is a commonly used assumption called the exclusion restriction (Angrist et al., 1996). To identify the causal treatment effect for one particular stratum in the four-strata setting, further restrictions are necessary. For example, Angrist et al. identified the causal treatment effect for the forward-improvers (“compliers,” in their terminology) by imposing the exclusion restriction for never-improvers (“never-takers”) and also for always-improvers (“always-takers”). Further, an additional assumption called monotonicity is imposed, which does not allow the existence of backward-improvers (“defiers”). These assumptions can be unrealistic when dealing with mediators, although they may be relatively more plausible in some special cases, for example, when the mediator is treatment receipt status. See Jo (2008) for further discussion of these assumptions in the mediation context.

Assume a two-strata setting considering the treatment as the reference condition. For individual i , the probabilities of belonging to the treatment improver and treatment nonimprover strata given a vector of pretreatment covariates W can be expressed using logistic regression as

$$\begin{aligned} P(C_{1i}=1|W_i) &= \pi_{1i}, \\ P(C_{1i}=0|W_i) &= 1 - \pi_{1i}, \\ \text{logit}(\pi_{1i}) &= \beta_{0C1} + \beta_{1C1} W_i, \end{aligned} \quad (5)$$

where π_{1i} is the probability that individual i would improve his or her mediator status under the treatment condition, β_{0C1} is the logit intercept, and β_{1C1} is a vector of logit coefficients that reflects the association between stratum membership C_{1i} and pretreatment covariates W . The average probability Π_1 is the proportion of mediator improvers in the treatment condition.

A continuous outcome Y for individual i with treatment assignment status X and mediator improvement type C_{1i} is then defined as

$$Y_i = \alpha_{0C1} + (\alpha_{1C1} - \alpha_{0C1}) C_{1i} + \gamma_{0C1} X_i + (\gamma_{1C1} - \gamma_{0C1}) C_{1i} X_i + \lambda_{0C1} W_i + (\lambda_{1C1} - \lambda_{0C1}) C_{1i} W_i + \varepsilon_{iC1}, \quad (6)$$

where α_{1C1} and α_{0C1} are the intercepts for those who would ($C_{1i} = 1$) and would not ($C_{1i} = 0$) improve their mediator status if the treatment is given. The difference, $\alpha_{1C1} - \alpha_{0C1}$, can be interpreted as the difference in the potential outcome under control between treatment improvers and treatment nonimprovers with the same value of W . If the mean of W is zero, α_{0C1} and α_{1C1} can be interpreted as the mean potential outcomes under control for the treatment improvers and treatment nonimprovers. The relationship between the covariates and the outcome is expressed by λ_{0C1} for treatment nonimprovers and λ_{1C1} for treatment improvers. The average effect of treatment assignment is λ_{0C1} for treatment nonimprovers and λ_{1C1} for treatment improvers. The residual ε_{iC1} is assumed to be normally distributed

with mean zero and variance σ_{0C1}^2 for treatment nonimprovers and σ_{1C1}^2 for treatment improvers. However, parametric assumptions such as normality are not essential for identifying causal treatment effects in the PS models discussed in this article.

We do not repeat the formulation of the principal stratification model using the control condition as the reference condition here. This can be done by replacing C_1 with C_0 in Equations (5) and (6).

The PS model described in Equations (5) and (6) is used for causal mediation analysis in this article. As shown in this model, the analytical structure of the PS approach is quite different from that of the SEM approach in identifying mediation effects. An important aspect of this approach is that the potential mediator improvement type C_1 or C_0 is unaffected by treatment assignment, and therefore it can be treated as a pretreatment covariate in the analysis. In other words, within each category of C_1 or C_0 , a comparison between treatment and control groups leads to a treatment effect estimate that can be interpreted as causal (just like a conventional ITT subgroup analysis). According to this framework, mediation is seen as having different effects of treatment assignment on the outcome across different potential mediator types (i.e., X by C_1 or X by C_0 interaction effect). In that sense, the perspective on mediation taken by the PS approach is similar to that of the approaches that include interaction effects (Judd & Kenny, 1981; Kraemer, Kiernan, Essex, & Kupfer, 2008). In principle, however, it is possible to take both perspectives (main/interaction or direct/indirect) and to translate results across them (Jo, 2008). Figure 2 illustrates the conceptual framework of mediation in the PS approach based on interaction between treatment assignment and the potential mediator improvement type (C , C_0 , or C_1).

Along with categorizing the potential mediator values, defining mediation as an interaction effect considerably helps the PS approach depend less on multiple identifying assumptions. In this setting, two identifying assumptions in the SEM approach, constant effect (owing to definition of mediation as an interaction) and linearity (owing to binary M), are unnecessary. Based on the two strata setting, which we focus on in this article, we discuss two PS methods that rely on one identifying assumption. Depending on the identifying assumption imposed, the mediator strata model shown in Equation (5) and the outcome model shown in Equation (6) are either separately (propensity score model) or jointly (joint model) estimated.

METHODS FOR ESTIMATING MEDIATIONAL EFFECTS IN THE PS FRAMEWORK

We discuss two methods for estimating mediational effects in the PS framework: propensity score-based methods and joint estimation methods. These methods rely on somewhat different assumptions, which also differ from the assumptions underlying the SEM approach, and thus the two methods can provide complementary information.

Propensity Score Methods

Propensity scores, first introduced by Rosenbaum and Rubin (1983), have traditionally been used in nonexperimental studies to help balance the treated and control groups on a set of confounders. The propensity score itself is defined as the probability of receiving the treatment given the observed covariates. Propensity scores are often estimated using logistic regression or nonparametric methods such as generalized boosted models (McCaffrey, Ridgeway, & Morral, 2004); the propensity scores themselves are the predicted probabilities of treatment generated from those procedures. The propensity scores are then used in matching, weighting, or subclassification to create treated and control groups that look only

randomly different from one another, at least with respect to the covariates used to estimate the propensity score (Stuart, 2010). Propensity score methods (and their use here, as we detail later), thus consist of two separate stages: first, estimation and application of the propensity score, and second, use of those propensity scores in estimating treatment effects on the outcomes of interest.

The propensity scores' ability to create well-matched groups come from the two key features of propensity scores, detailed in Rosenbaum and Rubin (1983). First, propensity scores are what are called “balancing scores,” which means that among individuals with similar propensity score values, the distribution of observed covariates will be similar between the treated and control groups. The second feature is that, if treatment assignment is unconfounded (i.e., does not depend on the outcomes) given the covariates, then treatment assignment is also unconfounded given just the propensity score. This is what enables collapsing all of the observed covariates into the scalar propensity score and using that for matching, weighting, or subclassification, rather than having to deal with each of the covariates individually. These two features imply that a comparison of outcomes among individuals with similar propensity scores can yield an unbiased estimate of the treatment effect, assuming no unmeasured confounders.

In this article we propose the use of propensity scores to model principal stratum membership, where the principal strata represent groups defined by mediators. This is useful for assessing mediation relations because the mediator value is not randomized. In particular, the propensity score will model mediator values under one treatment condition, which in the reference stratification framework corresponds to reference stratum membership (e.g., treatment improvers or nonimprovers). Following Hill et al. (2003), we will use the term “principal propensity scores” hereafter to refer to propensity scores that model principal stratum membership, for example, $P(C_1 = 1|W)$.

The crucial assumption underlying the use of propensity scores to assess mediation is principal ignorability:

Assumption PS-1. Principal Ignorability (PI)—Principal stratum membership is independent of the potential outcomes given the observed covariates (Jo & Stuart, 2009). When using the control condition as the reference condition, this assumption applies only to the potential outcome under treatment (C_{0i} is unobserved only in the treatment group): $Y_i(1) \perp C_{0i} | W_i$, meaning that there are no differences in the outcome across reference strata $C_0 = 0$ and $C_0 = 1$ given the observed covariates. The potential outcome under the control condition is allowed to vary across the two strata $C_0 = 0$ and $C_0 = 1$. When using the treatment condition as the reference condition, this assumption applies only to the potential outcome under control (C_{1i} is unobserved only in the control group): $Y_i(0) \perp C_{1i} | W_i$. Like the other ignorability assumptions considered in this article (SEM-1 and Common-1), this assumption is inherently untestable.

In more intuitive terms, principal ignorability assumes that we can use the observed covariates to identify the likely “treatment improvers” (and nonimprovers) in the control group and the likely “control improvers” (and nonimprovers) in the treatment group. As defined earlier, the principal ignorability assumption involves outcomes, although estimation of principal propensity scores does not use any outcome information. The use of outcome information is restricted to the second stage of the propensity score method. This assumes that the observed covariates are sufficient for identifying stratum membership. This is a crucial assumption, in parallel with the standard ignorability of treatment assignment assumption made commonly in the traditional propensity score context (expressed here as assumption Common-1). Unfortunately principal ignorability is untestable, although it can

be made more believable by including a large set of predictors in the propensity score model, especially those that (in this context) are particularly predictive of the mediator status. See Jo and Stuart (2009) and Stuart and Jo (in press) for further information on the sensitivity of principal causal effect estimates to deviation from principal ignorability.

There is a subtle but important implication of using the principal ignorability assumption in the context of reference stratification. Reference stratification collapses the four principal strata into two. In the case of stratification using C_0 , the group with $C_0 = 0$ consists of never-improvers and forward-improvers whereas the group with $C_0 = 1$ consists of backward-improvers and always-improvers. At the same time, principal ignorability assumes that $Y_i(1) \perp C_{0i} | W_i$, which essentially implies that the average outcome under treatment is the same for the $C_0 = 0$ and $C_0 = 1$ groups given W . In other words, this assumes that the average outcome under treatment for the combined never-improver and forward-improver group is equal to the average outcome under treatment for the combined backward-improver and always-improver group. Although it may be very reasonable to assume that the average outcome under treatment would be equal for never-improvers and backward-improvers (because neither group would improve under treatment) given W , and for forward-improvers and always-improvers (because both groups would improve under treatment) given W , it is not clear that the averages within reference strata would be equal given possibly different mixing proportions. The same restriction would hold for the reference strata defined using C_1 , but there the relevant outcome is the potential outcome under control. Implications of this are further discussed later.

The principal propensity score method takes advantage of random assignment to treatment and control groups. Because of randomization, we can expect that the composition of principal strata (mediator improvement types) and the relationship between principal stratum membership and covariates will be the same across treatment groups. Given this condition, principal propensity scores $P(C_1 = 1 | W)$ can be predicted for the control group using the relationship between covariates and the observed mediator in the treatment group. Similarly, principal propensity scores $P(C_0 = 1 | W)$ can be predicted for the treatment group using the relationship between covariates and the observed mediator in the control group. The specific steps we employ to estimate principal propensity scores and principal effects in this article are described here.

We describe the approach for estimating causal treatment effects for treatment improvers ($C_1 = 1$) and treatment nonimprovers ($C_1 = 0$). Note that when estimating the treatment effect for treatment improvers, the treatment nonimprovers in the treatment condition are not used, and likewise, when estimating the treatment effect for treatment nonimprovers, the treatment improvers in the treatment condition are not used. The same methods can be applied to estimate causal treatment effects for control improvers ($C_0 = 1$) and control nonimprovers ($C_0 = 0$).

Step 1: We first fit the principal propensity score model using the treatment group members only, predicting the potential mediator improvement type C_1 given the covariates. The logistic regression model shown in Equation (5) is used for this purpose.

Step 2: From the model estimated in Step 1, predicted probabilities of being a treatment improver (i.e., principal propensity scores), P , are generated for treatment and control group members.

Step 3: Apply the principal propensity scores using methods such as matching, weighting, or subclassification to equate the groups. In this article we illustrate the approach using a weighting technique. In particular, treatment improvers in the treatment group receive a weight of 1, so that they represent themselves. Treatment nonimprovers in the treatment

group are not used when estimating γ_{1C1} and thus receive a weight of 0. The control group members each receive a weight of $w = P$. This weighting serves to make the control group look like the set of treatment improvers. Control group members with small values of P (low probabilities of being a treatment improver) will receive small weights. Control group members with large values of P (and who thus have characteristics similar to the treatment improvers in the treatment group) will receive larger weights. The principal causal effect γ_{1C1} is then estimated using a weighted regression model of outcome on treatment status and the covariates. This is the first time that the outcome values are used, and doing this regression adjustment combined with propensity score matching or weighting has been shown to have very good performance (Ho, Imai, King, & Stuart, 2007; Robins & Rotnitzky, 2001). The same procedure is used to estimate γ_{0C1} except that the treatment nonimprovers in the treatment group receive a weight of 1, treatment improvers in the treatment group receive a weight of 0, and the control group members receive a weight of $(1 - P)$. This weighting serves to make the control group members weight up to the group of treatment nonimprovers.

There have been some concerns expressed about weighting approaches such as inverse probability of treatment weighting and weighting by the odds (Schafer & Kang, 2008) because they can be sensitive to the propensity score estimation. In this article we focus on weighting for illustrative purposes and because Jo and Stuart (2009) found that it worked well when estimating complier average causal effects, but other matching methods may be more appropriate for other data sets and settings. See Jo and Stuart for an example of using full matching to estimate complier average causal effects and Stuart (2010) for a discussion of matching methods more generally.

Joint Estimation

The propensity score methods have a two-step process of (1) principal propensity score estimation and weighting, and (2) effect estimation using weighted regression adjustment. The method does not require a large joint parametric model of covariates, mediator improvement type, and outcome. In contrast, more commonly used principal effect estimation methods use joint estimation by simultaneously modeling the mediator type (principal stratum membership) and outcomes to estimate principal causal effects. We use the term “joint estimation methods” to refer to these methods. These methods require explicit assumptions about the relationships between principal stratum membership, outcomes, and covariates. In this article, we employ a joint estimation method using a commonly used assumption called the exclusion restriction (Angrist et al., 1996).

Assumption PS-2. Exclusion Restriction—In the four-strata setting defined earlier, this assumption means that for never-improvers and always-improvers, there is no effect of treatment assignment on the outcome. In other words, the effect of treatment assignment on the outcome is completely mediated through the intended mediator (no direct effect of treatment assignment on the outcome). In a two-strata setting, this assumption applies to one of the two coarse strata. Because all reference strata contain at least one stratum of individuals who would change their improvement status in response to treatment assignment, the exclusion restriction no longer holds its pure meaning that there is no direct effect. However, the rationale behind the exclusion restriction is still similar in the sense that we assume that forward-improvers' outcomes are affected the most from treatment assignment. Under this common belief, we can impose the exclusion restriction on one of the reference strata that does not include forward-improvers. For example, when using the control as the reference condition, the exclusion restriction is likely to hold if treatment assignment has a minimal impact on always-improvers and backward-improvers. When using the treatment as the reference condition, the exclusion restriction is likely to hold if

treatment assignment has a minimal impact on never-improvers and backward-improvers. As in the conventional exclusion restriction assumption, some violation is possible due to residual effects of treatment assignment on the outcome mediated by other mediators. It is also possible that 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 such as depression). Further, treatment assignment may have some negative effect on backward-improvers.

Note that the exclusion restriction assumption replaces principal ignorability in the joint estimation method given that we do not need both assumptions to identify principal causal effects. For the joint estimation method, we use maximum likelihood estimation using the expectation-maximization algorithm (ML-EM), treating the unknown principal stratum membership (mediator improvement type) as missing data (C_1 is unknown among individuals assigned to the control, and C_0 is unknown among individuals assigned to the treatment). Details on identification of principal causal effects in the two-strata setting and the use of ML-EM can be found in several places (e.g., Jo, 2002; Little & Yau, 1998) and therefore is not repeated here.

In the outcome model shown in Equation (6), the exclusion restriction means that $\gamma_{0C1} = 0$. Under the exclusion restriction, the outcome model in (6) can be written as

$$Y_i = \alpha_{0C1} + (\alpha_{1C1} - \alpha_{0C1}) C_{1i} + \gamma_{1C1} C_{1i} X_i + \lambda_{0C1} W_i + (\lambda_{1C1} - \lambda_{0C1}) C_{1i} W_i + \varepsilon_{iC1}, \quad (7)$$

where the exclusion restriction is a sufficient assumption to identify the causal treatment effect for the treatment improver stratum (i.e., γ_{1C1}).

APPLICATION OF THE PS APPROACH TO JOBS II

We analyzed the JOBS II data based on the PS approach described in Equations (5) and (6). The same dichotomized observed sense of mastery M used in the SEM analysis reported in Table 2 is used here as a basis for formulating potential mediator improvement types, where, as described earlier, we use reference stratification to define four types of individuals: treatment improvers and nonimprovers and control improvers and nonimprovers.

Tables 3 and 4 show the estimated effects for the principal propensity score and joint model methods. In each table, the top panel shows the effects estimated using the principal propensity score approach and the bottom panel shows the same effects but estimated using the joint estimation method. The principal propensity score methods were implemented using R Version 2.7.1 (R Development Core Team, 2008) and the joint methods were run in Mplus Version 6.1 (Muthén & Muthén, 1998–2010).

Table 3 shows that treatment improvers benefited from being assigned to the treatment condition, although the size of the estimated effect varies somewhat, with the joint method implying a larger effect than the principal propensity score approach. That is, the treatment assignment led to a considerable decrease in depression levels (effect size = 0.35 [principal propensity score method] and 0.61 [joint model method], based on the pooled standard deviation of 0.78) for those who would have improved sense of mastery when the treatment is given. Baseline depression and sense of mastery were found to be significant predictors of the outcome for treatment improvers and nonimprovers. Baseline depression was also a strong predictor of the mediator improvement type. Individuals with lower depression were more likely to improve their sense of mastery under the treatment condition. Although the joint method assumes the effect for treatment nonimprovers is 0, the principal propensity score method allows us to also estimate the effect for treatment nonimprovers but finds no significant effect.

The two methods allow us to examine the potential validity of the other methods' key assumption. From the joint model results in Table 3, the outcome model intercepts are about half a standard deviation apart, which may help explain the size difference in the treatment effect estimate between the propensity score and the joint estimation approaches (0.270 vs. 0.475) in that it shows a potential violation of principal ignorability (Jo & Stuart, 2009). In parallel, the fact that the point estimate of the effect for treatment nonimprovers is 0.127, as estimated using the principal propensity score approach, indicates that there may be some violation of the exclusion restriction (although that estimate is not significant). Because both the exclusion restriction and principal ignorability are assumptions regarding unknown quantities, we cannot tell for sure which assumption is indeed violated. It is also possible that both assumptions are somewhat violated. Nonetheless, the two approaches are consistent in terms of the conclusion that the JOBS II intervention led to a meaningful decrease in depression levels for treatment improvers. The next step would be to formally conduct systematic sensitivity analyses considering plausible ranges of deviation from each identifying assumption (e.g., Jo & Vinokur, in press; Stuart & Jo, in press).

Table 4 shows the estimated effects of treatment assignment on the change in depression when individuals are stratified based on their potential mediator improvement status under the control condition. For the joint method used here, we imposed the exclusion restriction on control improvers, which includes two substrata, always-improvers and backward-improvers. In the mediational process hypothesized in the JOBS II intervention, the group of individuals who would positively change their sense of mastery in response to treatment assignment (i.e., forward-improvers) was the intervention target. Little impact of the treatment was expected for the other three improvement types (never-improvers, always-improvers, backward-improvers). This may partially explain why the largest effects are seen for the two reference strata that consist of forward-improvers: the control nonimprovers and treatment improvers. However, as in the analysis with the exclusion restriction imposed on treatment nonimprovers, the assumption is not fully testable based on observed data, leaving a possibility of deviation from the assumption and subsequent bias in causal effect estimates.

Again the two methods are consistent in showing that control nonimprovers benefited from being assigned to the treatment condition. That is, the treatment assignment led to a considerable decrease in depression (effect size = 0.48 for the joint model and 0.35 for the principal propensity score method) for those who would not have improved sense of mastery when in the control condition. Baseline motivation and sense of mastery were found to be significant predictors of the outcome for control improvers, and only the baseline mastery score was a significant predictor of the outcome for control nonimprovers. Baseline depression, motivation, and assertiveness were found to be significant predictors of the mediator improvement type. Individuals with lower depression, lower motivation, and higher assertiveness were more likely to improve their sense of mastery under the control condition. It is also shown that the outcome model intercepts are quite close in this analysis (about 0.2 standard deviations apart), which may explain the similarity in the treatment effect estimates between the propensity score and the joint estimation approaches (0.270 vs. 0.378). As in the analyses using the treatment as the reference condition, the two approaches are again consistent in terms of the conclusion that being assigned to the treatment intervention was useful to individuals who would not improve their sense of mastery otherwise. The principal propensity score approach also allows us to conclude that there is no effect of the intervention for control improvers: those individuals who would have improved sense of mastery even without the intervention.

CONCLUSIONS

This article has presented a new way of estimating mediation effects, using propensity scores to estimate effects within strata defined by potential mediator values under treatment and control conditions. The approaches presented here used coarsened (reference) strata, where each stratum includes mixed groups of individuals. For example, the treatment improver class consists of individuals who all would improve their sense of mastery when in the treatment condition, but some would also improve their sense of mastery when in control whereas others would not. In other words, our intention was not in identifying purely direct or purely indirect effects of treatment assignment. Therefore, a direct translation of our results to the structural equation model parameters is not as straightforward as in Jo (2008), where the four principal strata and their treatment effects were considered and connected to the structural equation model parameters. However, the current approach still provides useful information about mediation that can be understood as an interaction effect. For example, the results of the JOBS II analyses using the principal stratification methods can be interpreted as that the treatment assignment differently affected the change in depression for individuals who would differently change their sense of mastery in response to the treatment assignment. In part the coarsening to reference strata is done so that strata membership is fully observed for at least some individuals; future work should investigate how to extend these ideas to estimate the effects for all four principal strata, defined by mediator status under both treatment and control.

The binary mediator setting we considered in this article provided a setting where different mediation analysis methods can be compared with the fewest unverifiable assumptions. In this context, the key identifying assumption in the SEM approach is conditional ignorability of observed M . In the principal stratification framework, the key assumption is principal ignorability for the principal propensity score method and the exclusion restriction for the joint estimation method. The three assumptions impose different restrictions on different unobservable parameters, which may be useful for sensitivity analysis in mediation analysis. In this article, we did not attempt to fully translate assumptions and resulting estimates across different methods. However, this seems to be a viable task and necessary for mediation modeling to fully benefit from multiple methods. It is also important to recognize that nearly all of the assumptions considered in this article are untestable; that is in part why we believe that using a variety of methods, with different underlying assumptions, can be particularly beneficial as sensitivity analyses.

In terms of the JOBS II intervention, both the SEM and PS analyses indicate a positive impact of the JOBS II intervention on depression via the targeted mediator, sense of mastery. By considering both analyses, using the control or the treatment as the reference condition, we can see that a sizable effect is associated with the common principal stratum, forward-improvers. This implies that a desirable and hypothesized mediational process occurred. Recall that the treatment improver group consists of forward-improvers and always improvers, whereas the treatment nonimprover group consists of backward improvers and never-improvers. Likewise, the control improver group consists of always-improvers and backward-improvers, whereas the control nonimprover group consists of never-improvers and forward-improvers. If monotonicity is assumed (i.e., that there are no backward-improvers), which is an implicit assumption of the SEM method, treatment nonimprovers include only never-improvers and control improvers include only always-improvers. Therefore, the treatment effect estimates for these two coarsened strata can be interpreted as the direct effect that corresponds to the c' effect in the SEM approach. Once we know the direct effect, we can roughly calculate the indirect effect based on the relation that total (overall ITT) = direct effect + indirect effect. Because the direct effect estimates based on the principal propensity score method are small, we can roughly expect a very

small direct effect. Interestingly, these calculations agree quite well with the structural equation model results of direct and indirect effect estimation, supporting our speculation on the similarity between the two ignorability assumptions (ignorability of M and principal ignorability).

There are several limitations of our approach to be addressed in future research. One limitation of this approach is that it requires a binary mediator, which sometimes involves dichotomizing a continuous measure, as we did here. A benefit of this is that it can simplify analyses and their interpretation and relies less on unverifiable assumptions. With binary mediators bounding approaches are also feasible for sensitivity analyses. Drawbacks of this dichotomization include loss of information (although not necessarily power) and sensitivity to the choice of dichotomizing threshold (MacCallum, Zhang, Preacher, & Rucker, 2002), which is a general problem in principal stratification. In current work we are investigating the use of generalized propensity scores (Imai & van Dyk, 2004) to deal with continuous measures of compliance, which may be able to be extended to the mediation context as well. The methods and assumptions are more complex in the case of a continuous mediator (or measure of compliance), and thus we feel that it is important to understand the binary mediator setting first.

A possible concern with the joint model approach is violation of the assumption of normally distributed potential outcomes. This may be a particular concern in the two-strata setting of reference stratification, where each stratum consists of a mixture of two possibly heterogeneous subpopulations. However, previous work has found that the joint estimation method is not too sensitive to deviation from normality unless the deviation is severe (Stuart & Jo, in press). The principal propensity score method is even more robust to deviation from normality. However, how well treatment effects can be estimated with coarsened strata needs further examination. Another interesting question regards the relationship between the assumptions of the ignorability of the observed mediator M in the SEM framework versus the principal ignorability assumption regarding the strata indicator C in the principal propensity score approach. Future work should also consider different ways of estimating effects within the principal propensity score method, especially given concerns about the potential instability of weighting approaches. One other possible approach is full matching as discussed in a noncompliance setting in Jo and Stuart (2009). Another possibility is principal score stratification, where individuals are stratified based on their predicted mediator values under one or both treatment conditions. Future work should investigate the relative performance of these approaches and which may be the most effective in practice. Future investigations should also more carefully consider the implications and possible validity of the principal ignorability assumption and in particular which types of predictors are most important to include in the principal propensity score model.

To implement the current propensity score approach, coarsening of principal strata based on observed mediator values is necessary. However, by mixing individuals who do and do not change in response to treatment assignment (e.g., by grouping based only on treatment group mediator status and not considering control group mediator status at the same time), we are basically mixing the direct and indirect effects and therefore do not clearly capture the effects that occur through change in the mediator. This makes comparison of the resulting effects difficult in terms of their concordance with the effects from traditional mediation analysis and theories of mediation. Interestingly, the same problem also exists in the conventional SEM approach (when we do analyses based on observed M , we are always doing this kind of coarsening), although the problem does not get explicitly discussed because the potential values of the mediator are not considered. The difference is that this problem can be more clearly seen in the potential outcomes framework, which we consider an advantage of the PS approach. In this article, we focused on introducing an alternative

method for mediation analysis and providing some techniques for sensitivity analysis. However, more effective and thorough sensitivity analyses need to be developed, which we leave as a topic for future study. Further research is also needed to analytically examine the relationship between the principal ignorability model and the SEM model and to examine potential benefits from integrating the two, such as conceptual proximity to mediation theory and possibilities for sensitivity analysis.

The identification of mediating processes is complex, requiring a program of research with information from many sources and designs (MacKinnon, 2008). And in many fields such as the treatment and prevention research example in this article, the development and refinement of successful interventions is based on integration of all available information including partial and confounded information to develop the best possible interventions. Propensity score methods such as the principal propensity score approach described here provide a way of assessing mediation that directly addresses untestable assumptions based on potential outcomes. A crucial topic for future work is to further investigate the assumptions underlying each approach and begin to understand when different modeling approaches are more appropriate and how they can be jointly used to improve causal inference in mediation modeling.

Acknowledgments

This research was supported by the National Institute of Mental Health (MH083846, PI Stuart; MH086043, PI Alongo; MH066319, PI Jo) and the National Institute of Drug Abuse (DA09757, PI MacKinnon).

REFERENCES

- Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. 1996; 91:444–455.
- Baron R, Kenny D. 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, K. Total, direct and indirect effects in structural equation models. In: Clogg, C., editor. *Sociological Methodology*. American Sociological Association; 1987. p. 37–69.
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
- Hill J, Waldfogel J, Brooks-Gunn J. Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology*. 2003; 39:730–744. [PubMed: 12859126]
- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*. 2007; 15:199–236. <http://pan.oxfordjournals.org/cgi/content/abstract/15/3/199?etoc>.
- 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. *Sociological Methodology*. 1988; 18:449–484.
- Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychological Methods*. 2010; 15:309–334. [PubMed: 20954780]
- Imai K, van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*. 2004; 99:854–866.
- Jin H, Rubin DB. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*. 2008; 103:101–111.
- Jo B. Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*. 2002; 27:385–409.

- Jo B. Causal inference in randomized experiments with mediational processes. *Psychological Methods*. 2008; 13:314–336. [PubMed: 19071997]
- Jo B, Asparouhov T, Muthén B. Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine*. 2008; 27:5565–5577. [PubMed: 18623608]
- Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*. 2009; 28:2857–2875. [PubMed: 19610131]
- Jo B, Vinokur A. Sensitivity analysis and bounding of causal effects with alternative identifying assumptions. *Journal of Educational and Behavioral Statistics*. (in press).
- Jo B, Wang C-P, Jalongo NS. Using latent outcome trajectory classes in causal inference. *Statistics and Its Interface*. 2009; 2:403–413. [PubMed: 20445809]
- Joffe MM, Small D, Hsu C. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*. 2007; 22:74–97.
- Judd C, Kenny D. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*. 1981; 5:602–619.
- Kraemer H, Kiernan M, Essex M, Kupfer D. How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*. 2008; 27:101–108.
- Little RJ, Yau LH. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*. 1998; 3:147–159.
- MacCallum R, Zhang S, Preacher K, Rucker D. On the practice of dichotomization of quantitative variables. *Psychological Methods*. 2002; 7:19–40. [PubMed: 11928888]
- MacKinnon, D. Introduction to statistical mediation analysis. Mahwah, NJ: Erlbaum; 2008.
- MacKinnon D, Dwyer J. Estimating mediating effects in prevention studies. *Evaluation Review*. 1993; 17:144–158.
- MacKinnon D, Goldberg L, Clarke G, Elliott D, Cheong J, Lapin A, Krull JL. Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self-efficacy and dietary behavior. *Prevention Science*. 2001; 2:15–28. [PubMed: 11519372]
- MacKinnon DP, Lockwood C, Hoffman J, West S, Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*. 2002; 7:83–104. [PubMed: 11928892]
- McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*. 2004; 9:403–425. [PubMed: 15598095]
- Meichenbaum, D. Stress inoculation training: A clinical guide book. New York, NY: Pergamon/Elsevier; 1985.
- Muthén, L.; Muthén, B. Mplus user's guide. Los Angeles, CA: Author; 1998–2010.
- Neyman J. On the application of probability theory to agricultural experiments: Essay on principles. Section 9. Translated in *Statistical Science* (with discussion). 1923; 5:465–480. (1990).
- Ozer E, Bandura A. Mechanisms governing empowerment effects: A self-efficacy analysis. *Journal of Personality and Social Psychology*. 1990; 58:472–486. [PubMed: 2324938]
- Price R, van Ryn M, Vinokur A. Impact of a preventive job search intervention on the likelihood of depression among the unemployed. *Journal of Health and Social Behavior*. 1992; 33:158–167. [PubMed: 1619263]
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
- Robins J, Rotnitzky A. Comment on Peter J. Bickel and Jaimyoung Kwon, inference for semiparametric models: Some questions and an answer. *Statistica Sinica*. 2001; 11:920–936.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
- Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*. 1978; 6:34–58.
- Rubin DB. Bias reduction using Mahalanobis metric matching. *Biometrics*. 1980; 36:293–298.

- 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.
- Schafer JL, Kang JD. Average causal effects from nonrandomized studies: A practical guide and simulated case study. *Psychological Methods*. 2008; 13:279–313. [PubMed: 19071996]
- Sobel, ME. Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt, S., editor. *Sociological methodology*, 1982. Washington, DC: American Sociological Association; 1982. p. 290-312.
- 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 studies with mediating variables. *Journal of Educational and Behavioral Statistics*. 2008; 33:230–251.
- Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science*. 2010; 25:1–21. [PubMed: 20871802]
- Stuart EA, Jo B. Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*. (in press).
- Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. Causal mediation analyses with rank preserving models. *Biometrics*. 2007; 63:926–934. [PubMed: 17825022]
- Vinokur A, Price R, 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 A, Schul Y. Mastery and inoculation against setbacks as active ingredients in intervention for the unemployed. *Journal of Counseling and Clinical Psychology*. 1997; 65:867–877.
- Wood R, Bandura A. Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology*. 1989; 56:407–415. [PubMed: 2926637]

FIGURE 1.
Structural equation model: conceptual framework of mediation.

FIGURE 2.
Principal stratification: conceptual framework of mediation.

TABLE 1Job Search Intervention Study (JOBS II): Sample Statistics ($N = 422$)

Variable	Control Group ($X = 0$)		Treatment Group ($X = 1$)	
	Mean	SD	Mean	SD
Y (depression $T0 - T6$)	0.367	0.826	0.497	0.758
M (mastery $T2 - T0$)	0.073	0.421	0.224	0.474
W_1 (depression at $T0$)	2.483	0.290	2.425	0.301
W_2 (mastery at $T0$)	3.537	0.472	3.509	0.558
W_3 (age in years at $T0$)	36.533	10.059	37.482	10.293
W_4 (school grade completed at $T0$)	13.440	1.957	13.440	2.071
W_5 (assertiveness at $T0$)	3.011	0.895	3.053	0.917
W_6 (motivation at $T0$)	5.306	0.787	5.335	0.819
W_7 (not married (0/1) at $T0$)	0.563	0.498	0.640	0.481
W_8 (non-White (0/1) at $T0$)	0.167	0.374	0.176	0.382
W_9 (female (0/1) at $T0$)	0.611	0.489	0.579	0.495

TABLE 2

Job Search Intervention Study (JOBS II): Structural Equation Model Using Sense of Mastery as a Mediator (*M*)

Parameter	Continuous M		Dichotomized M	
	Estimate	SE	Estimate	SE
<i>a</i>	0.130	0.040	0.107	0.046
<i>b</i>	0.374	0.087	0.219	0.077
<i>c'</i> (direct)	0.149	0.072	0.174	0.073
<i>ab</i> (indirect)	0.049	0.019	0.023	0.013
<i>ab + c'</i> (total)	0.197	0.073	0.197	0.073
β_{mW_2} (mastery)	-0.465	0.039	-0.436	0.045
β_{yW_1} (depression)	1.026	0.114	1.034	0.116
β_{yW_2} (mastery)	0.472	0.081	0.393	0.078

JOBS II: Principal Stratification Model Estimation for Treatment Improvers and Nonimprovers

TABLE 3

Treatment Improvers ($C_1 = 1$, $N = 223$)			Treatment Nonimprovers ($C_1 = 0$, $N = 199$)		
Parameter	Estimate	SE	Parameter	Estimate	SE
Principal Propensity Scores Method					
γ_{1C1} (treatment effect)	0.270	0.098	γ_{0C1} (treatment effect)	0.127	0.087
α_{1C1} (intercept)	0.339	0.069	α_{0C1} (intercept)	0.321	0.069
λ_{1C1} (depression)	0.417	0.108	λ_{0C1} (depression)	0.252	0.110
λ_{1C1} (mastery)	0.999	0.142	λ_{0C1} (mastery)	1.119	0.168
σ^2_{1C1} (res variance)	0.507	NA	σ^2_{0C1} (res variance)	0.513	NA
<i>Logistic regression of C_1 on W ($C_1 = 1$ vs. $C_1 = 0$)</i>					
β_{0C1} (logit intercept)	0.172	0.140			
β_{1C1} (depression)	-2.102	0.334			
Joint Estimation Method					
γ_{1C1} (treatment effect)	0.475	0.154	γ_{0C1} (treatment effect)	fixed at 0	
α_{1C1} (intercept)	0.115	0.151	α_{0C1} (intercept)	0.496	0.064
λ_{1C1} (depression)	0.392	0.126	λ_{0C1} (depression)	0.142	0.121
λ_{1C1} (mastery)	0.996	0.162	λ_{0C1} (mastery)	1.118	0.189
σ^2_{1C1} (res variance)	0.515	0.060	σ^2_{0C1} (res variance)	0.394	0.048
<i>Logistic regression of C_1 on W ($C_1 = 1$ vs. $C_1 = 0$)</i>					
β_{0C1} (logit intercept)	0.132	0.145			
β_{1C1} (depression)	-2.007	0.315			

TABLE 4
JOBS II: Principal Stratification Model Estimation for Control Improvers and Nonimprovers

Control Improvers ($C_0 = 1$, $N = 169$)			Control Nonimprovers ($C_0 = 0$, $N = 253$)		
Parameter	Estimate	SE	Parameter	Estimate	SE
Principal Propensity Scores Method					
γ_{1c0} (treatment effect)	0.108	0.098	γ_{0c0} (treatment effect)	0.270	0.100
α_{1c0} (intercept)	0.438	0.088	α_{0c0} (intercept)	0.249	0.092
λ_{1c0} (motivation)	-0.030	0.050	λ_{0c0} (motivation)	-0.012	0.052
λ_{1c0} (mastery)	0.921	0.128	λ_{0c0} (mastery)	0.959	0.131
σ^2_{1c0} (res variance)	0.455	NA	σ^2_{0c0} (res variance)	0.503	NA
<i>Logistic regression of C_0 on W ($C_0 = 1$ vs. $C_0 = 0$)</i>					
β_{0c0} (logit intercept)	-0.420	0.202			
β_{1c0} (depression)	-2.555	0.540			
β_{1c0} (motivation)	-0.549	0.266			
β_{1c0} (assertiveness)	0.753	0.265			
Joint Estimation Method					
γ_{1c0} (treatment effect)	fixed at 0	.	γ_{0c0} (treatment effect)	0.378	0.108
α_{1c0} (intercept)	0.375	0.139	α_{0c0} (intercept)	0.202	0.113
λ_{1c0} (motivation)	-0.272	0.118	λ_{0c0} (motivation)	0.084	0.063
λ_{1c0} (mastery)	1.356	0.236	λ_{0c0} (mastery)	0.901	0.188
σ^2_{1c0} (res variance)	0.446	0.085	σ^2_{0c0} (res variance)	0.444	0.090
<i>Logistic regression of C_0 on W ($C_0 = 1$ vs. $C_0 = 0$)</i>					
β_{0c0} (logit intercept)	-0.574	0.225			
β_{1c0} (depression)	-2.320	0.577			
β_{1c0} (motivation)	-0.536	0.251			
β_{1c0} (assertiveness)	0.646	0.286			