Multivariate Behavioral Research, 46:842–854, 2011

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.606743

Psychology Press
Taylor & Francis Group

Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis

Kosuke Imai
Princeton University

Booil Jo
Stanford University

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health

In this commentary, we demonstrate how the potential outcomes framework can help understand the key identification assumptions underlying causal mediation analysis. We show that this framework can lead to the development of alternative research design and statistical analysis strategies applicable to the longitudinal data settings considered by Maxwell, Cole, and Mitchell (2011).

We begin our discussion by congratulating Maxwell, Cole, and Mitchell (2011; hereafter MCM) for providing a thought-provoking article that tackles the challenging problem of conducting mediation analysis in longitudinal settings. As MCM clearly demonstrated, although mediation analysis plays an essential role in psychological research, conducting such an analysis requires researchers to understand the underlying assumptions and adopt appropriate statistical analysis and research design strategies.

In this commentary, we use the potential outcomes framework to highlight the fundamental issues that arise in mediation analysis. Although MCM did

Correspondence concerning this article should be addressed to Kosuke Imai, Department of Politics, Princeton University, Princeton, NJ, 08544. E-mail: kimai@princeton.edu

not use it, we believe that this framework helps researchers to better understand the advantages and disadvantages of various mediation analysis methods. Within this framework, we discuss the values of longitudinal data for mediation analysis and offer alternative approaches to the method proposed by MCM.

The work and ideas we describe here are connected to other recent work in mediation analysis and represent a bridging of two literatures. Holland (1988) explored possible assumptions necessary to make causal inferences in the mediation analysis context more than two decades ago. Since then, growing interest in causal mediation analysis has been seen in the causal modeling literature (e.g., Mealli & Rubin, 2003; Pearl, 2001; Petersen, Sinisi, & van der Laan, 2006; Robins & Greenland, 1992, 1994; Rubin, 2004; Ten Have, Elliott, Joffe, Zanutto, & Datto, 2004). In the meantime, most mediation analyses in psychological studies have been conducted using the structural equation modeling (SEM) approach (Baron & Kenny, 1986; Bollen, 1987; Judd & Kenny, 1981; MacKinnon, 2008) with less emphasis on explicit causal assumptions and modeling. Methods to improve estimation and inferential procedures for SEM-based mediation analyses have continued to develop (e.g., Kraemer, Kiernan, Essex, & Kupfer, 2008; MacKinnon, 2008; Shrout & Bolger, 2002; Sobel, 1982, 1988).

However, this trend appears to be changing in recent years with the explosion of causal mediation papers that attempt to understand the SEM approach within the formal frameworks of causal inference (e.g., Albert, 2008; Imai, Keele, Tingley, & Yamamoto, 2011; Imai, Keele, & Yamamoto, 2010; Jo, 2008; Jo, Stuart, MacKinnon, & Vinokur, 2011; Pearl, in press; Sobel, 2008; VanderWeele, 2008). In this commentary, we demonstrate how the potential outcomes framework can help understand the key identification assumptions underlying causal mediation analysis. We show that this framework can lead to the development of alternative research design and statistical analysis strategies applicable to the longitudinal data settings considered by MCM.

MEDIATION ANALYSIS IN THE POTENTIAL OUTCOMES FRAMEWORK

As formalized by Rubin (1974), in the potential outcomes framework, the effect of some treatment $T = 1$ (vs. a control condition $T = 0$) on an outcome Y for individual i can be expressed as the difference between two potential outcomes, $Y_i(1) - Y_i(0)$, where $Y_i(1)$ is the value of the outcome the individual would experience if exposed to the treatment, and $Y_i(0)$ represents the outcome the individual would experience if exposed to the control. What is called the fundamental problem of causal inference (Holland, 1986) is that individuals can only observe one potential outcome for each person. Estimation of causal effects can thus be thought of as inferring the missing potential outcomes in

a reasonable way. Randomized experiments are desirable in part because, by ensuring comparability of the treatment and control groups, they can provide valid estimates of the missing potential outcomes, and thus unbiased estimates of the treatment effects.

Defining and Identifying Mediation Effects

Potential outcomes are also useful for defining effects related to intermediate posttreatment variables such as mediators or measures of compliance that occur (or are measured) in between treatment assignment and the ultimate outcomes. Suppose we use $M_i(1)$ and $M_i(0)$ to denote potential mediator values for unit i that would realize under the treatment and control conditions, respectively. We define the potential outcomes as functions of treatment and mediator where $Y_i(t, m)$ represents the outcome that would result if the treatment variable T_i takes the value t and the value of the mediator M_i equals m. For example, $Y_i(0, M_i(1))$ refers to the potential value of the outcome for unit i who is assigned to the control condition but takes on a value of the mediator that would be realized under the treatment condition. Then, the quantity MCM referred to as the indirect effect can be defined as $Y_i(t, M_i(1)) - Y_i(t, M_i(0))$ for $t = 0$ or 1, which can be read as the difference between the potential value of the outcome under the two scenarios where unit i whose treatment status is set to t takes on values for the mediator of $M_i(1)$ and $M_i(0)$ under the treatment and control conditions, respectively (Pearl, 2001; Robins & Greenland, 1992).

In MCM's example, this quantity represents the causal effect on children's depression attributable to the change in problematic parenting induced by mother's depression. Holding the treatment variable constant at t eliminates the direct effect of the treatment, thereby isolating the indirect effect of mother's depression on children's depression that transmits through problematic parenting from other alternative causal mechanisms. Using potential outcomes, we can define the direct effect as $Y_i(1, M_i(t)) - Y_i(0, M_i(t))$. These definitions formalize the intuition held by researchers about direct and indirect effects.

What assumptions are required to identify these direct and indirect effects? The potential outcomes framework makes it clear that these quantities involve purely counterfactual outcomes that can never be observed. For example, in order to infer an indirect effect, defined previously as $Y_i(0, M_i(1)) - Y_i(0, M_i(0))$, researchers must make inferences about $Y_i(0, M_i(1))$, which is an inherently unobservable quantity.[1] Imai, Keele, and Yamamoto (2010) showed that the se-

[1] We note that some distinguish counterfactuals from potential outcomes (Rubin, 2005). While potential outcomes (e.g., $Y_i(1, M_i(1))$) are potentially observable for everyone, counterfactuals (e.g., $Y_i(1, M_i(0))$) are inherently unobservable.

quential ignorability assumption must be satisfied in order to identify the average causal mediation effects. This key assumption implies that the treatment assignment is essentially random after adjusting for observed pretreatment covariates and that the assignment of mediator values is also essentially random once both observed treatment and the same set of observed pretreatment covariates are adjusted for. The validity of commonly used mediation analysis based on SEM also critically relies upon this sequential ignorability assumption (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010).

Although this assumption may appear to be similar to the usual exogeneity or no-omitted-variable assumption, in contrast to commonly held belief (e.g., Spencer, Zanna, & Fong, 2005), the randomization of both treatment and mediator in general does not satisfy this sequential ignorability assumption. This means that even when the treatment and mediator are randomized, we cannot identify mediation effects unless an additional assumption (e.g., no interaction effect between treatment and mediator) is imposed (for details, see e.g., Imai, Keele, Tingley, & Yamamoto, 2011; Imai & Yamamoto, 2011; Robins, 2003). Therefore, the potential outcomes framework clarifies the challenge researchers need to overcome when conducting mediation analysis, which is not necessarily apparent in the traditional SEM framework.

An alternative way to think about mediation analysis within the potential outcomes framework is to use the concept of principal strata (Frangakis & Rubin, 2002). It is easiest to discuss the main ideas of principal stratification with a binary mediator; the concepts and methods could potentially be extended to continuous mediators, such as by modifying the existing methods for continuous compliance measures (Imai & van Dyk, 2004; Jin & Rubin, 2008) or by using hybrid methods that combine the propensity score and latent variable approaches (Stuart, Warkentien, & Jo, 2011). In this commentary, we do not discuss the details of these methods that are presently under development. In general, however, stronger and untestable assumptions are necessary to identify causal effects with continuous mediators under the principal stratification framework.

We focus our discussion on a simpler but important setting of binary mediator. In MCM's example, $M_i = 1$ and $M_i = 0$ indicate problematic and nonproblematic parenting, respectively. In this scenario, we can consider the following four principal strata, defined by the combination of mediator values under the treatment, $M_i(1)$, and control, $M_i(0)$, conditions (in analogy with groups defined in Jo [2008] and Jo et al. [2011]):

- Never problematic (not problematic whether depressed or not): $(M_i(0), M_i(1)) = (0, 0)$
- Depressed problematic (problematic only when depressed): $(M_i(0), M_i(1)) = (0, 1)$

- Not depressed problematic (problematic only when not depressed): $(M_i(0), M_i(1)) = (1, 0)$
- Always problematic (problematic whether depressed or not): $(M_i(0), M_i(1)) = (1, 1)$

Under this setup, for the units that are classified as never problematic or always problematic, the causal effect of mother's depression on children's depression can be considered as a direct effect because the mediator is not affected by the treatment received by these units. For the units that are classified as depressed problematic or not depressed problematic, the causal effect of mother's depression on children's depression can be considered as a combination of the direct and indirect effects, where the indirect effect is attributed to the change in parenting style induced by mothers' depression (Rubin, 2004, 2005).

The causal effects defined for these principal strata can be translated more closely into the direct and indirect effects defined in Baron and Kenny's (1986) approach under certain conditions (Jo, 2008). For example, if the treatment effect is constant across never problematic and always problematic strata, we can interpret the effect for those two strata as the direct effect (i.e., c' in Baron and Kenny's approach). Additionally, if we can assume that there are no not depressed problematic individuals, the difference between this direct effect and the treatment effect on the depressed problematic category can be interpreted as the effect of the mediator on the outcome (i.e., b) in Baron and Kenny's approach. The percentage of depressed problematic individuals corresponds to the treatment effect on the mediator (i.e., a) in Baron and Kenny's approach.

These definitions of direct and indirect effects also involve quantities that cannot be directly observed. However, under certain conditions, the direct and indirect effects can be identified for a subset (or subsets) of the population. A typical set of identifying assumptions (monotonicity and exclusion restriction) in this setting can be found in Angrist, Imbens, and Rubin (1996). The monotonicity assumption excludes the existence of mothers who would develop as problematic parents only when they are not depressed (no not depressed problematic stratum of women). This assumption seems quite plausible given the context. The exclusion restriction assumes that there exists no direct effects of mothers' depression on children. The assumption implies that any indirect effects mediated by unobserved mediators (or mediators that are not included in the analysis) also do not exist. The plausibility of the exclusion restriction is more questionable, and furthermore it is not quite compatible with the common intention of mediation analysis (Jo, 2008). Thus, researchers may explore alternative identification assumptions. For example, under the sequential ignorability assumption, causal effects for all four principal strata can be identified and these effects can be translated into the direct and indirect effects as defined in the conventional SEM approach.

In sum, the potential outcomes framework makes it clear that strong and untestable assumptions are required for identifying causal mediation effects. Thus, a primary methodological challenge is to devise alternative research designs that require less stringent assumptions. Next, we discuss the extent to which the longitudinal setting considered by MCM may help overcome this challenge.

ALTERNATIVE APPROACHES TO MEDIATION ANALYSIS USING LONGITUDINAL DATA

Given the aforementioned difficulty in identifying mediation effects, it is important to consider alternative research designs and analytical strategies. MCM showed one strategy to exploit the availability of longitudinal data to overcome this difficulty. Here, we discuss two additional ways to take advantage of longitudinal data.

Crossover Design

One experimental design that can be used in the longitudinal data setting is the crossover design (Jones & Kenward, 2003). The simplest crossover design consists of a binary treatment and two time periods. Under this 2×2 design, the treatment is randomly assigned in the first period and then in the second period the treatment status is switched for each unit: the units that received the treatment (control) condition in the first period receive the control (treatment) condition in the second period. The key idea of the crossover design is that under the assumption of no carryover effects, both of the potential outcomes, $Y_i(1)$ and $Y_i(0)$, can be observed for each unit.

Imai, Tingley, and Yamamoto (2009) modified the standard crossover design so that it could be applicable to mediation analysis. Under the proposed design, the first period remains identical to that of the standard crossover design. During the second period, however, the mediator value is set to the observed mediator value from the first period while the treatment is switched as in the standard design. This means that if a unit belongs to the treatment group in the first period, $M_i(1)$ and $Y_i(1, M_i(1))$ are observed while in the second period we observe $Y_i(0, M_i(1))$, identifying the direct effect $Y_i(1, M_i(1)) - Y_i(0, M_i(1))$ for this unit. Because the direct and indirect effects sum up to the total effect, all quantities are identifiable provided that the assumption of no carryover effects is satisfied.

As a potential application of this crossover design, consider a randomized field experiment about labor market discrimination of African Americans in

which fictitious resumes are sent to potential employers with either African American– or White-sounding applicant names (Bertrand & Mullainathan, 2004). The quantity of interest is the difference in callback rates (Y) between applications with African American– and White-sounding names (T). Suppose that researchers use the crossover design to estimate the direct effect of perceived race of applicants (T) while adjusting for their perceived qualification (M). Under this design, a resume with an African American–sounding name that was sent in the first period will be sent again with a White-sounding name in the second period.

Alternatively, the same resume can be sent to multiple employers at the same time to estimate the same quantity of interest, and such a design may make the assumption of no carryover effects more plausible. The success of this design then depends critically on whether perceived qualification stays the same between the two periods, despite the fact that the applicant names have been changed. To ensure the credibility of this assumption, researchers may make the information that potential employers may use when evaluating qualifications of applicants as concrete and detailed as possible (e.g., providing GPA and academic honors in addition to the name of college attended).

Although this crossover design is promising, in many situations it may be difficult to directly manipulate the mediator. Imai et al. (2009) showed that this basic crossover design can be extended to the situation where the mediator can be manipulated only imperfectly. Furthermore, the key idea of this crossover design can be applied to observational research by finding natural experiments where the mediator is held constant across two periods. Imai, Keele, Tingley, and Yamamoto (2011) described examples of such research designs in observational data settings.

Principal Stratification Approach Using Propensity Scores

In this section we describe an approach to mediation analysis that uses propensity scores. Propensity scores have traditionally been used to estimate causal effects in nonexperimental settings. The propensity score itself is formally defined as the probability of receiving treatment given the observed covariates; the propensity scores are then used to create similar groups through matching, subclassification, or weighting (Stuart, 2010). One benefit of propensity score methods in general is that they have a clear separation of design and analysis (Rosenbaum, 2010), as does a randomized experiment. In particular, the matching is done using baseline (pretreatment) covariates, and the outcome is only used to estimate treatment effects after well-matched groups have been obtained. In fact, the propensity score matching can be done before the outcome data are even available, thus limiting the potential for bias in selecting a matched sample to obtain a desired result, as well as providing an opportunity to focus resources on following up

only those individuals who are good matches, as discussed in Stuart and Ialongo (2010).

Longitudinal data with covariates (realized and measured before treatment assignment) and treatment assignment (realized and measured before outcomes) eliminates the possibility of reverse causality and thus provides a clear way to adhere to this prescription of design followed by analysis. In contrast, with cross-sectional data, it is often unclear whether variables are in fact pre- or posttreatment. Careful data investigation and analysis (e.g., by examining the correlation of variables with the treatment as in Kraemer et al. [2008]) as well as subject matter knowledge will help improve understanding of the causal ordering of these variables in cross-sectional designs, although longitudinal designs have advantages in this regard. Of course the same problem can arise with longitudinal data that contains retrospective questions or variables with unclear temporal ordering, in which case the same cautions apply as for cross-sectional data.

An approach to mediation analysis using propensity score methods is described in detail in Jo et al. (2011). This approach works best when we simplify the four principal strata by defining strata based only on the potential mediator values under one treatment condition, labeled *reference stratification* (Jo, Wang, & Ialongo, 2009). For example, we may be particularly interested in children whose mothers would (or would not) be poor parents when depressed, regardless of their parenting practices when not depressed. If we use the depressed condition as the reference condition, there would be two strata, each of which combines two of the four strata we previously defined:

- Problematic under depression (depressed problematic and always problematic): $M_i(1) = 1$
- Not problematic under depression (never problematic and not depressed problematic): $M_i(1) = 0$.

Similar reference strata can be defined under the not-depressed condition. These strata might be of interest because they provide prognostic classes or represent baseline conditions of parenting behavior when a mother is depressed (or not depressed). As discussed in Jo et al. (2011), a benefit of reference stratification is that strata membership is fully observed under one treatment condition. In the previous example, we know to which stratum each of the mothers in the treatment condition (the depressed condition) belongs. On the basis of this property, the basic idea is to use propensity scores to model mediator status under one condition (e.g., parenting skills when depressed), using the subjects in that group (the depressed group). Then use that model to generate predictions of what the parenting levels of the nondepressed group would have been if they had actually been depressed (Hill, Waldfogel, & Brooks-Gunn [2003] termed this the *principal score,* linking principal stratification and

propensity scores). Subjects with poor parenting in the depressed group are then matched to nondepressed group women using this principal score, to estimate the effect of maternal depression on their children's depression levels, given poor parenting.

The effect of mothers' depression for the previous reference strata does not necessarily have a direct analog in traditional mediation analysis (e.g., direct and indirect effects), but nonetheless can provide insight into mediational processes. In particular, when monotonicity (which we believe is very plausible) is also assumed, the second reference stratum in the previous example only includes never problematic women. Then, the effect of treatment assignment (depression) on that stratum of individuals can be interpreted as the direct effect of mothers' depression on children's depression (without going through parenting). See Jo et al. (2011) for further discussion of the links and interpretation of these effects.

The primary assumption of the principal score method is what Jo and Stuart (2009) termed *principal ignorability*. This assumption states that the value of the mediator is independent of the potential outcomes given the observed covariates. Another way of saying this is that the observed covariates are sufficient in terms of identifying individuals in one group who would have had a particular value of the mediator had they been in the other group. Of course the validity of principal ignorability depends on any particular study. Jo & Stuart (2009) and Jo et al. (2011) provided a thorough discussion of this assumption and its plausibility. In MCM's example, this would mean that we could well predict whether a mother would have problematic parenting skills when depressed and when not depressed, given baseline characteristics. This assumption is more credible when there are strong baseline predictors of parenting practices, such as information on the mothers' own childhood and attitudes toward children and parenting.

The suggested principal stratification approach based on propensity scores can also be extended by taking advantage of repeated measures of mediators and outcomes. For example, individuals can be stratified into a few heterogeneous groups (latent growth trajectory types) based on their longitudinal trajectory patterns. Some individuals may develop increasing trends and others may develop stable or decreasing trends over time. One possibility of incorporating this information in the principal stratification framework is to use the growth trajectory types of the outcome under a particular treatment condition as reference strata (Jo et al., 2009). Another possible extension is to use the growth trajectory types of a mediator as reference strata (Wang, Jo, & Brown, 2011). In both extensions, identifying heterogeneous trajectory strata under a reference condition would involve exploratory methods such as growth mixture modeling (Muthén, 2004). Once the trajectory strata are identified under one condition, the same propensity score method we described previously can be applied treating the strata membership as known for those assigned to the reference condition and unknown (missing) for those assigned to the other condition.

Identifying causal effects in this latent variable framework is in principle not any different from doing so in the usual potential outcomes based approaches. The common goal is to recover causal treatment effects for heterogeneous strata, where individuals' membership in these strata is intrinsic and not affected by treatment. The difference is that the standard principal stratification framework defines these intrinsic strata based on prespecified rules (e.g., the four principal strata in MCM's example of mother's depression and problematic parenting), whereas in the latent variable model framework, we empirically estimate strata membership, making the exercise inherently exploratory.

An advantage of this latent variable framework is the ability to construct strata when prespecified rules are not so evident (e.g., repeated measures of mediator or the outcome or continuous mediators with subjective cut points). A drawback is that this exploratory process requires identifying assumptions related to functional and distributional forms, which we normally do not use in causal modeling practice. Because of this, interpreting causal effects based on these assumptions and conducting sensitivity analysis is potentially more difficult in the latent variable framework. Whether incorporating exploratory components in causal modeling undermines philosophical tenets of the potential outcomes approach is not so clear at this point (although we do not believe it does) and the issue has not been discussed much. This is a largely underexplored and challenging area in causal modeling. We believe that, if successfully developed, causal models incorporating latent variable strategies are likely to benefit causal modeling and latent variable modeling practice.

CONCLUDING REMARKS

MCM made an important contribution to the methodological literature on mediation analysis by discussing some of the complexities in the context of longitudinal data analysis. As is evident from their discussion, any mediation analysis requires researchers to understand the key underlying assumptions and carefully interpret the results. We believe that the potential outcomes framework provides a useful way to understand the advantages and disadvantages of different methods for mediation analysis. As shown previously, the framework also clarifies different definitions of mediation effects, thereby facilitating the interpretation of empirical results.

As demonstrated by MCM, the use of longitudinal data provides potential ways to overcome some of the difficulties encountered by mediation analysis in cross-sectional settings because repeated observations in general provide more information useful for inferring mediation effects. In this commentary, we discussed a few additional ways to take advantage of longitudinal data in experimental and observational studies. The use of the potential outcomes framework

led to these new statistical methods and research designs. For example, under the proposed crossover design, the advantage of longitudinal data can be clearly seen in the potential outcomes settings in that for the same unit the two potential outcomes that are required for causal mediation analysis can both be observed (under the assumption of no carryover effects). We believe that the potential outcomes framework can bring about further methodological developments for mediation analysis in the future.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Mental Health (MH083846, E. A. Stuart; MH086043, N.S. Ialongo) and the National Science Foundation (SES–0918968, K. Imai).

REFERENCES

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine, 27,* 1282–1304.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association, 91,* 444–455.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal?: A field experiment on labor market discrimination. *American Economic Review, 94,* 991–1013.

Bollen, K. (1987). Total, direct and indirect effects in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 37–69). Washington, DC: American Sociological Association.

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

Hill, J., Waldfogel, J., & Brooks-Gunn, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39,* 730–744.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association, 81,* 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 449–484). Washington, DC: American Sociological Association.

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

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of casualty: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105,* forthcoming.

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

Imai, K., Tingley, D., & Yamamoto, T. (2009). *Experimental designs for identifying causal mechanisms.* Technical report, Department of Politics, Princeton University. Retrieved from http://imai.princeton.edu/research/Design.html

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99,* 854–866.

Imai, K., & Yamamoto, T. (2011). *Sensitivity analysis for causal mediation effects under alternative exogeneity assumptions.* Technical report, Department of Politics, Princeton University. Retrieved from http://imai.princeton.edu/research/medsens.html

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association, 103,* 101–111.

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

Jo, B., & Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine, 28,* 2857–2875.

Jo, B., Stuart, E. A., MacKinnon, D. P., & Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research, 46,* 425–452.

Jo, B., Wang, C.-P., & Ialongo, N. S. (2009). Using latent outcome trajectory classes in causal inference. *Statistics and Its Interface, 2,* 403–412.

Jones, B., & Kenward, M. G. (2003). *Design and analysis of cross-over trials* (2nd ed.). London, England: Chapman & Hall.

Judd, C., & Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Kraemer, H., Kiernan, M., Essex, M., & Kupfer, D. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27,* 101–108.

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

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research, 46,* 816–841.

Mealli, F., & Rubin, D. B. (2003). Comment on Adams et al. and assumptions allowing the estimation of direct causal effects. *Journal of Econometrics, 112,* 79–87.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.

Pearl, J. (in press). The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*.

Petersen, M. L., Sinisi, S. E., & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology, 17,* 276–284.

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

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

Robins, J. M., & Greenland, S. (1994). 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, 89,* 737–749.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31,* 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100,* 322–331.

Shrout, P., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). Washington, DC: American Sociological Association.

Sobel, M. E. (1988). Direct and indirect effects in linear structural equation models. In J. S. Long (Ed.), *Common problems/proper solutions* (pp. 46–64). Beverly Hills, CA: Sage.

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

Spencer, S., Zanna, M., & Fong, G. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89,* 845–851.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25,* 1–21.

Stuart, E. A., & Ialongo, N. S. (2010). Matching methods for selection of subjects for follow-up. *Multivariate Behavioral Research, 45,* 746–765.

Stuart, E. A., Warkentien, S., & Jo, B. (2011). *Using propensity scores to account for varying levels of program participation in randomized controlled trials.* Manuscript in preparation.

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

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

Wang, C.-P., Jo, B., & Brown, C. H. (2011). *Causal inference in longitudinal comparative effectiveness studies with repeated measures of a continuous intermediate variable.* Manuscript submitted for publication.