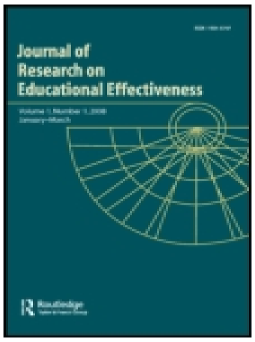

Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

# Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder

Nicole Bohme Carnegie, Masataka Harada & Jennifer L. Hill

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2016, VOL. 9, NO. 3, 395–420
http://dx.doi.org/10.1080/19345747.2015.1078862

METHODOLOGICAL STUDIES

OPEN ACCESS

# Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder

Nicole Bohme Carnegie[a], Masataka Harada[b], and Jennifer L. Hill[c]

**ABSTRACT**

A major obstacle to developing evidenced-based policy is the difficulty of implementing randomized experiments to answer all causal questions of interest. When using a nonexperimental study, it is critical to assess how much the results could be affected by unmeasured confounding. We present a set of graphical and numeric tools to explore the sensitivity of causal estimates to the presence of an unmeasured confounder. We characterize the confounder through two parameters that describe the relationships between (a) the confounder and the treatment assignment and (b) the confounder and the outcome variable. Our approach has two primary advantages over similar approaches that are currently implemented in standard software. First, it can be applied to both continuous and binary treatment variables. Second, our method for binary treatment variables allows the researcher to specify three possible estimands (average treatment effect, effect of the treatment on the treated, effect of the treatment on the controls). These options are all implemented in an R package called treatSens. We demonstrate the efficacy of the method through simulations. We illustrate its potential usefulness in practice in the context of two policy applications.

**KEYWORDS**
causal inference
sensitivity analysis
unmeasured confounder
hidden bias
omitted variable

One of the biggest struggles in policy analysis is the inability to directly study many questions of interest using randomized experiments. Attempts to infer causality using nonexperimental studies are vulnerable to violations of the assumption that requires, colloquially speaking, that all confounders have been measured. Rather than abandoning the goal of causal inference using nonexperimental data, methods that allow researchers to explore sensitivity of their inferences to violations of this assumption can act as a middle ground. These approaches help to build confidence in the results of nonexperimental studies by differentiating between studies whose results are relatively immune to potential unmeasured confounders versus those whose results might change drastically.

This article presents a set of graphical and numeric tools to investigate the sensitivity of causal estimates in nonexperimental studies to the presence of an unmeasured confounder.[1]

**CONTACT** Nicole Bohme Carnegie carnegin@uwm.edu Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201-0413, USA.

[a]University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA
[b]National Graduate Institute for Policy Studies, Tokyo, Japan
[c]New York University, New York, New York, USA

[1]Throughout the article we use the term "unmeasured confounder" rather than using terms such as "omitted variable" or "unobserved covariate" for the sake of consistency and clarity. In practice it is possible that this variable was measured but was simply omitted from the analysis.

We characterize the confounder through two parameters which describe (a) the relationship between the confounder and the treatment assignment and (b) the relationship between the confounder and the outcome variable. Our approach to assessing sensitivity to an unmeasured confounder has two primary advantages over similar approaches (with two sensitivity parameters) that are currently implemented in standard software. First, it can be applied to both continuous and binary treatment variables. Second, our method for binary treatment variables allows the researcher to specify three possible estimands (average treatment effect, effect of the treatment on the treated, effect of the treatment on the controls). All tools described in this article are available in an R package called treatSens.

## Brief Overview of Approaches to Assess Sensitivity to Unmeasured Confounding

The term "sensitivity analysis" has several meanings in empirical data analysis. In the field of causal inference this term typically refers to approaches that assess the sensitivity of causal estimates to the presence of unmeasured confounders. This is sometimes referred to as sensitivity to "unmeasured confounding" or to "hidden bias" (Rosenbaum, 2002a). These methods have a long history in applied statistics, beginning perhaps with the exploration by Cornfield et al. (1959) into the robustness of the causal link between smoking and cancer, in which the magnitude of hidden bias necessary to explain away the observed relationship between smoking and risk of cancer was quantified. Other seminal work in this area includes sensitivity analysis for 2 × 2 tables developed by Bross (1966, 1967) and a method for a binary outcome and a categorical covariate by Rosenbaum and Rubin (1983).[2] Nevertheless, comparatively little attention has been paid to this important topic compared to other topics in causal inference.

Extant methods for evaluation of sensitivity to hidden bias can be roughly divided into semi- or nonparametric approaches and parametric approaches. Work on semi- or nonparametric approaches to sensitivity analysis has been dominated by the contributions of Rosenbaum and colleagues (chapter 4 in Rosenbaum, 2002b provides an overview) although others (for example, Steenland & Greenland, 2004) have made important contributions as well. Rosenbaum presents several methods of assessing sensitivity in studies with matched pairs of treated and control observations (Rosenbaum, 1987), for analyses with strata containing multiple matched controls to each treated observation (Rosenbaum, 1988), for matched case-control studies (Rosenbaum, 1991), and for unmatched, two-sample observational studies (Rosenbaum & Krieger, 1990).[3] These rely on nonparametric randomization tests such as McNemar's test for binary treatments and outcomes (Greenland, 1996; Rosenbaum, 1987) or Wilcoxon signed-rank test for continuous outcomes.[4] As an applied example, Buckley and Schneider (2007, chapter 9) assess the sensitivity to unmeasured confounding of the causal effect of charter schools on parent satisfaction using Rosenbaum's methods for the Wilcoxon signed-rank test.

[2] Ichino, Mealli, and Nannicini (2008) developed a computational approach along this line; sensatt (Nannicini, 2007) performs their sensitivity analysis in Stata.

[3] In Stata, rbounds (DiPrete & Gangl, 2004) performs the sensitivity analysis of Rosenbaum (2002b) for continuous outcome variables, while mhbounds (Becker & Caliendo, 2007) performs the counterpart for binary outcome variables. In R, rbounds (Keele, 2010) performs the sensitivity analysis of Rosenbaum (2002b).

[4] Some of these methods are nonetheless categorized as a semiparametric due to their reliance on a parametric model to estimate propensity scores used to match or stratify.

Although these methods have the appeal of avoiding parametric assumptions, they also have several limitations. First and foremost the bulk of the nonparametric methods require the researcher to first create matched samples (typically based on the propensity score), which itself can be a labor-intensive process with uncertain results (see, e.g., Hill, Weiss, & Zhai, 2011). Another concern with these methods is that they have been shown to be sensitive to the choice of test statistic used (Rosenbaum, 2010).[5]

Traditionally, a primary drawback of these approaches was that they only allowed the researcher to explore sensitivity based on a range of assumptions about the relationship between the unmeasured confounder and the treatment assignment. For instance, in the seminal work by Rosenbaum (2002a), the sensitivity parameter $\Gamma$ sets bounds on the odds of one member of a perfectly matched pair receiving the treatment relative to the other receiving the treatment; the association between the confounder and the outcome, however, is assumed to be nearly collinear. This simplification is overly conservative and may lead researchers to be more concerned with violations of ignorability than is merited. In more recent work, Rosenbaum and Silber (2009) have extended this paradigm to decompose the $\Gamma$ parameter into a constrained set of corresponding parameters that reflect both the association between the unobserved confounder and the treatment as well as the association between the unobserved confounder and the sign of the difference in outcomes across members of the matched pairs. This extension has the merit of allowing the researcher to consider a wider range of potential unobserved confounders. However, it loses information about a continuous outcome variable by dichotomizing the treatment versus control difference into a simple categorization of a positive or negative difference in outcomes (ignoring the magnitude of this difference).

Parametric approaches to sensitivity analysis have their own strengths and weaknesses. These generally begin with a model familiar to applied researchers, such as a *t* test for the difference in means across treatment groups, a standard linear regression, a logistic or probit regression for dichotomous dependent variables, or one of several survival models, and then make additional assumptions about the relationship between hypothetical unmeasured confounders and both the selection to treatment and the outcome measure (for examples, see, Altonji, Elder, & Taber, 2005; Copas & Li, 1997; Frank, 2000; Lin, Psaty, & Kronmal, 1998). These models typically allow for specification of two sensitivity parameters: one that describes the relationship between the unmeasured confounder and the treatment assignment and one that describes the relationship between the unmeasured confounder and the outcome.[6] They have the drawback of making potentially restrictive assumptions about the parametric form of the models for the outcome and the treatment assignment. Of course, for researchers with a taste for more robust model estimation, this drawback could be addressed by using these approaches on matched samples. In other words, these approaches allow for, but do not require, use of matching to address concerns regarding parametric assumptions.

Given the difficulty in implementing randomized experiments and the widespread concern with the assumptions required in nonexperimental studies, it is curious that

[5] However, recent work by Rosenbaum (2011) presents a promising alternative statistic.

[6] Some methods in this realm (McCandless, Gustafson, & Levy, 2007; Rosenbaum & Rubin, 1983 for instance,) rely on still more sensitivity parameters, which we find to be prohibitively complicated for standard usage. A few strategies, such as the one suggested by Altonji et al. (2005) require only one sensitivity parameter but this restriction unfortunately makes it quite difficult to interpret.

sensitivity analysis methods have not gained more traction among applied researchers interested in pursuing causal questions. This persists despite the recent availability of software (within Stata and R) to implement the nonparametric methods. We posit that one of the biggest barriers has been the difficulty of understanding the sensitivity parameters in the nonparametric approaches and the related difficulty of calibrating those to the empirical context. For this reason we have pursued the parametric approaches that, as we shall demonstrate, allow for sensitivity parameters that are easier to understand and calibrate, do not require (but do allow for) matched samples, and allow for separate specification of associations between the unmeasured confounder and the treatment and the unmeasured confounder and the outcome.

Among the parametric methods currently available in standard software, our approach most resembles two existing methods. The first was proposed by Imbens (2003) and is a likelihood-based approach specific to a context with a continuous outcome (with a linear, additive relationship with the covariates and treatment and normally distributed errors), a binary treatment variable, and a binary unmeasured confounder. Sensitivity parameters are specified as the partial correlations between the unmeasured confounder and the treatment and between the unmeasured confounder and the outcome. Optimization methods are used to estimate the treatment effect conditional on specific values of the sensitivity parameters.[7]

The second method was proposed by Harada (2013). It uses residualized versions of the outcome (from a linear model conditional on the treatment and covariates) and the treatment variable (from a linear model conditional on the covariates) to generate candidate values of the unmeasured confounder in a constrained range. The algorithm saves the candidate values of the unmeasured confounder that change the estimate of the treatment effect by a specified amount. The corresponding values of the sensitivity parameters (partial correlations from linear models) can then be obtained by fitting the appropriate linear models with the unmeasured confounder.[8] Both approaches, as currently implemented, target the average treatment effect.

The algorithm we propose in this article is likelihood-based like the Imbens method but handles either a binary or a continuous treatment variable. Similar to Harada we generate candidate values of the unmeasured confounder; however, we draw these directly from the distribution of the confounder conditional on observed data (as derived from the specified complete-data likelihood); consequently our approach is more computationally efficient than that of Harada. Our approach is not as computationally efficient as Imbens's approach; however, it fits within a framework that will be easier to generalize in future work to accommodate complications such as multilevel data and nonlinear/nonparametric models. Moreover, our approach can target any of three different causal estimands: the effect of the treatment on the treated, the effect of the treatment on the controls, or the average treatment effect. The other two approaches, as currently implemented, can only estimate the average treatment effect. Similar to the other two approaches, we have built software that will be available to researchers; ours is implemented in the freely available R software package.

[7] isa (Harada, 2011) performs the sensitivity analysis of Imbens (2003) in Stata.
[8] gsa (Harada, 2012) performs the sensitivity analysis of Harada (2013) in Stata.

## Causal Inference Notation and Theoretical Framework

This work is motivated to address a lack of confidence in the assumption, commonly invoked with observational data, that we have measured all confounders. In statistics, this assumption is formalized as a statement about the independence between the potential outcomes and the treatment variable conditional on a set of covariates, $X$,

$$Y(z) \perp Z|X \quad \forall z \in \mathbf{Z} \tag{1}$$

and is referred to as the ignorability assumption.[9] This is a critical assumption, because if it is satisfied, we can identify causal effects simply by appropriately adjusting for the covariates in $X$ (for instance by an appropriate model for $Y$ conditional on the covariates and $Z$ or by a propensity score approach).

It is rare, however, that researchers have confidence that this assumption holds. As a strategy for exploring sensitivity to this assumption, we posit an unmeasured variable, $U$, that, if included, would serve to satisfy ignorability. That is, we assume that our set of potential outcomes, $Y(Z)$, is conditionally independent of the treatment assignment, $Z$, given both the vector of observed covariates, $X$, and an additional unmeasured confounder $U$. Thus, the ignorability assumption would relax to the following version,

$$Y(z) \perp Z|X, U \quad \forall z \in \mathbf{Z} \tag{2}$$

We illustrate the basic structure of our sensitivity analysis first assuming that our data comprise independent, identically distributed units. In addition, we make modeling assumptions. Specifically, we assume that, if ignorability is satisfied given $X$ and $U$, the following equation holds

$$E[Y(Z)|X, U] = \beta^y X + \zeta^y U + \tau Z \tag{3}$$

The observed $Y$ for individual $i$ is then given by $Y_i = E[Y_i(Z_i)|X_i, U_i] + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma_y)$ and where $\beta^y$ is the vector of regression coefficients for the covariate matrix $X$ and $\tau$ is the true population-level treatment effect.

Given the assumed correspondence between potential outcomes and the linear model and the fact that we will be using linear models for the outcome throughout, we will henceforth refer only to observed outcomes in our discussion of the statistical distributions required for inference.

## Complete Data Likelihoods and Conditional Distribution of the Unmeasured Confounder

The goal of our enterprise is to understand what treatment effect estimate we might have obtained had we conditioned on a posited unmeasured confounder $U$. To do this, as described in more detail in the following section, we (a) specify a model for the way the world works (in statistical terms, a complete-data likelihood), (b) use this model to derive hypothetical values of $U$ based on our assumptions about its relationship with the outcome and the treatment (through sensitivity parameters), and (c) estimate the treatment effect

[9] Ignorability is typically referred to as "selection on observables" in economics and "all confounders measured" or "exchangeability" in the epidemiology literature.

given these potential values of $U$. This section describes the information pertinent to these first two steps. We start by specifying complete-data likelihoods for each of two scenarios, corresponding to continuous and binary treatments, respectively.[10] We then describe the conditional distribution of $U$ that corresponds to each of these models.

First, we motivate the (nonparametric) factorization of the joint complete data likelihood,

$$\mathrm{p}(Y, Z, U|X) = \mathrm{p}_1(Y|Z, U, X)\mathrm{p}_2(Z|U, X)\mathrm{p}_3(U|X) \tag{4}$$

This factorization allows us to specify sensitivity parameters that correspond to conditional associations between the unmeasured confounder, $U$, and both the outcome and the treatment variable in a way that is familiar. In particular (as is made specific below) we can specify these sensitivity parameters as the coefficient on $U$ in the regression of $Y$ on $Z$, $U$, and $X$ and the coefficient on $U$ in the regression of $Z$ on $U$ and $X$. We make the additional assumption that $U \perp X$, giving $\mathrm{p}_3(U|X) = \mathrm{p}_3(U)$. Although some authors (Gustafson, McCandless, Levy, & Richardson, 2010) have argued that this is an overly restrictive assumption, we argue for a conceptualization of $U$ as that portion of the confounder that is independent of $X$. Because most researchers are used to thinking about conditional associations (e.g., regression coefficients) rather than marginal associations, we feel that this leads to more understandable and more easily calibrated sensitivity parameters. Moreover, it reduces the number of sensitivity parameters required. It also allows for a relatively simple specification for the distribution of $U$.

Our strategy for recovering treatment effects conditional on $U$ requires generating a potential confounder $\dot{U}$ consistent with the stated complete data likelihood so we can estimate the true treatment effect by conditioning on both $X$ and the generated $\dot{U}$. In the next sections we derive the distribution of $U$ given $Y$, $Z$, $X$, and our sensitivity parameters (defined below) for each of two complete-data likelihoods corresponding to different parametric models.

### Continuous Treatment Variable

A reasonable starting point for pursuing this approach is to consider a scenario with continuous outcome, treatment variable, and unmeasured confounder modeled with normal distributions. In particular we posit

$$\begin{aligned} &Y|X, U, Z \sim N\left(X\beta^y + , \zeta^y U + \tau Z, \sigma^2_{\mathrm{y \cdot xuz}}\right) \\ &Z|X, U \sim N\left(X\beta^z + , \zeta^z U, \sigma^2_{\mathrm{z \cdot xu}}\right) \\ &U \sim N\left(0, \sigma^2_u\right) \end{aligned} \tag{5}$$

where $\beta^y$ and $\beta^z$ represent the vectors of regression coefficients for the covariate matrix $X$, and $\tau$ is the population-level treatment effect. It is natural to think of $U$ as being normally distributed if we conceptualize $U$ as a linear combination of all unmeasured confounders, which is approximately normally distributed for a sufficient number of confounders or if the

[10] We avoid creating a separate set of models for binary outcomes due to the issues with noncollapsibility inherent in such models (Greenland, Robins, & Pearl, 1999).

confounders are themselves normally distributed. With no loss of generality we can assume that ${\sigma^2}_u = 1$. $\zeta^y$ and $\zeta^z$ act as sensitivity parameters in the form of regression coefficients that define the relationship between the unmeasured confounder and the outcome (conditional on $X$ and $Z$) and the relationship between the unmeasured confounder and the treatment variable (conditional on $X$), respectively.

This complete data likelihood implies the following distribution for $U$ conditional on the other variables (for derivation see the appendix).

$$\begin{aligned} U|X,Z,Y &\sim N\left(\mu_{\mathrm{u\cdot xzy}}, \sigma^2_{\mathrm{u\cdot xzy}}\right) \\ \mu_{\mathrm{u\cdot xzy}} &= \frac{\zeta^z}{\sigma^2_{\tilde{z}}}\tilde{z} + \frac{\left(\sigma^2_{\tilde{z}} - \zeta^{z2}\right)\zeta^y}{\sigma^2_{\tilde{z}}\sigma^2_{\tilde{y}}}\tilde{y} \\ \sigma^2_{\mathrm{u\cdot xzy}} &= \frac{\sigma^2_{\tilde{z}} - \zeta^{z2}}{\sigma^2_{\tilde{z}}\sigma^2_{\tilde{y}}}\left(\sigma^2_{\tilde{y}} - \zeta^{y2} + \frac{\zeta^{z2}\zeta^{y2}}{\sigma^2_{\tilde{z}}}\right) \end{aligned} \tag{6}$$

where $\tilde{z}, \tilde{y}, \sigma^2_{\tilde{z}}$ and $\sigma^2_{\tilde{y}}$ denote the realized residuals and their respective variances from regressions run for $Z$ conditional on $X$ and for $Y$ conditional on $Z$ and $X$, respectively.

We can draw $\dot{U}$ from an estimated version of this conditional distribution for any valid combination of values for the sensitivity parameters (discussed in the "General Algorithm" section) by plugging these values into the formula along with estimates of the required residual variances.

### Binary Treatment Variable

Although it is possible to use the above likelihood as an approximate solution for the situation with binary $Z$, a more appropriate solution is to posit an alternate complete data likelihood for this scenario

$$Y|X,U,Z \sim \mathrm{N}\left(X\beta^y + \zeta^y U + \tau Z, \sigma^2_{\mathrm{y\cdot xuz}}\right) \tag{7}$$

$$Z|X,U \sim \mathrm{Bernoulli}(\Phi(X\beta^z + \zeta^z U)) \tag{8}$$

$$U \sim \mathrm{Bernoulli}(\pi^u) \tag{9}$$

where $\Phi$ denotes the probit link. We posit a binary $U$ in this scenario for reasons of mathematical convenience.

The distribution of $U$ conditional on the sensitivity parameters and observed data corresponding to this complete data likelihood is not as readily available as in the multivariate normal case. Thus we rely on a computational trick to draw from the correct distribution. This strategy capitalizes on the fact that if the parameters from the complete data likelihood were known, the conditional distribution of $U$ would be straightforward to define as a

Bernoulli distribution with probabilities equal to the ratio of appropriate likelihoods.

$$U|Y,Z,X \sim \text{Bernoulli}\left(\frac{\pi^{y,z,x,u=1}}{\pi^{y,z,x}}\right) \tag{10}$$

where the numerator (representing the joint likelihood of the variables with U set equal to 1) and denominator (representing the likelihood after marginalizing over U) can be calculated using the following formulas

$$\begin{aligned}\pi^{y,z,x,u=1} = \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2}\exp\left(-\frac{\left(Y - X\beta^y - \zeta^y - \tau Z\right)^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\ \cdot\left(1-\Phi\left(X\beta^z+\zeta^z\right)\right)^{(1-Z)}\left(\Phi\left(X\beta^z+\zeta^z\right)\right)^Z\pi^u\end{aligned} \tag{11}$$

and

$$\begin{aligned}\pi^{y,z,x} = {} & \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2}\exp\left(-\frac{\left(Y - X\beta^y - \tau Z\right)^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\ & \cdot\left(1-\Phi\left(X\beta^z\right)\right)^{(1-Z)}\left(\Phi\left(X\beta^z\right)\right)^Z\left(1-\pi^u\right) \\ & + \left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2}\exp\left(-\frac{\left(Y - X\beta^y - \zeta^y - \tau Z\right)^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \\ & \cdot\left(1-\Phi\left(X\beta^z+\zeta^z\right)\right)^{(1-Z)}\left(\Phi\left(X\beta^z+\zeta^z\right)\right)^Z\pi^u\end{aligned} \tag{12}$$

The challenge in making use of this distribution is that we not only do not know the true parameters needed to calculate the probabilities for this conditional distribution, we cannot even estimate all of them unbiasedly. In order to properly estimate $\beta^y$, $\tau$, and $\sigma^2_{\mathrm{y\cdot xuz}}$, we would need to know $U$. This motivates an iterative estimation strategy that is a form of Stochastic EM (expectation-maximization) algorithm (Neilsen, 2000). We iterate between two steps. In the one step we estimate the model parameters in Equation 7 and Equation 8 conditional on the observed data and a draw of the vector $U$. In the other step we draw $\dot{U}$ from its conditional distribution (as in Equation 10) conditional on the data and the parameter value estimates from the previous step. We repeat these steps enough times that we are confident that we are drawing it from the correct distribution; in our applications this typically took less than 10 steps.

## Sensitivity Analysis Algorithm and Treatment Effect Estimates

This section describes the algorithm we use to estimate treatment effects corresponding to a set of assumptions about our unmeasured $U$, codified in our sensitivity parameters. It also describes some extensions of the general algorithm and the plot that can be produced to display results. The algorithms and plots described here are implemented in the R package treatSens.

### General Algorithm

For any valid combination of sensitivity parameters, $\zeta^z$ and $\zeta^y$, we generate an estimate of the treatment effect, $\tau$, and the corresponding standard error through a multistep process based on generation of a simulated potential confounder, $\dot{U}$, drawn from an estimate of the conditional distribution of $U|Y$, $Z$, $X$. In the case of continuous treatment, this involves first obtaining residuals from linear models of $\mathrm{E}[Y|X, Z]$ and $\mathrm{E}[Z|X]$, then plugging the residuals and their empirical variances into the closed-form formulas for the mean and variance of $\dot{U}$, and, finally, drawing $\dot{U}$ from the corresponding normal distribution with the sensitivity parameters set to the given values. Once we obtain a realization of $\dot{U}$, we then fit a regression of $Y$ on $Z$, $X$ and $\dot{U}$, and record the estimate of $\tau$ and the standard error of the estimate; we denote these estimates as $\widehat{\tau}_k(\zeta^z, \zeta^y)$ and $\hat{\sigma}_{\hat{\tau}_k(\zeta^z,\zeta^y)}$.

In the case of binary treatment, we do not need to estimate residuals, but rather iterate between estimating the parameters of $\mathrm{E}\left[Y|X, Z, \dot{U}_k\right]$ and $\mathrm{E}\left[Z|X, \dot{U}_k\right]$ and drawing $\dot{U}_{k+1}$ from the distribution of $U|Y$, $Z$, $X$, given those parameters. We find that the algorithm typically converges in a small number of steps (less than 20 is usually sufficient). Once again the realization of $\dot{U}$ is used to obtain $\widehat{\tau}_k(\zeta^z, \zeta^y)$ and $\hat{\sigma}_{\hat{\tau}_k(\zeta^z,\zeta^y)}$.

Of course any single such estimate of $\tau$ corresponds to just one realization from the distribution of $U$ and as such will reflect the idiosyncratic nature of that draw. A more accurate estimate is obtained by averaging over $K$ such estimates to obtain $\widehat{\tau}(\zeta^z, \zeta^y) = \frac{1}{K}\sum_{k=1}^{K} \widehat{\tau}_k(\zeta^z, \zeta^y)$. A corresponding standard error estimate[11] that reflects both the uncertainty conditional on a given draw of $\dot{U}$ as well as uncertainty across draws of $\dot{U}$ is obtained as $\hat{\sigma}_{\hat{\tau}(\zeta^z,\zeta^y)} = \sqrt{W + (1 + \frac{1}{K})B}$, where $W = \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}^2_{\hat{\tau}_k(\zeta^z,\zeta^y)}$ and $B = \frac{1}{K-1}\sum_{k=1}^{K} \left(\widehat{\tau}_k(\zeta^z, \zeta^y) - \widehat{\tau}(\zeta^z, \zeta^y)\right)^2$.

To gain an overall picture of the sensitivity of the treatment effect estimate, we divide the range of valid sensitivity parameters into a grid, and compute $\widehat{\tau}$ for each cell of the grid. Valid sensitivity parameters are determined by examining the residual variance after conditioning on observed variables; intuitively, an unmeasured confounder cannot account for any more variability in treatment or outcome than is observed. When $Z$ is binary, there is no such hard constraint, but a probit coefficient on a binary or standardized continuous variable in excess of $\pm 2$ is highly unlikely to be observed in practice, so we restrict our sensitivity analyses to this range when using a probit model.

### Visualization and Interpretation of Results

We visually summarize the grid of sensitivity-parameter-specific treatment effect estimates computed by our sensitivity analysis algorithm using a contour plot such as the one displayed in Figure 1. For demonstration, we use a simulated data set with a continuous outcome, a binary treatment, four observed continuous confounders (all with the same association with the treatment but with varying associations with the outcome), and a binary unobserved confounder. We performed the sensitivity analysis across an 8 by 4 grid with 40 repetitions for each combination of sensitivity parameters.

[11] This is analagous to the variance formula used for multiple imputation estimates (Rubin, 1987).

The visualization combines four different definitions of contours, each represented by a different color. The primary result is given by the black contours. Each of these represents the combinations of sensitivity parameters for $U$ that lead to the same estimated treatment effect (that is, the estimated coefficient on $Z$ in the regression of $Y$ on $Z$, $X$, and $U$, noted along the contour). For instance, sensitivity parameter pairs ($\zeta^z = 2$, $\zeta^y = 1$) and ($\zeta^z = 1$, $\zeta^y = 1.6$) both (approximately) fall on a contour corresponding to a treatment effect estimate of .17. The red curve represents the contour along which the treatment effect estimate is reduced to zero. For this treatment effect estimate to be driven to zero by a single confounder, that confounder would have to have a combination of sensitivity parameters that would fall on the line labeled 0—for instance, a confounder whose coefficient in the treatment assignment model was 2 and whose coefficient in the outcome model was 1.3. The gray contour is used for calibration and is discussed below.

The blue curves bracket the region in which the treatment effect estimate is not significant at the 5% level using the aggregated standard error estimate $\hat{\sigma}_{\hat{\tau}(\zeta^z,\zeta^y)}$. As such, they no longer represent combinations of sensitivity parameters that lead to the same estimated treatment effect, but rather combinations that lead to the same $t$-statistic $\hat{\tau}(\zeta^z,\zeta^y)/\hat{\sigma}_{\hat{\tau}(\zeta^z,\zeta^y)}$. They may cross the black contours if the standard error varies strongly with the sensitivity parameters. Thus the plot indicates that for a positive treatment effect estimate to be driven to nonsignificance in our example the confounder would have to have sensitivity parameters at least as large as those that fall on the lower line labeled "N.S."—for instance, a coefficient of about 2 in the treatment assignment model and a coefficient of about .9 in the outcome model. A confounder with sensitivity parameters at least as large as those on the upper "N.S." line would produce a significant *negative* treatment effect (e.g., coefficients of about 2 in the treatment model and about 1.8 in the outcome model).

We restrict the range of sensitivity parameters for the coefficient in the model for $Y$ to be positive; results in the negative range are a simple reflection of the positive, and thus provide no additional information. The x- and y-axes represent regions of no confounding, because one or both of the sensitivity parameters is set to zero; thus the observed treatment effect estimate (0.79 in this simulated example) is used as a label for the x-axis.

Thus far we have been interpreting this plot from the vantage point of a researcher with no knowledge of the "truth." Because the data are simulated we know that the true $U$ for these data was generated using sensitivity parameters equal to 1 in both dimensions and that the population effect is .5. We see that the point (1,1) falls reasonably close to the contour corresponding to a treatment effect estimate of .51 (given that this combination of sensitivity parameters has a standard error of approximately .15).

### Calibration

As other scholars have before us (e.g., Imbens, 2003), we suggest calibrating the strength of these sensitivity parameters by using estimates of the regression coefficients for the observed covariates, $X$, in the data (from the regression of $Y$ on $Z$ and $X$ and from the regression of $Z$ on $X$) as a benchmark. Any confounder with a negative partial association with the outcome is reverse-scaled so that this coefficient estimate will be positive; these estimates are tagged by displaying them as $\bigtriangledown$. To create a common scale for these coefficients (among themselves and relative to the coefficients that serve as our sensitivity parameters) we first standardize all continuous covariates and the outcome to have mean of zero and standard deviation of one. As an additional calibration

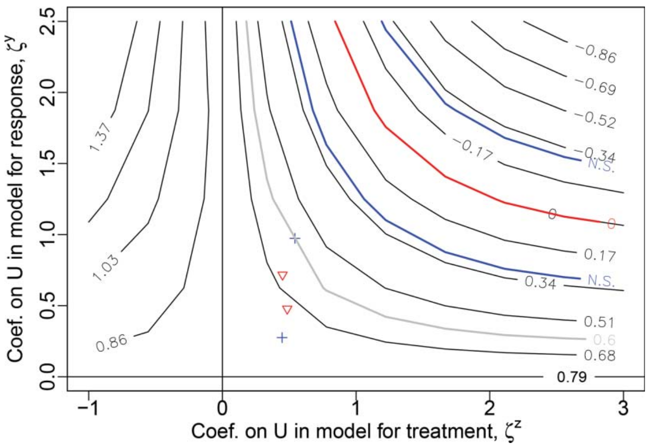

**Figure 1.** Each contour on the plot reflects the combinations of sensitivity parameters that would lead to the treatment effect estimate used as the label for that contour. The lines labeled "N.S." correspond to the treatment effect estimates that forfeit statistical significance. The pale contour corresponds to the effect estimate that would be yielded by an unmeasured confounder with properties equivalent to that of the observed confounder that is farthest from the origin. The plus signs and triangles give the coefficient estimates for the observed covariates. The ▽ sign reflects a covariate whose coefficient for the outcome model was negative, prompting a rescaling of the variable by − 1.

aid, we plot a gray contour corresponding to the treatment effect estimate obtained with sensitivity parameters equivalent to the pair of observed coefficients that are farthest from the origin.[12] The display of the coefficients for the observed confounders in this example allows us to see that, in order for either no treatment effect or a nonsignificant treatment effect to represent the truth, the unmeasured confounder would have to have considerably greater predictive power than the observed confounders.

## *Use of Modification Weights to Target Specific Estimands*

Thus far in the literature, methods that are similar to ours—that is, parametric methods that specify both the relationship between $U$ and $Z$ as well as that between $U$ and $Y$ (both conditional on $X$)—have focused on the average treatment effect as the estimand of interest. However, there has been increasing interest in the past two decades in other estimands, perhaps most notably the effect of the treatment on the treated (ATT) and the effect of the treatment on the controls (ATC). These estimands may differ from the average treatment effect when treatment effects are not constant (or additive). When the heterogeneity in treatment effects is driven by the observed covariates, $X$, these estimands can be recovered by methods such

[12] It may seem odd that this gray line will not necessarily intersect the symbol for the covariate with the most extreme set of coefficients. This is driven by two considerations. The first is the fact that the plot function provides contours that represent a smoothed fit to the estimated treatment effect estimates at our specified grid points. The second is that the estimates of the coefficients on the observed covariates from the $Y$ model will not be precisely equivalent to estimates from the $\beta^y$ in the model formulation that includes $U$.

as inverse probability of treatment weighting (IPTW; see, e.g., Imbens, 2004; Kurth et al., 2006).[13]

To target estimands such as ATT and ATC we have incorporated "modification weights" into the algorithm. These are similar to the weights used in IPTW with the distinction that they do not condition on $U$ and thus are not functions of the estimates of what we assume to be the true propensity score, $\Pr(Z=1|X,U)$. Rather, they are functions of estimates of what we, for clarity, refer to as the "modification score," $g(x)=\Pr(Z=1|X)$. Otherwise these weights will be formed similarly to those in traditional IPTW, with modification scores estimated using probit regression (as a default, in practice any of a variety of prediction models for a binary response could be used). Specifically, if we want to estimate the effect of the treatment on the treated we weight the control observations by $\widehat{g}(x)/(1-\widehat{g}(x))$ (normalized so that the sum of these weights is equal to the number of control observations). If we want to estimate the effect of the treatment on the controls we weight treated observations by $(1-\widehat{g}(x))/\widehat{g}(x)$ (again normalized so that the sum of these weights is equal to the number of treated observations). Use of modification scores rather than propensity scores assumes that treatment effects do not vary with levels of U after conditioning on X (if they did the entire sensitivity analysis strategy would need to change).

## Simulation Study

We present a simulation study to demonstrate the performance of our approach in the more complex setting, binary treatment with weighted estimators.[14] We can use this setting to investigate whether our algorithm is able to recover estimands such as ATE and ATT in the setting where covariate-driven treatment effect heterogeneity exists.

### Data-Generating Process

We adopt a fairly simple simulation setup that nonetheless captures the key features of the problem. We begin by assuming four independent covariates, $X_1, \ldots, X_4$, each distributed as an independent standard normal variable. We also posit a covariate $M$ that follows a Bernoulli distribution with $\Pr(M=1)=.5$. We distinguish this covariate because of its role as moderator of the treatment effect (made explicit below). Finally we assume an unmeasured covariate, $U$, that follows a Bernoulli distribution with $\Pr(U=1)=.5$.

Because this simulation is designed to verify that the algorithms behave as expected, we generate $Z$ and $Y$ according to a slight variation on our specified complete-data likelihood described in Equation 7,

$$\begin{aligned} Y|Z,U,X,M &\sim \mathrm{N}\left(\alpha_y + X\beta_y + \theta_y M + \zeta^y U + \tau Z + \tau_{\text{int}} MZ, \sigma^2_{\text{y}\cdot\text{xuz}}\right) \\ Z|X,U &\sim \text{Bernoulli}(\Phi(\alpha_z + X\beta_z + \theta_z M + \zeta^z U)), \end{aligned} \tag{13}$$

with $\alpha_y = -1.5$, $\beta_y = (.2, .4, .6, .8)$, $\theta_y = 1$, $\tau = -3$, $\tau_{\text{int}} = 6$, $\sigma^2_{\text{y}\cdot\text{xuz}} = 4$, $\alpha_z = -1.5$, $\beta_z = (.25, .25, .25, .25)$, and $\theta_z = 1$. This specification allows for treatment effects to vary based on the level of observed covariate $M$; in particular, observations with $M = 1$ have a treatment effect

[13] As far as we know the combination of sensitivity analysis and IPTW has only been used in one applied article (Frank et al., 2008); that article did not discuss the specific assumptions that accompany this technique.

[14] Simulations for the continuous treatment scenario are not shown given that the mathematical derivations for that scenario are straightforward.

of 3 and those with $M = 0$ have a treatment effect of $-3$. In turn it implies that ATE (which is always equal to 0 in expectation) is not equal to ATT (which varies with the level of $\zeta^z$). We vary the simulation over combinations of choices for the sensitivity parameters: $\zeta^z = (-2, -1, 0, 1, 2)$ and $\zeta^y = (0, 1, 2)$.

## *Estimands*

There are several estimands of interest in this simulation. One set of estimands represents different ways of conceptualizing the average treatment effect. One such estimand is the population average treatment effect (PATE; which is 0). We also consider the model-based sample approximation to this, which is the estimated coefficient on $Z$ from a regression of $Y$ on $Z$, $X$, $M$, and $U$ for our sample; in other words, this estimand represents the coefficient we would have estimated had we had access to $U$ as an observed covariate. We refer to this estimand as the regression average treatment effect (RATE).

A second set of estimands represents ways of conceptualizing the average effect of the treatment on the treated, which differs from PATE because of the treatment effect modification. This includes the population average treatment effect for the treated (PATT). PATT varies by $\zeta^z$; when $\zeta^z = (-2, -1, 0, 1, 2)$, $PATT = (.71, .94, 1.04, .94, .71)$. We also consider the model-based sample approximation to these, which is the estimated coefficient on $Z$ from a weighted regression of $Y$ on $Z$, $X$, $M$, and $U$ for our sample, with weights corresponding to those described in the "Use of Modification Weights to Target Specific Estimands" section. This is the coefficient we would have estimated using a standard inverse probability of treatment weighting (IPTW) approach had we had access to $U$ as an observed covariate; we call it the regression average treatment effect for the treated (RATT).[15]

## *Results*

Figures 2 and 3 present the simulation results that demonstrate the performance of our approach targeting two different types of estimands, ATE and ATT, respectively. Simulations are performed across 15 different combinations of the sensitivity parameters, which are presented in the top ($\zeta^z$) and the right ($\zeta^y$) of each figure. For each distinct combination of $\zeta^z$ and $\zeta^y$, we present three box plots. The box plot in the middle is a summary of our sensitivity analysis estimates, $\widehat{\tau}(\zeta^z, \zeta^y)$, over 1,000 simulations. The box plot on the left is a summary of the estimated treatment effects from a regression of Y on Z, X, M, and $U$, had we had access to $U$, RATE in Figure 2 and RATT in Figure 3. The box plots to the right are the counterparts from the regression of Y on Z, X, and M assuming no access to the unmeasured confounder; that is, the regression estimate based purely on observed data. Thus, the left box plots can be thought of as the best possible results, while the right box plots are the worst case scenario. Finally, PATE and PATT, displayed as red solid line segments in Figure 2 and Figure 3, respectively, represent the true population level parameters.

[15] We choose not to present results for PATC and RATC for the sake of parsimony since, due to the nature of the problem (e.g., PATE is just a weighted average of PATT and PATC) they provide no additional information in this setup.

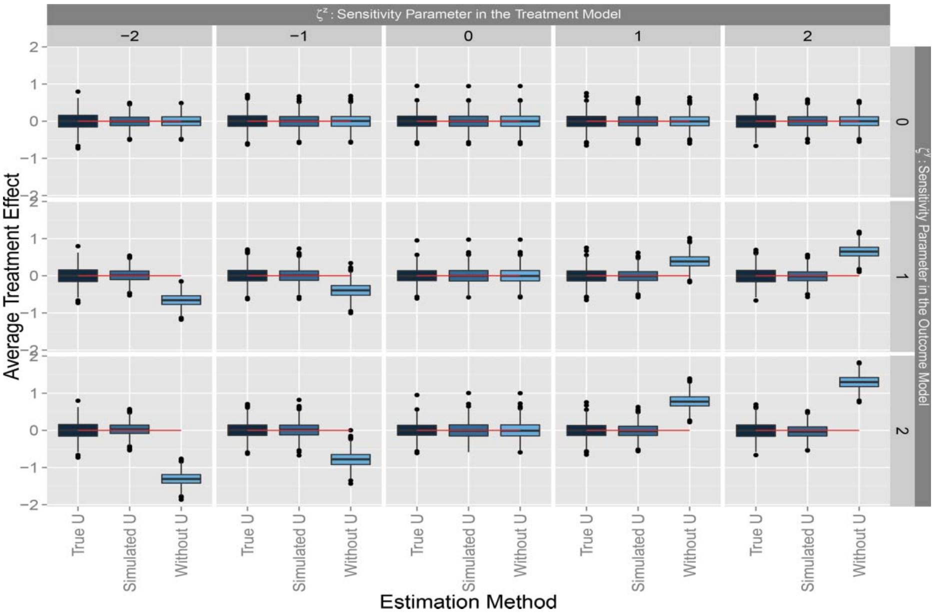

**Figure 2.** Sensitivity analysis results from the simulation for estimates targeting ATE estimands (PATE, RATE). Each box plot displays the distribution of estimates/estimands across 1,000 simulations. The left box plot in each cell displays RATE. The middle box plot displays the sensitivity analysis results. The right box plot displays regression estimates that exclude *U*. Each set of three box plots varies across combinations of the sensitivity parameters (labels on the top and to the right).

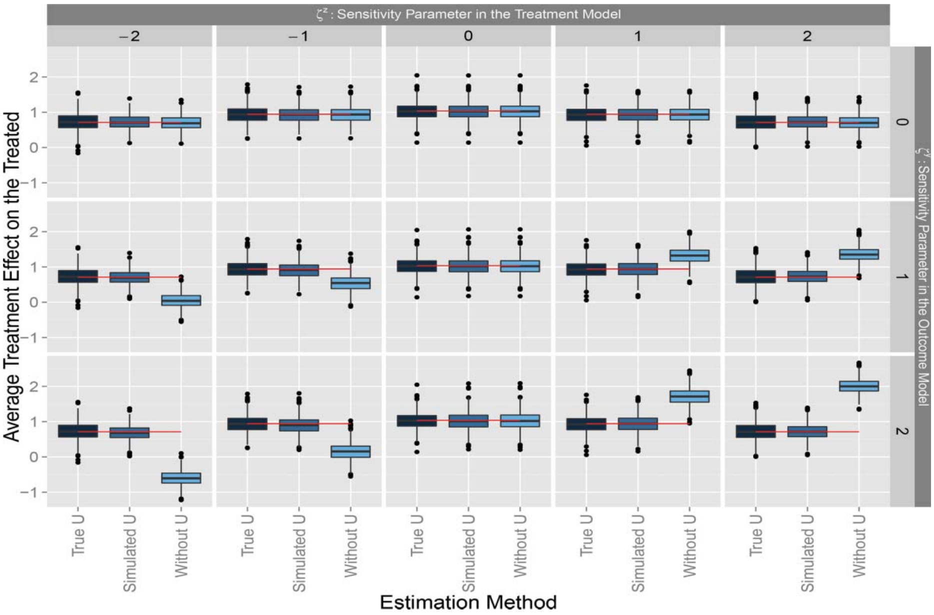

**Figure 3.** Sensitivity analysis results from the simulation for estimates targeting ATT estimands (PATT, RATT). Each box plot displays the distribution of estimates/estimands across 1,000 simulations. The left box plot in each cell displays RATT. The middle box plot displays the sensitivity analysis results. The right box plot displays regression estimates that exclude *U*. Each set of three box plots varies across combinations of the sensitivity parameters (labels on the top and to the right).

These figures show that our approach precisely recovers estimands corresponding to a given set of sensitivity parameters. In both figures, as long as either $\zeta^z$ or $\zeta^y$ is 0, the medians of all three box plots correspond with the red line. This indicates that the treatment effects are estimated without bias from their population treatment effects. When $\zeta^y$ is positive, the median of the right box plot is below the red line when $\zeta^z$ is negative and above the red line when $\zeta^z$ is positive. This indicates that the treatment effects estimated from a regression of Y on Z, X, and M are estimated with bias when both of the sensitivity parameters are nonzero. On the other hand, the left and middle box plots are centered on the red line representing the population parameter. Furthermore, the dispersion of our estimates is roughly equivalent to that of RATE or RATT, which is reassuring. This indicates that our approach performs as well as the regression estimate using the true model including *U* with regard to these measures, and the treatment effect estimates from our approach are both unbiased for both of these estimands.

## Continuous Treatment Variable Example: Child Care and Externalizing Behavior in Young Children

Given changing patterns of maternal employment in the United States over the past few decades, there is a great deal of interest in understanding the effects that increasing amounts of nonparental care in early childhood might have on children's development. One area that has been particularly controversial is the long-standing debate in developmental psychology over whether placement in child care leads to higher levels of "externalizing behavior"—aggression, impulsivity, or disobedience—in young children and in the early school years. Excessive externalizing behavior has been seen to be associated with poor academic performance and adverse social consequences from school age through adulthood (Campbell, 2002; Farmer, Bierman, & The Conduct Problems Prevention Research Group, 2002; Levenston, 2002; Zahn-Waxler, Usher, Suomi, & Cole, 2005).

Empirical results on this question are inconsistent, with some studies showing adverse effects on externalizing behavior of early initiation of nonparental care or greater hours in nonparental care (Bates et al., 1994; Baydar & Brooks-Gunn, 1991; Belsky, 1999; Crockenberg & Litman, 1991; Egeland & Hiester, 1995; Han, Waldfogel, & Brooks-Gunn, 2001; Haskins, 1985; Hofferth, 1999; Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Rabinovich, Zaslow, Berman, & Hyman, 1987; Rubenstein, Howes, & Boyle, 1981; Schwarz, Strickland, & Krolick, 1974; Vandell & Corasaniti, 1990; Youngblade, Kovacs, & Hoffman, 1999), while others show no effect or positive effects (Anme & Segal, 2004; Bacharach & Baumeister, 2003; Crockenberg & Litman, 1991; Field, Masi, Goldstein, & Perry, 1988; Greenstein, 1993; Harvey, 1999; Howes, 1988; Macrae & Herbert-Jackson, 1976; McCartney & Rosenthal, 1991). These studies must naturally be observational, as the need for care is determined by the situation of any given family.[16] However, it seems implausible that we could measure all covariates that might affect both the child care choice and subsequent externalizing behaviors. Thus, this question is a prime candidate for sensitivity analysis. It is important to understand how sensitive results from observational studies might be to a potential unmeasured confounder; for instance, how predictive would such a confounder

[16] Even if we could randomize some families to receive center-based care, we could not force them to participate in this care or force those in the control group to not participate in center-based care.

have to be to remove support for the hypothesis that nonparental child care impacts externalizing behavior?

We explore this question using data from the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (SECCYD) because several influential studies on this topic were conducted using these data. Moreover, the NICHD study is one of the largest, most comprehensive studies of the developmental effects of early child care experience. It followed from birth a cohort of more than 1,300 children from 10 communities across the United States, recording detailed information on the quantity and quality of care experienced, family and socioeconomic variables, and cognitive and behavioral outcomes. The design of the study has been described in detail elsewhere (see, e.g., National Institute of Child Health and Human Development Early Child Care Research Network, 1998).

### Data and Variables

Earlier work considering externalizing behavior from the SECCYD data has shown some evidence for an increase in problem behavior with more time in nonparental care (McCartney et al., 2010; National Institute of Child Health and Human Development Early Child Care Research Network, 1998, 2003). Following the later studies, we focus on externalizing behavior measured by caregivers at 54 months as an outcome.[17] To demonstrate the performance of the sensitivity analysis method developed for continuous treatment variables, we use the average hours per week in nonparental care from age 24 to 50 months as our treatment.[18]

Following the convention from previous studies, we include the following pretreatment variables as potential confounders: child externalizing behavior measured at age 24 months, an indicator for whether the child is female, mother's educational attainment, mother's anxiety/depression score at 24 months, total income-to-needs ratio at 24 months, poverty (total income-to-needs ratio less than 2) at 24 months, whether the father is living in the same household at age 24 months, and indicators for data collection locations. Although previous attempts to address this question often included values of time-varying covariates at posttreatment surveys in the analysis, we did not do so due to the potential for bias that inclusion of posttreatment covariates can yield (Gelman & Hill, 2007; Rosenbaum, 1984). We also standardize all continuous variables to create a common scale for interpreting and comparing the magnitudes of the coefficients on the observed covariates relative to the sensitivity parameters for the unmeasured confounder, as discussed in the "Visualization and Interpretation of Results" section.

### Results

Figure 4 contains an analysis of the sensitivity of the estimated coefficient of the continuous treatment to potential unmeasured confounding.[19] Most of the observed covariates have a

[17] When a child's behavior was evaluated by multiple caregivers, the average of their evaluations was used.

[18] Because of a seven-month gap between the Phase I and Phase II studies and the fact that hours per week in care are recorded every four months in Phase II, the variable for average hours in care is created with the following formula: $\overline{care}_{24\sim 50} = \frac{1}{27}\left(\sum_{m=24}^{34} care_m + 8\cdot care_{42} + 4\cdot care_{46} + 4\cdot care_{50}\right)$. Because we observed excessively large entries for this continuous treatment in some cases, we capped the maximum value of time in care at 70 hours.

[19] We use a 10 by 20 grid with 100 draws of $\dot{U}$ per cell.

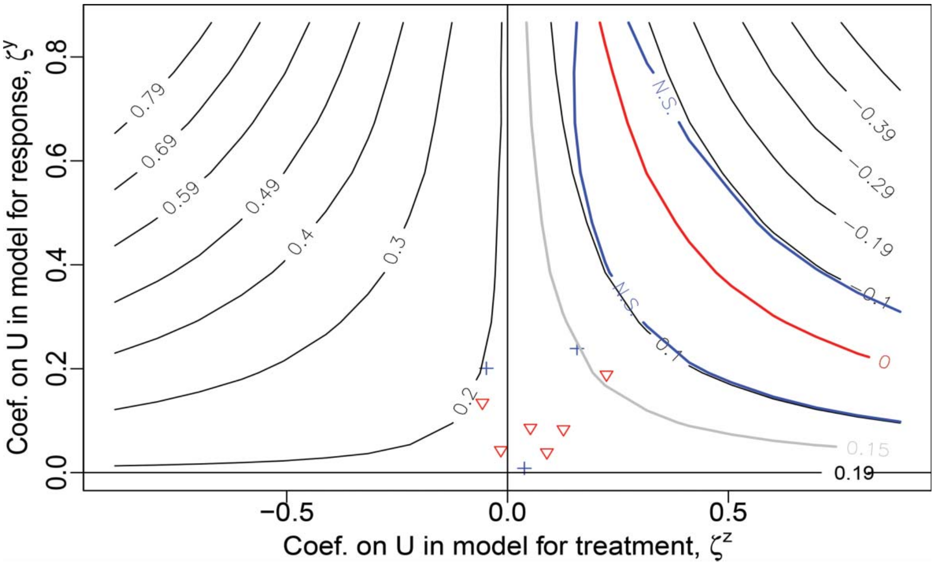

**Figure 4.** Sensitivity analysis results from an analysis of the effect of average hours of nonparental care on externalizing behavior using data from the NICHD study.

negative association with externalizing behavior and have been rescaled (multiplied by −1) so that the estimated coefficients will be positive; these are represented by ▽ on the plot. All continuous covariates have also been standardized to have a mean of 0 and standard deviation of 1 to facilitate comparisons of their magnitude to that of the sensitivity parameters.

This figure shows that, over a range of plausible sensitivity parameters, the estimated treatment effects range from roughly −.5 to .8, with the majority of the range consistent with a significant, positive treatment effect. The blue reference lines indicate that the sensitivity parameters required for an unmeasured confounder to drive the estimate to nonsignificance are larger than the coefficients for the observed confounders by about 50%. To drive the estimate to zero (red reference line), these parameters would have to be about 200% larger than the largest coefficient estimate for the observed covariates. Such values are not implausible. For instance, consider a confounder with $\zeta^z = .5$ and $\zeta^y = .4$. This corresponds to a continuous confounder with the following property: when comparing two groups that differ by one standard deviation on this confounder they would also differ by .5 standard deviations on the treatment variable (which corresponds to about 6 hours per week of nonparental care). This confounder would also have the property that when comparing two groups that differ by one standard deviation on this confounder they would also differ by .4 standard deviations on the outcome variable (which corresponds to 4 on the externalizing behavior index).

## Binary Treatment Variable Example: National Supported Work Constructed Observational Study

Labor economists have long been interested in whether job training has the potential to improve workers' human capital. However, the answer to this question is often hampered by selection bias. That is, those who are likely to participate in a job training program tend to have different earnings potential than those who do not choose to participate. This earnings potential may or may not be proxied by observed confounders. Thus, several social experiments have been conducted. The National Supported Work Demonstration (NSW) was one such experiment. In the NSW evaluation, the samples were randomly assigned to participate in a job training program or not, and thus it was straightforward to unbiasedly estimate the effect of participation in this program on subsequent yearly earnings (in 1978).

### Constructed Observational Study Data

LaLonde (1986) created a unique constructed observational data set by merging the experimental treatment group from NSW with a nonexperimental control group drawn from the Population Survey of Income Dynamics (PSID). Both surveys measured the following baseline covariates that can be used as potential confounders: education, age, black, Hispanic, married, earning in 1974, earning in 1975, unemployed in 1974, and unemployed in 1975. These variables are also representative of the types of covariates that were used in standard observational analyses of this research question using other data during that time period. LaLonde examined how well the sophisticated econometric techniques at that time could recover the experimental estimates obtained using the NSW data set and discovered that none of them were able to do so reliably. Since then, the data set has been used as a touchstone for new statistical methods that attempt to remove selection bias due to observables in observational studies.

Perhaps not surprisingly, simple estimators such as the coefficient on the treatment variable in a linear regression of the outcome on the treatment indicator and observed confounders fail to recover the experimental benchmark as well. Two potential explanations for this failure present themselves. The first is that linear regression is targeting a different estimand, a form of ATE (Angrist & Pischke, 2008), rather than the effect of the treatment on those who actually participate in the program (ATT in this constructed observational study); assuming there are differences in the characteristics between the treatment and control groups and these differences also drive differences in the sizes of the effects of the programs, these estimands will not be the same. The second explanation is that linear regression is producing biased results (either because it does not satisfy ignorability or because it relies on incorrect parametric assumptions); that is, even if ATE=ATT for these data, linear regression would not be estimating the ATT unbiasedly.

We know from subsequent work on this topic that methods that provide robustness to model misspecification and that are able to target estimands such as the ATT—such as propensity score matching and IPTW—have been reasonably successful at reliably recovering the experimental benchmarks in this example when the full set of controls is used (see, e.g., Dehejia, 2005; Dehejia & Wahba, 1999).[20] However, one drawback of evidence obtained from constructed observational studies such as this one is that we cannot ascertain whether a “success” (which occurs when a method closely mirrors the experimental benchmark) reflect a situation in which ignorability is satisfied or rather an example of bias cancellation.

[20] Notably, removal of 1974 earnings is particularly problematic for these data most probably due to the so-called “Ashenfelter dip” (due to Ashenfelter, 1978), which describes the phenomenon that those who enroll in job training programs are more likely to have experienced job loss right before the participation (implying that conditioning on only one period of pretreatment earnings is likely not sufficient for establishing ignorability).

Imbens (2003) used these data to illustrate his approach to sensitivity analysis as well. However, his sensitivity analysis strategy cannot differentially target the effect of the treatment on the treated. Yet the estimand of interest in the NSW-constructed observational study (the one that corresponds to the experimental benchmark) is the effect of the treatment on the treated (because in the original experiment, the randomized control group is not systematically different from the treatment group, so estimating the effect for the treated is the same as estimating the average effect for the randomized experiment). We compare the ATE and ATT results obtained from our sensitivity analysis strategy to observe the differences in our conclusions based on targeting the proper estimand.

## *Results*

Figure 5 shows the results of the sensitivity analysis that examines the sensitivity to unmeasured confounding for each of two estimates of the effect of participation in the NSW job training program. The left panel does not use weights and thus targets an average treatment effect (ATE). The right panel uses modification weights to target the ATT. We use standardized coefficients to facilitate comparisons between the coefficients on observed confounders and the sensitivity parameters for our posited binary unmeasured confounder. These benchmarking coefficients are more strongly predictive of the treatment assignment in the unweighted sample than in the weighted sample, which is not surprising since the weighted sample creates greater balance between the treatment and (weighted) control groups. With one exception, the observed covariates have either similar or reduced power to predict the outcome variable in the weighted sample as well.

The left panel suggests that absent unmeasured confounding the effect estimate would be about 0. It further suggests that it would take a reasonably strong unmeasured confounder that was negatively associated with treatment but positively associated with the outcome to yield a significant positive effect (such effects correspond to the upper-left corner of the plot). One such confounder does exist in the observed data, however. On the other hand, if the unmeasured confounder were strongly positively associated with both treatment and outcome this would suggest a significant negative treatment effect estimate (upper-right corner of the plot).

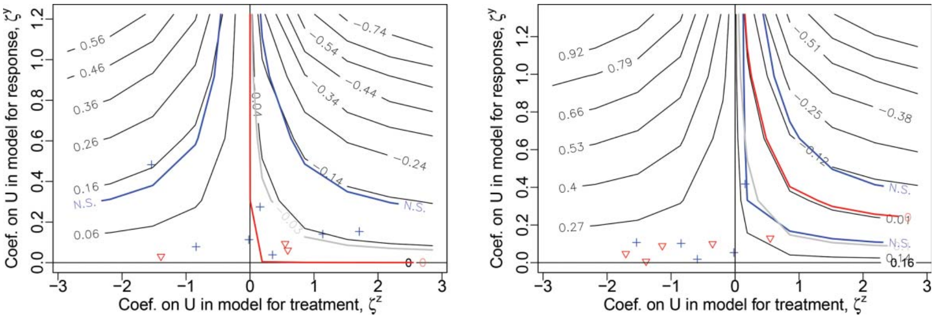

**Figure 5.** Sensitivity analysis results from the NSW. Left panel displays results from algorithm that uses no weights. Right panel displays results from algorithm that uses modification weights to target the effect of the treatment on the treated.

Consider instead the right panel with sensitivity analyses that target the ATT. Here more than half of the plot corresponds to significant, positive treatment effect estimates. The level of unobserved confounding (as operationalized by the sensitivity parameters) required to drive this estimate to nonsignificance is represented by the blue lines labeled N.S. As a comparison to these values, consider that the strongest observed confounder in the data has coefficients of 0.4 and 0.2 in the response and treatment models, respectively. Therefore it seems plausible that the omitted confounder might be equally strong. On the other hand, the omitted confounder required to drive the treatment effect estimate to zero, this confounder (as represented by the red line labeled "0") would have had to be almost twice this strong.

Furthermore, the "naive" treatment effect estimate (that is the regression estimate on observed data), which is close to zero when targeting the ATE, is significant and positive when targeting the ATT. The larger effects for ATT suggest that the job training program was more effective for the workers who were more likely to participate in the program. The standardized coefficient that targets the ATT corresponds to a treatment effect estimate of about 2,400 dollars per year with a standard error of approximately 700 dollars. This compares favorably with the treatment effect estimate from the original randomized experiment (which as previously noted targets the same estimand—the effect of the program on those who participated) which was approximately 1,800 dollars per year. The different conclusions reached from these two plots highlights the importance of targeting an appropriate estimand when performing sensitivity analysis.

## Discussion

This article describes a general approach to sensitivity analysis with specific implementations for models with continuous and binary treatment variables. Our approach makes use of the conditional distribution of the unmeasured confounder implied by each of the two models to generate candidate realizations of the unmeasured confounder. This allows us to estimate the distribution of treatment effect estimates that correspond to an analysis that conditions on that confounder. Conveniently, the unmeasured confounder is characterized by sensitivity parameters that can be interpreted as regression coefficients (from the regression of the outcome on the treatment, observed covariates, and unmeasured covariate and from the regression of the treatment on the observed covariates and the unmeasured covariate). These coefficients are interpretable to applied researchers who have substantial experience in interpreting regression coefficients. In addition, this specification allows us to benchmark the magnitude of the sensitivity parameters relative to the corresponding coefficients for observed confounders in our data.

An important contribution of our approach is the extension of the binary treatment variable implementation to allow for targeting of the effect of two causal estimands that are often of strong interest for policy and practice: the effect of the treatment on the treated and the effect of the treatment on the controls. Given the complexity of the algorithm for the binary treatment, particularly combined with the weighted estimation needed for targeting ATT and ATC, we have performed simulations to verify the properties of the algorithm. Across a wide range of combinations of sensitivity parameters, the algorithm reliably recovers a distribution nearly equivalent to that of the regression estimates in a regression controlling for the corresponding unmeasured confounder (i.e., it mimics the sampling distributions of RATE or RATT). The NSW example highlights the potential practical implications of this feature of our strategy.

This approach to assessing sensitivity to unmeasured confounding described in this article is a useful complement to any observational study for which it is unclear if ignorability has been satisfied. Software that implements these algorithms is currently available in the treatSens package on the CRAN archive (for the R software program).

## Funding

This research was partially supported by Institute of Education Sciences grant R305D110037.

## ARTICLE HISTORY

Received 16 December 2013
Revised 17 December 2014
Accepted 29 July 2015

## EDITORS

This article was reviewed and accepted under the editorship of Carol McDonald Connor and Spyros Konstantopoulos.

## References

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, *113*, 151–184.

Angrist, J., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Anme, T., & Segal, U. A. (2004). Implications for the development of children in over 11 hours of centre-based care. *Child: Care, health and development*, *30*, 345–352.

Ashenfelter, O. (1978). Estimating the effects of training programs on earnings. *Review of Economics and Statistics*, *60*, 47–57.

Bacharach, V. R., & Baumeister, A. A. (2003). Child care and severe externalizing behavior in kindergarten children. *Journal of Applied Developmental Psychology*, *23*, 527–537.

Bates, J. E., Marvinney, D., Kelly, T., Dodge, K. A., Bennett, D. S., & Pettit, G. S. (1994). Child care history and kindergarten adjustment. *Developmental Psychology*, *30*, 690–700.

Baydar, N., & Brooks-Gunn, J. (1991). Effects of maternal employment and child-care arrangements on preschoolers' cognitive and behavioral outcomes: Evidence from the Children of the National Longitudinal Survey of Youth. *Developmental psychology*, *27*, 932–945.

Becker, S. O., & Caliendo, M. (2007). Sensitivity analysis for average treatment effects. *Stata Journal*, *7*, 71–83.

Belsky, J. (1999). Quantity of nonmaternal care and boys' problem behavior/adjustment at ages 3 and 5: Exploring the mediating role of parenting. *Psychiatry: Interpersonal and Biological Processes*, *62*, 1–20.

Bross, I. D. (1966). Spurious effects from an extraneous variable. *Journal of Chronic Diseases*, *19*, 637–647.

Bross, I. D. (1967). Pertinency of an extraneous variable. *Journal of Chronic Diseases*, *20*, 487–495.

Buckley, J., & Schneider, M. (2007). *Charter schools: Hype or hope?*. Princeton, NJ: Princeton University Press.

Campbell, S. B. (2002). *Behavior problems in preschool children: Clinical and developmental issues*. New York, NY: Guilford Press.

Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*, 55–95.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, *22*, 173–203.

Crockenberg, S., & Litman, C. (1991). Effects of maternal employment on maternal and two-year-old child behavior. *Child Development*, *62*, 930–953.

Dehejia, R. H. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics*, *120*, 355–364.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *JASA*, *94*, 1053–1062.

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, *34*, 271–310.

Egeland, B., & Hiester, M. (1995). The long-term consequences of infant day-care and mother-infant attachment. *Child Development*, *66*, 474–485.

Farmer, A. D., Bierman, K. L., & The Conduct Problems Prevention Research Group (2002). Predictors and consequences of aggressive-withdrawn problem profiles in early grade school. *Journal of Clinical Child and Adolescent Psychology*, *31*, 299–311.

Field, T., Masi, W., Goldstein, S., & Perry, S. (1988). Infant day care facilitates preschool social behavior. *Early Childhood Research Quarterly*, *3*, 341–359.

Frank, K. A. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research*, *29*, 147–194.

Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Extended influence: National board certified teachers as help providers. *Education, Evaluation, and Policy Analysis*, *30*, 3–30.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, *25*, 1107–1116.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*, 29–46.

Greenstein, T. N. (1993). Maternal employment and child behavioral outcomes: A household economics analysis. *Journal of Family Issues*, *14*, 323–354.

Gustafson, P., McCandless, L., Levy, A., & Richardson, S. (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics*, *66*, 1129–1137.

Han, W.-J., Waldfogel, J., & Brooks-Gunn, J. (2001). The effects of early maternal employment on later cognitive and behavioral outcomes. *Journal of Marriage and Family*, *63*, 336–354.

Harada, M. (2011). *ISA: Stata module to perform Imbens' (2003) sensitivity analysis*. Boston, MA: Statistical Software Components, Boston College Department of Economics.

Harada, M. (2012). *GSA: Stata module to perform generalized sensitivity analysis*. Boston, MA: Statistical Software Components, Boston College Department of Economics.

Harada, M. (2013). *Generalized sensitivity analysis* (Technical report). New York, NY: New York University.

Harvey, E. (1999). Short-term and long-term effects of early parental employment on children of the National Longitudinal Survey of Youth. *Developmental Psychology*, *35*, 445–459.

Haskins, R. (1985). Public school aggression among children with varying day-care experience. *Child Development*, *56*, 689–703.

Hill, J. L., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*, 477–513.

Hofferth, S. (1999). *Child care in the first three years of life and preschoolers' language and behavior*. Paper presented at the biennial meeting of the Society for Research in Child Development, Albuquerque, NM.

Howes, C. (1988). Relations between early child care and schooling. *Developmental Psychology*, *24*, 53–57.

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, *23*, 305–327.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, *93*(2), 126–132.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29.

Keele, L. (2010). *An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data*. White Paper. Retrieved from http://www.personal.psu.edu/ljk20/rbounds%20vignette.pdf

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of non-uniform effect. *American Journal of Epidemiology*, *163*, 262–270.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs. *American Economic Review*, *76*, 604–620.

Levenston, G. K. (2002). A longitudinal analysis of childhood externalizing and internalizing psychopathology as risk factors for criminal offending in adulthood. *Dissertation Abstracts International: Section B: Sciences and Engineering*, *62*, 5381.

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*, 948–963.

Macrae, J. W., & Herbert-Jackson, E. (1976). Are behavioral effects of infant day care program specific? *Developmental Psychology*, *12*, 269–270.

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal*, *41*, 115–157.

McCandless, L. C., Gustafson, P., & Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, *26*, 2331–2347.

McCartney, K., Burchinal, M., Clarke-Stewart, A., Bub, K. L., Owen, M. T., & Belsky, J. (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology*, *46*, 1–17.

McCartney, K., & Rosenthal, S. (1991). Maternal employment should be studied within social ecologies. *Journal of Marriage and Family*, *53*, 1103–1107.

Nannicini, T. (2007). Simulation-based sensitivity analysis for matching estimators. *Stata Journal*, *7*, 334.

National Institute of Child Health and Human Development Early Child Care Research Network. (1998). Early child care and self-control, compliance, and problem behavior at twenty-four and thirty-six months. *Child Development*, *69*, 976–1005.

National Institute of Child Health and Human Development Early Child Care Research Network. (2003). Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten? *Child Development*, *74*, 976–1005.

Neilsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*, 457–489.

Rabinovich, B., Zaslow, M., Berman, P., & Hyman, R. (1987). *Employed and homemaker mothers' perceptions of their toddlers compliance behavior in the home*. Poster presented at the meeting of the Society for Research in Child Development, Baltimore, MD.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A, General*, *147*, 656–666.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*, 13–26.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika*, *75*, 577–581.

Rosenbaum, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics*, *47*, 87–100.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*, 286–327.

Rosenbaum, P. R. (2002b). *Observational studies*. New York, NY: Springer.

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, *105*, 692–702.

Rosenbaum, P. R. (2011). A new u-statistic with superior design sensitivity in observational studies. *Biometrics*, *67*, 1017–1027.

Rosenbaum, P. R., & Krieger, A. M. (1990). Sensitivity of two-sample permutation inferences in observational studies. *Journal of American Statistical Association*, *85*, 493–498.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, *45*, 212–218.

Rosenbaum, P. R., & Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, *104*, 1398–1405.

Rubenstein, J. L., Howes, C., & Boyle, P. (1981). A two-year follow-up of infants in community-based day care. *Journal of Child Psychology and Psychiatry*, *22*, 209–218.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Schwarz, J. C., Strickland, R. G., & Krolick, G. (1974). Infant day care: Behavioral effects at preschool age. *Developmental Psychology*, *10*, 502–506.

Steenland, K., & Greenland, S. (2004). Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology*, *160*, 384–392.

Vandell, D. L., & Corasaniti, M. A. (1990). Child care and the family: Complex contributors to child development. *New Directions for Child and Adolescent Development*, *49*, 23–37.

Youngblade, L., Kovacs, D., & Hoffman, L. (1999). *The effects of early maternal employment on 3rd and 4th grade children's social development*. Poster presented at the biennial meeting of the Society for Research in Child Development, Albuquerque, NM.

Zahn-Waxler, C., Usher, B., Suomi, S., & Cole, P. M. (2005). Intersections of biology and behavior in young children's antisocial patterns: The role of development, gender and socialization. In D. M. Stoff & E. J. Susman (Eds.), *Developmental psychobiology of aggression* (pp. 141–160). New York, NY: Cambridge University Press.

## Appendix: Derivation of Conditional Distribution of U in the Multivariate Normal Case

Recall that when $U$, $Z$, and $Y$ are normally distributed, we specify the complete-data likelihood as:

$$\begin{aligned} U &\sim \mathrm{N}\left(0,\sigma_u^2\right) \\ Z|X,U &\sim N\left(\beta^z x+\zeta^z u,\sigma_{\mathrm{z\cdot xu}}^2\right) \\ Y|X,U,Z &\sim N\left(\beta^y x+\zeta^y u+\tau z,\sigma_{\mathrm{y\cdot xuz}}^2\right) \end{aligned} \tag{A1}$$

Varying $\sigma^2{}_u$ only changes the range of possible coefficients, not the qualitative results of the sensitivity analysis. The mean and variance of $Y|X$, $Z$ are modified equivalently to defining sensitivity parameters $\zeta^y$ and $\zeta^z$ by setting $\zeta' = \sigma_u\zeta$. Thus, the shape of the distribution of $\widehat{\tau}$ will be unchanged, but varying $\sigma^2{}_u$ has implications for benchmarking according to coefficients of $X$, because changing the assumed variance will change the relationship to the observed $X$. For this reason, we take $\sigma^2{}_u = 1$ and standardize all variables in the analysis.

This yields a joint distribution

$$\begin{aligned} \mathrm{p}(U, Y, Z|X) = (2\pi)^{-1/2}\exp\left(-\frac{u^2}{2}\right) \\ \cdot\left(2\pi\sigma^2_{\mathrm{z\cdot xu}}\right)^{-1/2}\exp\left(-\frac{(z-\beta^z x-\zeta^z u)^2}{2\sigma^2_{\mathrm{z\cdot xu}}}\right) \\ \cdot\left(2\pi\sigma^2_{\mathrm{y\cdot xuz}}\right)^{-1/2}\exp\left(-\frac{(y-\beta^y x-\zeta^y u-\tau z)^2}{2\sigma^2_{\mathrm{y\cdot xuz}}}\right) \end{aligned} \tag{A2}$$

After factoring $\mathrm{p}(U, Y, Z|X)$, we obtain the following conditional distributions for $U$ and $(Y, Z)$:

$$\begin{aligned} U|X, Z, Y &\sim N\left(\mu_{\mathrm{u\cdot xzy}}, \sigma^2_{\mathrm{u\cdot xzy}}\right) \\ \mu_{\mathrm{u\cdot xzy}} &= \frac{\sigma^2_{\mathrm{y\cdot xuz}}\zeta^z(z-\beta^z x)+\sigma^2_{\mathrm{z\cdot xu}}\zeta^y(y-\beta^y x-\tau z)}{\sigma^2_{\mathrm{z\cdot xu}}\sigma^2_{\mathrm{y\cdot xuz}}+\sigma^2_{\mathrm{y\cdot xuz}}\zeta^{z2}+\sigma^2_{\mathrm{z\cdot xu}}\zeta^{y2}} \\ \sigma^2_{\mathrm{u\cdot xzy}} &= \frac{\sigma^2_{\mathrm{z\cdot xu}}\sigma^2_{\mathrm{y\cdot xuz}}}{\sigma^2_{\mathrm{z\cdot xu}}\sigma^2_{\mathrm{y\cdot xuz}}+\sigma^2_{\mathrm{y\cdot xuz}}\zeta^{z2}+\sigma^2_{\mathrm{z\cdot xu}}\zeta^{y2}} \end{aligned} \tag{A3}$$

and

$$\begin{aligned} Y, Z|X &\sim N\left(\mu_{\mathrm{yz\cdot x}}, \textstyle\sum_{\mathrm{yz\cdot x}}\right) \\ \mu_{\mathrm{yz\cdot x}} &= (\beta^y x+\tau z, \beta^z x) \\ \Sigma_{\mathrm{yz\cdot x}} &= \begin{bmatrix} \sigma^2_{\mathrm{y\cdot xuz}}+\zeta^{y2} & \zeta^y\zeta^z \\ \zeta^y\zeta^z & \sigma^2_{\mathrm{z\cdot xu}}+\zeta^{z2} \end{bmatrix} \end{aligned} \tag{A4}$$

We see that the distribution in (A3) is a function of the residuals, conditional variances, and coefficients from the underlying true models for $Y$ and $Z$ including $U$. Because $U$ is unobservable, we would like to rewrite this distribution in terms of what we can observe in our data, plus sensitivity parameters that the analyst can vary. To do so, we need the conditional distributions of $Y$ and $Z$ without $U$. We can obtain the distribution of $Z$ given $X$ in a similar manner to that of $\mathrm{p}(Y, Z|X)$, which yields

$$Z|X \sim N\left(\beta^z x, \sigma^2_{\mathrm{z\cdot xu}}+\zeta^{z2}\right) \tag{A5}$$

Thus, the observed residuals $\tilde{z}$ have mean $z-\beta^z x$ and variance $\sigma^2_{\tilde{z}}=\sigma^2_{\mathrm{z\cdot xu}}+\zeta^{z2}$.

From the bivariate normal distribution of $Y$ and $Z$ given $X$ in A4, we see that the distribution of $Y$ given $Z$ and $X$ is

$$Y|X,Z \sim N\left(\beta^y x + \tau z + \frac{\zeta^y \zeta^z}{\sigma^2_{\mathrm{z\cdot xu}} + \zeta^{z2}}(z - \beta^z x), \frac{\sigma^2_{\mathrm{z\cdot xu}}\sigma^2_{\mathrm{y\cdot xuz}} + \sigma^2_{\mathrm{y\cdot xuz}}\zeta^{z2} + \sigma^2_{\mathrm{z\cdot xu}}\zeta^{y2}}{\sigma^2_{\mathrm{z\cdot xu}} + \zeta^{z2}}\right) \tag{A6}$$

This implies that the residuals $\tilde{y}$ from a linear model of $Y$ as a function only of the observed data ($Z$, $X$) have mean $(y - \beta^y x + \tau z) - c\,\tilde{z}$and variance $\sigma^2_{\tilde{y}} = \frac{\sigma^2_{\mathrm{z\cdot xu}}\sigma^2_{\mathrm{y\cdot xuz}} + \sigma^2_{\mathrm{y\cdot xuz}}\zeta^{z2} + \sigma^2_{\mathrm{z\cdot xu}}\zeta^{y2}}{\sigma^2_{\mathrm{z\cdot xu}} + \zeta^{z2}}$. Note that $c$ also quantifies the bias in the estimate of $\tau$ for given coefficients $\zeta^y$ and $\zeta^z$, with denominator given by $\sigma^2_{\tilde{z}}$. Thus, in this limited case we can directly calculate the bias in $\widehat{\tau}$ associated with particular values of $\zeta^z$ and $\zeta^y$ as a function of the variance in $\tilde{z}$.

Given these distributions, we can now rewrite the conditional distribution of $U$ in terms of observables and the coefficients of $U$:

$$\begin{aligned}
&U|X,Z,Y \sim N\left(\mu_{\mathrm{u\cdot xzy}}, \sigma^2_{\mathrm{u\cdot xzy}}\right)\\
&\mu_{\mathrm{u\cdot xzy}} = \frac{\sigma^2_{\tilde{y}}\zeta^z\,\tilde{z} + \left(\sigma^2_{\tilde{z}} - \zeta^{z2}\right)\zeta^y\,\tilde{y}}{\sigma^2_{\tilde{z}}\sigma^2_{\tilde{y}}}\\
&\sigma^2_{\mathrm{u\cdot xzy}} = \frac{1}{\sigma^2_{\tilde{z}}\sigma^2_{\tilde{y}}}\left(\frac{\sigma^2_{\tilde{z}}\left(\sigma^2_{y} - \zeta y^2\right) + \zeta^{z2}\zeta^{y2}}{\sigma^2_{\tilde{z}}}\right)\left(\sigma^2_{\tilde{z}} - \zeta^{z2}\right)
\end{aligned} \tag{A7}$$

So if we treat $\zeta^z$ and $\zeta^y$ as sensitivity parameters, we can generate $\dot{U}$ from this distribution.

As an alternative, we consider using the partial correlations of $U$ with $Y$ and $Z$ as sensitivity parameters. Note that $\rho_{u\tilde{z}} = \mathrm{Cor}(U,Z|X) = \zeta^z/\sqrt{\zeta^{z2} + \sigma^2_{\mathrm{z\cdot xu}}} = \zeta^z\sigma_u/\sigma_{\tilde{z}}$ and $\rho_{u\tilde{y}} = \mathrm{Cor}(U, Y|X) = \left(\zeta^y/\sigma_{\tilde{y}}\right)\cdot\left(1 - \zeta^{z2}/\sigma^2_{\tilde{z}}\right) = \left(\zeta^y/\sigma_{\tilde{y}}\right)\left(1 - \rho^2_{u\tilde{z}}\right)$. Using these we can reparameterize the conditional distribution of $U$ as:

$$\begin{aligned}
&U|X,Z,Y \sim N\left(\mu_{\mathrm{u\cdot xzy}}, \sigma^2_{\mathrm{u\cdot xzy}}\right)\\
&\mu_{\mathrm{u\cdot xzy}} = \frac{\rho u\,\tilde{z}}{\sigma_{\tilde{z}}}\tilde{z} + \frac{\rho u\,\tilde{y}}{\sigma\,\tilde{y}}\tilde{y}\\
&\sigma^2_{\mathrm{u\cdot xzy}} = 1 - \rho^2_{u\,\tilde{z}} - \rho^2_{u\,\tilde{y}}
\end{aligned} \tag{A8}$$