Simulation-Extrapolatino for Estimating Principal Causal Effect Surfaces

Simulation Extrapolation for Estimating Principal Causal Effect Surfaces

Qualifying Paper

Submitted by

Marcus Waldman

October, 2017

Acknowledgements

This work would not be possible without the support, guidance, and advice from my committee members: Dr. Andrew Ho, Dr. Stephanie Jones, and Dr. Katherine Masyn. I also thank Dr. Joseph McIntyre and Dr. Luke Miratrix for their feedback and insights.

Table of Contents

Abstract 5

Introduction 6

Motivating Example 10

Background and Notation 12

Application to the MDRC Experiment 14

Modeling Assumptions 15

Relationship to Instrumental Variables 19

Bias in TSLS Estimates 21

Conditional Mean Imputation: Local Bias and a Correction 24

Conditional Mean Estimator and Omitted Variable Bias 26

Local Bias in Conditional Mean Imputations 27

Correcting Local Bias in Conditional Mean Imputations 29

SIMEX 31

General Background 31

Application to Estimating a PCES 33

Simulation Study Design 34

Measures 35

Simulation Setup 38

Sampling Scheme 40

Simulation 1: Evaluating TSLS & Conditional Mean Imputation Estimators 40

Simulation 2: Imbalance Resulting from a Univariate Imputation Estimator 42

Simulation 3: TSLS-Assisted SIMEX PCES Estimators 43

Simulation Results 44

Simulation 1 Results 44

Simulation 2 Results 46

Simulation 3 Results 47

Discussion and Future Directions 48

References 53

Appendix: Figures and Tables 58

Abstract

Amid the "big data" revolution, background information on participants is becoming ever more available for experimental researchers to predict treatment effect heterogeneity, including heterogeneity on some intermediate variable collected post-treatment. At the same time, the recently developed principal stratification framework allows researchers to assess heterogeneity on an intermediate variable in a manner that maintains causal interpretations. This paper details the shortcomings of two-stage least squares and imputation methods as viable estimators if used to assess treatment effect heterogeneity when the intermediate variable is continuous and traditional assumptions are not tenable. Results from an alternative estimator that relies on simulation-extrapolation is evaluated to inform future research.

Introduction

Researchers are increasingly interested in identifying heterogeneity in experimental treatment effects, including whether some post-treatment intermediate variable D moderates a causal relationship. Common applications of intermediate moderating variables include compliance status (Angrist, Imbens, & Rubin, 1996), censoring due to death (Rubin, 2006), surrogate outcomes (Gilbert & Hudgens, 2008), and level of exposure to opportunities that influence downstream outcomes of interest (Page, 2012). Because it is unlikely that post-treatment variables can be randomly assigned, assessing heterogeneity by conditioning on observed values of D results in inferences that are subject to self-selection bias. It is for this reason that some view assessing post-treatment moderators as a problem more akin to estimating causal effects in observational studies than in randomized experiments (Ding and Lu, 2016).

Principal stratification (PS; Frangakis & Rubin, 2002) applies the Rubin Causal Model (Rubin, 1974; Holland, 1986) to clarify causal inference in the presence of an intermediate variable. Both the observed values for the outcome of interest, Y, and moderating post-treatment variable, D, are understood to be realizations of potential outcomes. Principal causal effects are defined as any comparison of Y's potential outcomes between units that share similar D potential outcomes. Unlike the observed value of the intermediate variable, randomization ensures that the D potential outcomes

can be treated as an exogenous covariate. Thus, comparisons remain "apples-to-apples" by conditioning on the potential outcomes rather than the endogenous observed values.

In many applications, treatment and the intermediate variable take on a discrete set of values only. This implies that units can be stratified into a finite set of principal strata corresponding to the joint distribution of D's potential outcomes, and principal causal effects are estimated within each stratum. The fundamental challenge of inference is that the joint distribution is never fully observed for any unit, making membership into principal strata a latent variable because it can only be ascertained indirectly from the observed data. Thus, principal strata represent latent classes, and principal causal effects within each class are not identified without imposing some combination of: (a) simplifying modeling assumptions, such as monotonicity and the exclusion restrictions (e.g., Angrist et al., 1996), (b) parametric assumptions, such as including informative a priori assumptions (Hirano, Imbens, Rubin, & Zhou, 2000; Conlon, Taylor, & Elliott, 2017) and finite mixture modelling (Page, 2012), or (c) a rich set of exogenous predictors (Ding & Lu, 2017).

In other applications, such as the application focused on in this paper, the intermediate variable is continuous, resulting in an infinite set of principal strata. To simplify this problem, some have chosen to discretize a continuous intermediate variable (e.g., Sjölander, 2009; Page, 2012). However, discretization may be sensitive to cutoff

point decisions, and aggregation obfuscates heterogeneity (Schwartz, Li, & Mealli, 2011).

To estimate causal effects across an infinite number of strata with finite sample sizes, the principal stratification framework with continuous intermediate variables requires that the researcher impose greater structure to achieve model identification. At most, only one of D's potential outcomes can ever be observed, causing the onerous statistical file-matching problem in which associational parameters describing the relationship between the potential outcomes are not estimable (see, e.g., Little & Rubin, 2002). Accordingly, the estimator must address the file-matching missing data problem directly, and I will show that inferences depend critically on these values. Additionally, dimensionality requires that the researcher specify a functional relationship between the D and Y potential outcomes. Comparisons of the Y potential outcomes may then be modelled by fitting a surface spanning the full joint distribution of the D potential outcomes. A principal causal effect surface (PCES) refers to a surface that models mean differences in the Y potential outcome.

Both frequentist and Bayesian PCES estimators that do not rely on unverifiable assumptions like monotonicity have been proposed in previous literature, but the asymptotic properties of the resulting estimates have not been studied. In the Bayesian framework, Schwartz et al. (2011) incorporate infinite mixture models and data augmentation (Tanner & Wong, 1987) to address the missing data problem. Yet, results

from Bayesian estimators are likely highly sensitive to alternative priors. This is especially true in the file-matching missing data case because the data provide no information on associational parameters for Bayesian updating of the corresponding priors to proceed (see Little & Rubin, 2002, p. 156-161).

Using copulas, Bartolucci & Grilli (2011) proposed a frequentist PCES estimator. However, previous literature conflicts as to whether a copula-based estimator can adequately resolve the file-matching problem. Bartolucci & Grilli (2011) do find that a value exists for the associational parameter that maximizes the profile log-likelihood in a real-world data example. In contrast, others maintain that copulas themselves cannot fix the identifiability issue and suggest that researchers include prior beliefs when fitting copula models (Conlon, Taylor, & Elliott, 2017). Moreover, even if copulas can identify associational parameter estimates in some samples, the consistency properties of the associational parameter estimates have not been studied, even though it is well established that estimating associational parameters under the file-matching problem is known to produce "misleading results" (Little & Rubin, 2002, p. 7).

The main contribution of this paper is to propose a simulation-extrapolation (SIMEX; Devanarayan & Stefanski, 2002) estimator that leverages ignorability and extrapolates to circumvent the file-drawer problem. In the "big data" revolution, researchers increasingly have access to information about pre-treatment covariates to explain heterogeneity, and machine learning algorithms can make highly accurate

predictions. I exploit the fact that the associational parameters—namely residual covariances—are zero in the limiting case that the covariates perfectly predict D's potential outcomes. I use a simulation study to evaluate whether a SIMEX estimator can produce consistent estimates in settings where a set of highly predictive covariates ($R^2 \approx 0.7$) exist to extrapolate parameter estimates in the hypothetical situation that $R^2 = 1$. I compare the estimation properties of the SIMEX estimator to two-stage least squares (TSLS) and conditional mean imputation.

I organize this paper as follows: First, I describe the illustrative example that motivates the adoption of a PCES estimator. Next, I expand on the PS framework and discuss underlying assumptions. Next, I relate PS to instrumental variables and show that TSLS produces biased and inconsistent estimates for principal causal effect surfaces. I then explain why regression to the mean results in biased estimates using standard imputation methods. Subsequently, I propose a simulation-extrapolation method to resolve the problems identified for the TSLS and imputation estimators. Finally, I conduct a simulation to study alternative estimators for estimating a PCES and discuss future directions.

Motivating Example

Without loss of generality, I provide an illustrative example to motivate and explain the PS framework. The data come from a 15-year experimental evaluation

conducted by MDRC of career academies in nine high schools located in cities throughout the United States (Kemple & Willner, 2008). A career academy was established within each high school that focused instruction on a vocational theme and fostered communication between school staff and employers through partnerships. Before enrolling in high school, students were randomly assigned to receive an offer to attend a career academy (treatment) instead of attending a traditional school. Kemple & Willner (2008) found that that the offer caused positive and significant impacts on wages eight years after graduation, even though standard determinants of wages—such as performance on standardized tests or postsecondary attainment—remained unaffected.

Page (2012) explored one mechanism to explain why wages increased with an attendance offer, even though indicators of human capital did not show gains. The curricula for the career academies afforded opportunities for students to become exposed to the work world as measured by a "world-of-work" scale constructed from student survey responses (Page, 2012). Indeed, students attending career academies reported engaging in more discussions about careers with their teachers, counselors, and parents. These students also more frequently engaged in internship and job-shadowing programs. Page (2012) proposed that the greater the effect of the offer on a student's exposure to the work world, the stronger the effect of the offer on future wages. Testing this hypothesis in a manner that is consistent with PS requires investigating treatment effect

heterogeneity on an endogenous intermediate variable, as exposure to the work world is strongly related to treatment assignment.

Background and Notation

I make the stable uniform treatment value assumption (SUTVA; Rubin, 1978) to simplify the analysis, in that I assume that there are no hidden forms of treatment and that a student's potential outcomes are independent of others' potential values. I define Z_i to be a binary treatment assignment indicator for student i taking on the value Z_i = 1 if given an offer to attend a career academy and $Z_i = 0$ if not given an offer. In addition, I assume assignment into treatment is nondeterministic for all students in the study, so that each student has a nonzero probability of being granted an attendance offer *and* a nonzero probability of not being granted an attendance offer (i.e., Regular Assignment Mechanism; Imbens & Rubin, 2015).

I denote a student's score on the world-of-work scale under treatment condition $Z_i = z$ as $D_i(z)$. The SUTVA assumption allows me to denote $D_i(1)$ and $D_i(0)$ as the student's reported world-of-work score if given an offer or not, respectively. The joint distribution of potential outcomes for the intermediate variable are defined as all $(D_i(0), D_i(1))$ pairs in the population. At most, one of the $D_i(z)$ is ever observed for any student, implying that "some of the parameters relating to the association between [$D_i(0)$ and $D_i(1)$] are not estimable from the data" (Little & Rubin, 2002, p. 7). This is called

the file-matching problem, and I later show that PCES estimates obtained from TSLS depend critically on the association between $D_i(0)$ and $D_i(1)$. I denote d_i^{obs} as the D potential outcome that is observed and d_i^{miss} as the potential outcome that is missing.

Alternative methods to TSLS, including standard methods for handling missing data, do not address the file-matching problem. Multiple-imputation and the Expectation-Maximization algorithm (EM; Dempster, Laird, & Rubin, 1977) both require that the associational values be fixed prior to estimation. Bayesian estimators, on the other hand, may include prior "beliefs" about the associational parameter as part of a data augmentation step (Tanner & Wong, 1987), but there is no information in the data to update these beliefs. Consequently, the posterior and the prior distributions on the associational parameters remain unchanged (Little & Rubin, 2002, p. 133-161).

I denote the potential outcomes representing student i's post-graduation wages as $Y_i(1)$ and $Y_i(0)$. Following standard practice in the PS framework, I ignore potential outcomes of the form $Y_i\big(z, D_i(z')\big)$ because I do not consider the world-of-work score to be a quantity that is subject to experimental manipulation. Therefore, I denote the principal causal effect surface (PCES) as the estimand:

$$\text{PCES}_i = \mathbb{E}[\Delta_i | D_i(1), D_i(0)] \tag{1}$$

where $\Delta_i = Y_i(1) - Y_i(0)$ represents the individual causal effect of the randomized offer on *Y*.

Application to the MDRC Experiment

Estimating the PCES is directly relevant to testing the hypothesis that students with the greatest gains in exposure to the work world by attending a career academy also exhibit the greatest gains in wages. To illustrate, consider the hypothetical situation where Δ_i, $D_i(0)$, and $D_i(1)$ potential values were known for everyone. Figures 1 and 2 plots the $D_i(0)$ and $D_i(1)$ values for two hypothetical samples. The size of the marker is proportional to the individual's Δ_i value.

In Figure 1, the markers are all the same size, indicating that there is no heterogeneity in treatment effects on wages. The PCES in this case would therefore be a flat three-dimensional surface. On the other hand, in Figure 2, the marker sizes increase the greater the difference in $D_i(1)$ and $D_i(0)$, indicating that the more positive the individual causal effect on D, the more positive the individual causal effect on Y. This hypothetical situation would be consistent with Page's (2012) hypothesis. Clearly, the PCES in this scenario cannot be modeled by a flat surface for this sample. Instead, the three-dimensional best-fit surface would increase in elevation most along the direction indicated by the arrow. This example shows that heterogeneity in Δ_i can be identified by the shape of the PCES.

<Insert Figure 1 & 2 about here>

Modeling Assumptions

I now describe the required assumptions to estimate parameters that define a principal causal effect surface. Randomization implies that all potential outcomes and all pre-treatment covariates, *X*, are ignorable with respect to treatment assignment. Strong ignorability is formalized as:

<u>Assumption 1:</u> *Strong ignorability*:

$$Z_i \perp (D_i(0), D_i(1), Y_i(0), Y_i(1), X_i)\ \forall i$$

Because of ignorability, *X* and the potential outcomes for *D* and *Y* are balanced across treatment arms. This means that all moments—means, variances, covariance, skew, etc.—in the distribution of the covariates are equal across treatment arms in expectation. Balance in the means of *X* is critical for unbiased predictions of missing *D* values (see Imbens & Rubin, 2015, p. 272).

I further assume that causal effect heterogeneity exists for the intermediate variable: the causal effect of treatment on the intermediate variable is not constant across all individuals:

<u>Assumption 2</u>**:** *Heterogeneity in causal effects on D:*

$$D_i(1) - D_i(0) = \Gamma_i,\ \mathbb{V}[\Gamma_i] \neq 0$$

where Γ_i represents the individual causal effect on *D.* Assumption 2 is assured if the distributions of potential outcomes cannot be reproduced by a simple mean-shifted

translation. Therefore, any differences in variances, skew, or higher order moments in the observed D values across treatment arms is indicative that heterogeneity is present.[1]

For the PCES estimators evaluated in this paper, I assume that the researcher has access to observed predictors that explain the heterogeneity:

<u>Assumption 3:</u> *Effectiveness of exogenous predictors:*

$$\Pr(\Gamma_i|X_i) \neq \Pr(\Gamma_i) \ \forall i$$

In words, Assumption 3 states that the individual causal effect on world-of-work scores is related to the observed pre-treatment covariates. Thus, the conditional mean in causal effects on D is some nontrivial function of the observed covariates. I denote the individual causal effect of treatment on Y as Δ_i, so that $\Delta_i = Y_i(1) - Y_i(0)$. I assume throughout this paper that there are no other post-treatment intermediate variables that are simultaneously correlated with both Γ_i and Δ_i. Such variables would be confounders, denoted C_i, and the assumption of no confounders is formalized as:

<u>Assumption 4:</u> *Unconfoundedness:*

$$\Pr(\Delta_i, \Gamma_i|X_i, C_i) = \Pr(\Delta_i, \Gamma_i|X_i) \ \forall i$$

Assumption 4 is directly akin to the instrumental variable assumption that the instrument Z_i is independent of D and Y residuals (Murnane & Willett, 2011; Morgan &

[1]Ding, Feller, and Miratrix (2016) discuss non-parametric statistical tests for assessing causal effect heterogeneity.

Winship, 2015). While a PCES estimator can be built to account for confounders by modeling the potential outcomes of these variables, that is beyond the scope of this paper.

Next, I assume that the pre-treatment predictors influence wages only indirectly through $D_i(1)$ and $D_i(0)$:

<u>Assumption 5</u>**:** *No direct effect of X on* Δ_i:

$$\Pr(\Delta_i | D_i(0), D_i(1), X_i) = \Pr\big(\Delta_i \big| D_i(0), D_i(1)\big) \ \forall i$$

On its face, Assumption 5 resembles the exclusion restriction assumption in instrumental variables, but this comparison requires caution. In instrumental variables, the instrument—often random assignment to treatment—is said to only induce a causal effect through its effect on the intermediate variable. Consequently, the exclusion restriction would imply that the causal effect $\Delta_i = 0$ for all units whose Γ_i value is zero. This, however, is not what is stated by Assumption 5. Indeed, there may be a nonzero causal effect on Y, even for units whose intermediate value is unchanged by treatment assignment.

An important consequence of Assumption 2 and Assumption 5 is that heterogeneity in causal effects on D translates directly to heterogeneity in causal effects on Y. Thus, a non-trivial PCES requires variance in individual causal effects on Y, i.e. $\mathbb{V}(\Delta_i) \neq 0$. The researcher can test for the existence of treatment effect heterogeneity on

this variable by looking for differences in variances, skew, or higher order moments in the observed Y values across treatment arms.

Finally, dimensionality requires the specification of a functional relationship relating the D potential outcomes to the Y potential outcomes. I make the rather restrictive functional assumption that the Y potential outcomes are a function of $D_i(0)$ and $D_i(1)$ main effects and a $D_i(0)$-by-$D_i(1)$ interaction. This necessarily implies that y_i^{obs} is a linear function of d_i^{obs} in both treatment arms. Denoting the ith observation in a design matrix as $\boldsymbol{v}_i = [1 \;\; D_i(0) \;\; D_i(1) \;\; D_i(0) \times D_i(1)]$, the linearity assumption is formalized as:

<u>Assumption 6:</u> *Linearity:*

$$Y_i(z) = v_i'\beta_z + \epsilon_{z,i}, \qquad \mathbb{E}\left[\epsilon_{z,i}\right] = 0$$

Under Assumption 6, a model for the principal causal effect surface is then given by the linear function:

$$\text{PCES}_i(\boldsymbol{\theta}) = \boldsymbol{v}_i'\boldsymbol{\theta} \tag{2}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$.

Relationship to Instrumental Variables

Reardon and Raudenbush (2013) show that the IV estimand can be interpreted as a compliance-weighted average treatment effect (CWATE) when the intermediate variable is continuous:

$$\text{CWATE}_i = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1] \tag{3}$$

While others have demonstrated that PS is a generalization of the instrumental variables (IV) framework when the intermediate variable is discrete (Frangakis & Rubin, 2002), I find no literature that generalizes IV to PS when the intermediate variable is continuous. In this section, I show that CWATE_i is simply a constrained principal causal effect surface and that the Wald IV estimator follows directly from the PS framework.

In addition to SUTVA, IV estimators assume that: (a) a linear relationship exists between differences in the D potential outcomes and the Y potential outcomes, (b) for all units, differences in potential outcomes in Y result only from differences in potential outcomes in D (i.e. *person-specific exclusion restrictions*), and (c) no covariance between person-specific effects on D and person-specific effects on Y (i.e. $\mathbb{C}\text{ov}(\Gamma_i, \Delta_i) = 0$) (Reardon & Raudenbush, 2013). Note that linearity with respect to the *potential outcomes* rules out the possibility that the PCES is a function of a $D_i(0)$-by-$D_i(1)$ interaction or higher order terms. Thus, the linearity assumption implies that the principal causal effect surface is a plane in $\mathbb{R}^3$, i.e.

$$Y_i(1) - Y_i(0) = \theta_0 + \theta_1 D_i(0) + \theta_2 D_i(1) + \epsilon_i \tag{4}$$

where $\mathbb{E}[\epsilon_i] = 0$.

In addition, assume that $\theta_2 = -\theta_1$ is true for the PCES in the population. Plugging into Equation 5 and rearranging implies that the PCES can be written as

$$Y_i(1) - Y_i(0) = \theta_0 + \theta_2(D_i(1) - D_i(0)) + \epsilon_i \tag{6}$$

The exclusion restrictions require that the plane goes through the origin, implying that $\theta_0 = 0$. Thus, the PCES can be rewritten as:

$$Y_i(1) - Y_i(0) = \theta_2(D_i(1) - D_i(0)) + \epsilon_i \tag{7}$$

Note that θ_2 represents the expected difference in the potential outcomes for a unit difference between $D_i(1) - D_i(0)$. This is equivalent to the CWATE estimand given in Equation (7).

To arrive at the Wald estimator, first take the expectation of Equation 7 and solve for θ_2.

$$\theta_2 = \frac{\mathbb{E}[Y_i(1) - Y_i(0)]}{\mathbb{E}[D_i(1) - D_i(0)]} = \frac{\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]}{\mathbb{E}[D_i(1)] - \mathbb{E}[D_i(0)]} \tag{8}$$

Substitute the naive estimates for the $\mathbb{E}[Y_i(z)]$ and $\mathbb{E}[D_i(z)]$ term

$$\hat{\theta}_2 = \frac{\bar{y}^{obs}_{\{i:\, Z_i=1\}} - \bar{y}^{obs}_{\{i:\, Z_i=0\}}}{\bar{d}^{obs}_{\{i:\, Z_i=1\}} - \bar{d}^{obs}_{\{i:\, Z_i=0\}}} \tag{9}$$

where $\bar{y}^{obs}_{\{i:\, Z_i=z\}}$ and $\bar{d}^{obs}_{\{i:\, Z_i=z\}}$ are the sample means of y_i^{obs} and d_i^{obs} in the $Z_i = z$ treatment arm. Equation 9 is equivalent to the Wald estimator used ubiquitously in the IV framework.

In summary, the Wald estimator can be derived by assuming that the PCES is a plane constrained to pass through the origin. The PCES estimand is defined more generally: it is a surface of average causal effects across the joint distribution of the intermediate's potential outcomes.

Bias in TSLS Estimates

In the previous section, I considered only the case where the researcher had access to one instrument (Z_i), and I discussed the assumptions needed for traditional IV estimators to produce unbiased estimates for a highly constrained PCES surface. I now consider the possibility that the researcher has access to pre-treatment covariates that explain heterogeneity in Γ_i, but these covariates have no direct effect on Δ_i. I show that the TSLS estimates of the PCES are biased in general and that the bias is proportional to the unobserved associational parameter that defines the file-matching problem.

With pre-treatment covariates, a TSLS estimate for a PCES can be obtained by conducting separate multivariate regressions of d_i^{obs} and X_i for both treatment arms to obtain $\widehat{D}_i(0)$ and $\widehat{D}_i(1)$ predictions for all individuals. Next, the predicted values are substituted for $D_i(0)$ and $D_i(1)$ in the design matrices when estimating the $\boldsymbol{\beta}$ parameters using the observed outcomes in each treatment arm.

Proposition: If (a) Assumptions 1-6 hold, (b) a nonzero covariance exists between the $D_i(0)$ and $D_i(1)$ residuals (i.e., $\mathbb{C}\text{ov}(D_i(0), D_i(1)|X_i) \neq 0$), and (c) the PCES is a function of an interaction between the *D* potential outcomes (i.e., $\Pr(\Delta_i|D_i(0), D_i(1), D_i(0) \times D_i(1)) \neq \Pr(\Delta_i|D_i(0), D_i(1))$, then the PCES is biased by a constant value across the joint distribution of *D* potential outcomes.

Proof: Under Assumptions 1-6, the joint distribution of all potential outcomes is generated from the following data generating mechanism:

$$D_i(0) = X_i\boldsymbol{\alpha}_0 + \epsilon_{0,i} \tag{10}$$

$$D_i(1) = X_i\boldsymbol{\alpha}_1 + \epsilon_{1,i} \tag{11}$$

$$Y_i(0) = \beta_{0,0} + \beta_{0,1}D_i(0) + \beta_{0,2}D_i(1) + \beta_{0,3}(D_i(0) \times D_i(1)) + \epsilon_{2,i} \tag{12}$$

$$Y_i(1) = \beta_{1,0} + \beta_{1,1}D_i(0) + \beta_{1,2}D_i(1) + \beta_{1,3}(D_i(0) \times D_i(1)) + \epsilon_{3,i} \tag{13}$$

such $\mathbb{C}\text{ov}(\epsilon_{0,i}, \epsilon_{1,i}) = \sigma_{0,1}$. In general, one would expect that $\sigma_{0,1} \neq 0$. A TSLS estimator would proceed by obtaining the predicted values $\widehat{D}_i(z) = X_i\boldsymbol{\alpha}_z$ in the first stage and substituting these predicted values in the second stage. Noting that $D_i(z) = \widehat{D}_i(z) + \epsilon_{z,i}$, the $Y_i(z)$ potential outcome in the second stage can be rewritten as:

$$\begin{aligned} Y_i(z) &= \beta_{z,0} + \beta_{z,1}\left(\widehat{D}_i(0) + \epsilon_{0,i}\right) + \beta_{z,2}\left(\widehat{D}_i(1) + \epsilon_{1,i}\right) + \beta_{z,3}\left\{\left(\widehat{D}_i(0) + \epsilon_{0,i}\right) \times \left(\widehat{D}_i(1) + \epsilon_{1,i}\right)\right\} + \epsilon_{z+2,i} \\ &= \beta_{z,0} + \beta_{z,1}\widehat{D}_i(0) + \beta_{z,2}\widehat{D}_i(1) + \beta_{z,3}\widehat{D}_i(0) \times \widehat{D}_i(1) + \epsilon^*_{z+2,i} \end{aligned} \tag{14}$$

where $\epsilon^*_{z+2,i} = \epsilon_{z+2,i} + \left(\beta_{z,1} + \beta_{z,3}\widehat{D}_i(1)\right)\epsilon_{0,i} + \left(\beta_{z,2} + \beta_{z,3}\widehat{D}_i(0)\right)\epsilon_{1,i} + \beta_{z,3}\left(\epsilon_{0,i} \times \epsilon_{1,i}\right)$. OLS is also used as the estimator for the β coefficients in the second

stage. The corresponding estimand for the OLS estimator is then given by $\mathbb{E}[Y_i(z)|\widehat{D}_i(0), \widehat{D}_i(1)]$. Taking the conditional expectation of Equation 14 and noting that randomization guarantees that the estimates for the predicted values are unbiased,

$$\mathbb{E}\left[Y_i(z)\middle|\widehat{D}_i(0), \widehat{D}_i(1),\right] = \beta_{z,0} + \beta_{z,1}\widehat{D}_i(0) + \beta_{z,2}\widehat{D}_i(1) + \beta_{z,3}\widehat{D}_i(0) \times \widehat{D}_i(1) + \mathbb{E}\left[\epsilon^*_{z+2,i}\middle|\widehat{D}_i(0), \widehat{D}_i(1)\right] \quad (15)$$

Note that $\mathbb{E}\left[\epsilon^*_{z+2,i}\middle|\widehat{D}_i(0), \widehat{D}_i(1),\right] = \beta_{z,3}\mathbb{E}\left[\epsilon_{0,i} \times \epsilon_{1,i}\right] = \beta_{z,3}\sigma_{0,1}$. This implies that the multiplication of the residuals results in a constant term, provided that $\beta_{z,3}$ and $\sigma_{0,1}$ are not zero. Thus, $\mathbb{E}\left[Y_i(z)\middle|\widehat{D}_i(0), \widehat{D}_i(1)\right] =$

$$\beta^*_{z,0} + \beta_{z,1}\widehat{D}_i(0) + \beta_{z,2}\widehat{D}_i(1) + \beta_{z,3}\left(\widehat{D}_i(0) \times \widehat{D}_i(1)\right). \quad (16)$$

where $\beta^*_{z,0} = \beta_{z,0} + \beta_{z,3}\sigma_{0,1}$. Consistent estimates for $\beta^*_{z,0}$, $\beta_{z,1}$, $\beta_{z,2}$, and $\beta_{z,3}$ in the second stage require that $\mathbb{C}\mathrm{ov}\left(\widehat{D}_i(0), \epsilon^*_{z+2,i}\right) = \mathbb{C}\mathrm{ov}\left(\widehat{D}_i(1), \epsilon^*_{z+2,i}\right) = \mathbb{C}\mathrm{ov}\left(\widehat{D}_i(0) \times \widehat{D}_i(1), \epsilon^*_{z+2,i}\right) = 0$. Because the predicted values are functions of X, it can be shown that this is only guaranteed if there are no direct effects of X on Y (i.e. Assumption 5). Thus, TSLS provides consistent estimates for all β terms, except for the coefficient of the intercept. In large samples, TSLS estimates will tend towards $\beta_{z,0} + \beta_{z,3}\sigma_{0,1}$, rather than $\beta_{z,0}$.

The bias in the intercept coefficients for the Y models translates directly to bias in the coefficients for the PCES estimated with TSLS. By applying properties of

expectation, it is straight forward to show that in large samples, TSLS estimates tend towards $\theta_3\sigma_{0,1}$ over repeated sampling.

In summary, a TSLS approach produces biased and inconsistent PCES intercept estimates if there exists a $D_i(0) \times D_i(1)$ interaction. The magnitude of this bias for the PCES estimator is proportional to the unobserved residual covariance between the D potential outcomes. It should be noted that I have only considered the situation where the PCES is known to have only linear main effects of $D_i(0)$ and $D_i(1)$ and an interaction term. It can also be shown that the bias is *not* dependent of the unobserved associational parameter for select terms if the PCES can be well approximated by a polynomial function with interaction terms and if the residuals are multivariate normal.

Conditional Mean Imputation: Local Bias and a Correction

As discussed in the previous section, the TSLS point estimates are highly sensitive to the associational parameter, $\sigma_{0,1}$. Here, I consider an alternative estimator that slightly modifies TSLS. In the second stage, TSLS involves substituting the predicted values for both $D_i(0)$ and $D_i(1)$, even though one of these potential outcomes is observed directly in the data. An alternative is to not substitute the predicted values for the observed value, but instead to use the observed value itself. Such a modification also more clearly treats the estimation problem as a missing data problem because it can be

understood as imputing the missing D potential outcome with the predicted value. Specifically, this type of imputation procedure is named conditional mean imputation because the missing D potential outcome is to be imputed with an estimate of the expected D value given a unit's covariate values, X_i (see Little & Rubin, 2002, p. 62-66). One advantage of conducting imputation is that it allows the researcher to work directly with complete data in a transparent and intuitive manner.

Despite being biased in its raw form due to omitted variable bias, I propose that conditional mean imputation can be modified in a manner that exploits ignorability to transform the problem into the more tractable terrain of estimating in the presence of covariates that are measured with error. The key insight is that regression to the mean results in imputations that are biased for all d_i^{miss} except for exactly at the population mean value. Therefore, the conditional mean imputation estimate is locally biased. Nevertheless, ignorability allows the researcher to correct the local bias using the observed data drawn from the opposite treatment arm. The correction ensures that the imputed values are unbiased at all d_i^{miss} values. If the variation in differences of the corrected imputation and the true missing value is assumed to be completely random, then the imputed value is a fallible indicator of d_i^{miss}. The resulting attenuation bias can be solved using existing estimators that address this problem.

Conditional Mean Estimator and Omitted Variable Bias

Without loss of generality, consider the population equation for the $Y_i(0)$ potential outcome to understand why the conditional mean imputation estimator is biased:

$$Y_i(0) = \beta_{0,0} + \beta_{0,1}D_i(0) + \beta_{0,2}D_i(1) + \beta_{0,3}(D_i(0) \times D_i(1)) + \epsilon_{3,i} \quad (17)$$

Noting that $D_i(1) = \widehat{D}_i(1) + \epsilon_{1,z}$, the above equation can be rewritten as

$$Y_i(0) = \beta_{0,0} + \beta_{0,1}D_i(0) + \beta_{0,2}\widehat{D}_i(1) + \beta_{0,2}\epsilon_{1,i} + \beta_{0,3}\left(D_i(0) \times \widehat{D}_i(1)\right) + \beta_{0,3}\left(D_i(0) \times \epsilon_{1,i}\right) + \epsilon_{3,i} \quad (18)$$

The conditional mean estimator proceeds by excluding any term with $\epsilon_{1,i}$ during estimation. Thus, $\epsilon_{1,i}$ and $D_i(0) \times \epsilon_{1,i}$ are omitted variables. Bias in the $\boldsymbol{\beta}_z$ estimate will only result from omitting these variables under two conditions: (a) at least one of the omitted terms is not independent of $Y_i(0)$, and (b) at least one of the omitted terms is associated with any of the observed $D_i(0)$, $\widehat{D}_i(0)$, or $D_i(0) \times \widehat{D}_i(1)$ variables.

Both necessary conditions for omitted variable bias are satisfied. In any practical application of a PS estimator, the researcher would hypothesize a non-zero population value for θ_2, θ_3, or both. This implies that at least one of the corresponding omitted terms is associated with at least one of the Y potential outcomes. Moreover, if $\sigma_{0,1} \neq 0$, then there exists a non-zero correlation between $D_i(0)$; also, the unobserved $D_i(0) \times \epsilon_{1,i}$ interaction term is clearly not independent of the observed main effect $D_i(0)$. This all suggests that conditional mean imputation is subject to omitted variable bias because the

omitted variables confound the relationship between the missing D potential outcome and the Y potential outcomes.

Local Bias in Conditional Mean Imputations

Although estimates from conditional mean imputation are susceptible to omitted variable bias, I propose that randomization makes it possible for the researcher to transform this problem into a more tractable latent variable problem. Suppose, for example, that one could obtain an imputed value that was only susceptible to the measurement error that defines classical test theory. In this hypothetical, let the imputed value be the "observed score" and the missing value represent the "true score" so that

$$d_i^{imp} = d_i^{miss} + u_i,\ \mathbb{E}[u_i] = 0\ \&\ \mathbb{V}\mathrm{ar}[u_i] = \mathrm{SEM}^2 \tag{19}$$

where d_i^{imp} is the imputed value, and the standard error of measurement (SEM) is a known and constant value (i.e., homoscedasticity). I note that the assumption that $\mathbb{E}[u_i]$ implies that the observed imputation is an unbiased estimate of the true missing value for all units, i.e. $\mathbb{E}\left[d_i^{imp} \middle| d_i^{miss}\right] = d_i^{miss}$. Supposing that such an imputed value could be obtained, then substituting that value for the missing D potential outcome when estimating the $\boldsymbol{\beta}$ coefficients brings us to the more tractable goal of disattenuating the bias induced by the measurement error.

A fundamental problem with using the conditional means when imputing is that regression to the mean results in unbiased estimates of d_i^{miss} only for units exactly at the population mean. As I now explain, $\mathbb{E}\left[d_i^{imp} \middle| d_i^{miss}\right] \neq d_i^{miss}$ for all other units. I show, however, that the conditional mean imputations can be easily transformed so that the corrected values are unconditionally unbiased and homoscedastic.

To understand how regression to the mean leads to local bias, consider Panel A in Figure 3. The bold, grey line represents the condition that the imputed value (y-axis) is exactly equal to the true missing value (x-axis). The origin of the Cartesian coordinates is centered at the population means. Because the pre-treatment covariates only explain about 70% of the variance in the missing D potential outcome, there exists random variation away from the grey line. Some of the imputed values are greater than the true missing value, while other imputed values are smaller than the true mean value. This variation is illustrated by contours that represent density in the joint distribution of the missing and imputed values. The thin blue line represents the *average* imputed value for a given missing value, i.e. $\mathbb{E}\left[d_i^{imp} | d_i^{miss}\right]$. One can see that the imputed values equal the true value on average only at the origin. For missing values greater than the mean, the imputed values are *on average* less than the true missing value. The opposite is true for missing values less than the mean. Consequently, random variation leads to *local* bias for any given d_i^{miss} value away from the origin, with the direction of this bias towards the

population mean. Note that it is straightforward to show that this local bias persists even if more advanced *univariate* imputation procedures are used, including rigorous multiple imputation procedures that approximate a posterior predictive distribution.

Correcting Local Bias in Conditional Mean Imputations

Randomization allows the researcher to estimate the expected local bias at each d_i^{miss} value by dividing the imputed values by the R^2 obtained from regressing d_i^{miss} on X_i. Although a direct estimate of this value cannot be obtained from the observed data, randomization guarantees that an unbiased estimate of this quantity can be calculated by regressing d_i^{obs} on X_i in the opposite treatment arm, denoted $\hat{R}^2_{D_i(z')}$. I denote the corrected imputations as $d_i^{imp^*}$. The effect of the correction on the imputed values is shown visually in Panel B of Figure 3; after applying the correction, the imputations equal the true missing value on average. Furthermore, it can be shown that at each d_i^{miss} value, the corrected imputation values are dispersed by an estimated value

$$\mathbb{V}\hat{\mathrm{a}}\mathrm{r}\left(d_i^{imp^*} | d_i^{miss}\right) = \frac{\mathbb{V}\hat{\mathrm{a}}\mathrm{r}\left(d_i^{imp}\right)}{\left(\hat{R}^2_{D_i(z')}\right)^2} - s^2_{D(z')} \tag{20}$$

where $s^2_{D(z')}$ is the estimated variance of the observed D values in the alternative treatment arm and d_i^{imp} is the conditional mean imputation. If deviations from the

missing value are independent sources of error, then Equation 20 represents the square of the standard error of measurement.

Randomization also implies that local bias can be attenuated using this correction procedure even if important pre-treatment predictors of Γ_i are omitted. Certainly, omitting predictors will necessarily reduce the amount of variance in d_i^{miss}. Nevertheless, ignorability implies that $\hat{R}^2_{D_i(z')}$ continues to remain an unbiased estimate of the R^2 estimate for d_i^{miss} on X_i. This is illustrated in Figure 4.

<Insert Figures 3 & 4 about here>

Despite relieving the local bias associated with regression to the mean, the random variation of the imputations away from the true d_i^{miss} complicate estimation using the corrected imputation values. Naively estimating the $\boldsymbol{\beta}$ parameters that define the PCES by substituting the conditional mean imputation value for d_i^{miss} assumes that the imputation values are perfect measures of the missing value—a condition that would only be true if X_i perfectly predicted d_i^{miss}. Otherwise, the corrected imputation value is a fallible indicator of d_i^{miss}, and the resulting $\boldsymbol{\beta}$ parameters would then be subject to the attenuation bias.

In summary, conditional mean imputation does not lead to "apples-to-apples" comparisons with respect to the joint distribution of the D potential outcomes because regression to the mean leads to local bias. However, a correction factor can be estimated using the observed data and applied to generate imputations that are, at best, fallible indicators of d_i^{miss} measured with error. Even so, a PCES estimator that naively substitutes the corrected imputation values is subject to attenuation bias provided that the deviations from the true missing value are independent sources of error.

SIMEX

General Background

It is well established that measurement error in the predictors biases OLS coefficient estimates. I have proposed a modification to conditional mean imputation that transforms the missing data problem into the more manageable territory of estimating regression coefficients when the covariates are measured with error. I assume that these departures share the essential qualities of simple measurement error to make use of known estimators that ameliorate the resulting attenuation bias, namely independence. Popular estimators in the presence of measurement error include: (a) calculating regression coefficients by first modifying the observed correlation matrix, (b) latent variable models, and (c) SIMEX (Cook & Stefanski, 1994). I chose SIMEX because manually deattenuating a correlation matrix in the presence of interactions and higher-

order terms is nontrivial. Similarly, it is far more computationally scalable to include interaction and higher-order terms in SIMEX than in traditional latent variable models. Additionally, SIMEX does not rely on parametric assumptions regarding the underlying latent variable, and the approach is highly intuitive for researchers with experience conducting sensitivity analyses. On the other hand, the standard error of measurement (SEM) must be known, and SIMEX is generally only effective when the measurement error is homoscedastic (Devanarayan & Stefanski, 2002; Lockwood & McCaffrey, 2017).

SIMEX involves a simulation phase and an extrapolation phase, and the results of the procedure are illustrated heuristically in Figure 5. First, the researcher identifies a grid of $l = 1, \dots, L$ non-negative λ values to conduct independent simulations. λ is scaled so that it represents the amount of measurement error added in SEM units. For example, $\lambda = 0$ represents the case where no simulated measurement error is added, and $\lambda = -1$ corresponds to the hypothetical situation that the "true scores" of the covariates were observed. The procedure requires simulating measurement error $b = 1, \dots, B$ times and recording coefficient estimates at each λ value. These parameter estimates are denoted as $\hat{\beta}_b(\lambda)$.

Next, the researcher fits a best-fit function relating the $\hat{\beta}_b(\lambda)$ estimates to λ. I denote the true best-fit line by $g(\lambda)$. Departures from the predicted $g(\lambda)$ values are assumed to be the result of sampling error when simulating measurement to generate

individual $\hat{\beta}_b(\lambda)$ estimates. If $g(\lambda)$ is known, then the researcher can obtain the SIMEX estimate by extrapolating away from the observed data and predicting $g(-1)$. This extrapolation process is shown visually by the dotted line and circular marker in Figure 5.

Extrapolation is the "Achilles' heel" of the SIMEX estimator. Estimates will only be unbiased estimates so long as the $g(\lambda)$ is correctly specified. Moreover, predictions resulting from extrapolation can be highly inefficient estimates, even if the true form of $g(\lambda)$ is known. This is especially true when the standard error of measurement is large relative to the variance in the true scores, resulting in extrapolation far from the observed data.

<Insert Figure 5 about here>

Application to Estimating a PCES

As applied to the problem of estimating a PCES, $\lambda = 0$ represents the coefficient estimates obtained by substituting the corrected conditional mean imputations for the missing D potential outcomes. The key insight is that $\lambda = -1$ correponds to the hypothetical situation where the imputed value *exactly* equals d_i^{miss}. Thus, a SIMEX estimator attempts to identify the parameter estimates that would be estimated in the hypothetical situation that the missing D potential outcome had been observed for all

units. Moreover, the only way one could possibly observe the missing D potential outcome is if the researcher had access to a rich set of predictors that perfectly predict the missing value so that there was no unexplained variance in d_i^{miss}. Under such a condition, the associational parameter that defines the file-matching problem becomes irrelevant.[2] In this way, the proposed SIMEX estimator bypasses the file-matching problem completely.

Simulation Study Design

To make the simulation more realistic, I designed the population parameters for the simulation study around the data from the MDRC experiment. In this section, I discuss the measures in the empirical sample used to assess Page's (2012) hypothesis: students who experienced the greatest boost in world-of-work scores also experienced the strongest treatment effects on wages. Next, I describe the specifics of the simulation procedure.

[2] If one of the D potential outcomes is perfectly predicted by the covariates, then the corresponding mean-squared error is zero, i.e. $\sigma_{0,0}^2 = 0$ or $\sigma_{1,1}^2 = 0$. Because $\sigma_{0,1} = \sqrt{\sigma_{0,0}^2 \sigma_{1,1}^2}\mathbb{C}\mathrm{orr}(D_i(0)|X_i, D_i(1)|X_i)$, it follows that $\sigma_{0,1} = 0$.

Measures

The sample includes N = 403 male participants across seven different school sites. Within this sample, 221 students were offered assignment to the career academy (treatment). At the end of high school, all students in the sample responded to a survey asking about their level of engagement in school-sponsored labor-market activities. For each student, a score was calculated from these survey items using the first component of a principal components analysis. I verified that individuals with higher scores tended to have greater levels of labor-market exposure. Moreover, treatment assignment offer had a strong, positive effect on the world-of-work exposure scores ($I\hat{T}T$ = 1.3 pts, Cohen's d = 0.65 sd, $p < 0.001$).

The data also provides suggestive evidence in support of the assumption that heterogeneity in treatment effects for world-of-work scores exists. The distributions of world-of-work scores across treatment conditions is shown in Figure 6, and QQ-plots are shown in Figure 7. The world-of-work scores for the control group exhibit right skew, while the scores for the treatment group display more symmetry. Thus, $D_i(0)$ and $D_i(1)$ differ regarding higher-order moments beyond simply the mean. This provides suggestive evidence that an offer to attend a career academy had a larger effect on subsequent exposure to the labor market for some students than for others (Ding, Feller, & Miratrix, 2016).

<Insert Figures 6 & 7 about here>

To identify predictors that explain heterogeneity on world-of-work scores, I applied a commonly used machine learning algorithm. Specifically, I first identified 44 pre-treatment covariates that showed rank-ordered pairwise correlations greater than 0.2, with the world-of-work scores in either the treatment arm or the control arm. Missingness on the covariates ran between 0-75%. For the purposes of identifying population values for the simulation parameters, I created one imputed data set using the R package missForest (Stekhoven & Buehlmann, 2011). I then assessed balance on all 44 pre-treatment covariates by comparing the means between treatment and control units for continuous covariates or conducting χ^2 tests of independence for ordinal and nominal covariates. I found that only mother's education displayed evidence of imbalance (χ^2 = 10.0, $df = 4$, $p = 0.03$). Although this result may be explained by Type-I error, mother's education is substantively important in these settings, so it is considered a confounder requiring adjustment. Inverse propensity score weighting (IPTW) was applied when fitting all statistical models to ensure balance on this variable within the sample.

Six predictors were formed by developing an ensemble of predictions.[3] Ensemble forecasting involves two stages and can be thought of as conducting a "poll of polls." In

[3] I refer the reader to Hastie, Efron, and Friedman (2009) for more information on the theory of ensemble modeling.

the first stage, the researcher collects predictions estimated from competing algorithms. I chose RandomForests, ridge regression, and bagged CART as competing estimators to form three alternative predictions of $D_i(0)$ and $D_i(1)$ for each individual. In the second stage, the researcher creates final predictions by using the first-stage predictions as the covariates. I used leave-one-out cross validation to assess out-of-sample prediction error. The ensemble method formed strong predictors for both the world-of-work scores in both treatment arms (cv-RMSE$_{D(0)}$ = 0.88, cv-$R^2_{D(0)}$ =0.76, cv-RMSE$_{D(1)}$ = 1.1, cv-$R^2_{D(1)}$ =0.70).

MDRC collected information on average monthly wages four to eight years after high school graduation. Consistent with previous studies that analyzed this data, I find that an offer to a career academy had a moderate, positive effect on post-graduation wages ($I\hat{T}T$= \$186 USD, Cohen's d = 0.18 sd, p = 0.06). To ensure compatibility with Page (2012), I applied a square-root transformation before modeling wages as the outcome of interest.

If there are no direct effects of the covariates on Δ_i (Assumption 5), the existence of a non-trivial PCES surface requires the existence of heterogeneous treatment effects on Y. The distributions of the observed scaled wage values are shown in Figure 8, a QQ-plot is presented in Figure 9 and the descriptive statistics are presented in Table 1. The difference in skew (Control: Est = -0.31; Treated: Est = 0.91) and kertosis (Control: Est =

-0.42; Treated: Est = 3.69) values across treatment arms suggest that the requisite treatment effect heterogeneity may be present.

<Insert Table 1 about here>

<Insert Figures 8 & 9 about here>

Simulation Setup

The data generating mechanism for the simulation is the same as that identified in Equations 10-13. Table 2 reports the population values for $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\theta}$, σ_0^2, σ_1^2, σ_2^2, and σ_3^2 used in the simulation. I now discuss how I determined these population values.

I used the principal causal effect estimates reported in Page (2012) to identify population values for the PCES. Page (2012) first discretized the intermediate variable by three equal quantiles across treatment arms. This implied that $D_i(z)$ could only take on three ordinal values: low, medium, and high. Latent classes are then defined by the joint distribution of potential outcomes for *D*. By imposing monotonicity for all individuals ($D_i(1) > D_i(0)\ \forall i$), six latent classes are possible. Specifically, $(D_i(0), D_i(1)) \in$ {(low, low), (low, medium), (low, high), (medium, medium), (medium, high), (high, high)}.

Table 3 reports the principal causal effect estimates estimated by Page (2012) for each latent class. To generate a continuous surface, these estimates must correspond to

some coordinate pair in $\mathbb{R}^2$. I identified each element in the coordinate pair by taking the midpoint between the largest and smallest observed $D_i(z)$ value for each class. The coordinate pairs corresponding to each latent class are also presented in Table 3. Next, I identified population $\boldsymbol{\theta}$ values by fitting multiple regression models to these data and weighting by the precision of the estimates. I assumed that a student's wage in the control condition only depends on $D_i(0)$. By fitting a simple linear regression model for this treatment arm with y_i^{obs} as the outcome and d_i^{obs} as the control, I determined $\boldsymbol{\beta}_0$. Finally, $\boldsymbol{\beta}_1$values were identified using the property $\boldsymbol{\beta}_1 = \boldsymbol{\theta} + \boldsymbol{\beta}_0$.

I now discuss how I identified $\boldsymbol{\alpha}_z$, σ_0^2,and σ_1^2 values for the simulation. Because each replication requires sampling new X_i values from a known multivariate distribution, I applied matrix algebra properties to the six ensemble predictors to form a new basis. Specifically, I first conducted a principal components analysis to generate an orthogonal basis for the predictors in $\mathbb{R}^6$, and I calculated all principal scores. After standardizing, I applied a linear transformation to the principal scores using a Cholesky decomposition so that the transformed predictors were defined to have pairwise correlations of r =0.25. I then identified $\boldsymbol{\alpha}_z$ values by fitting multiple regressions using the six transformed predictors to the d_i^{obs}, separately across treatment arms. The mean squared error estimates from the multiple regression models were also taken as the population values σ_0^2 and σ_1^2.

<Insert Tables 2 & 3 about here>

Sampling Scheme

Given a specified sample size, *N*, I first generated correlated covariates by sampling X_i from a multivariate normal distribution, i.e.

$$X_i \sim \mathrm{MVN}_6(\mathbf{0}, \Sigma_{\boldsymbol{x}}) \tag{21}$$

such that diagonals of $\Sigma_{\boldsymbol{x}}$ are all equal to 1 and the off-diagonal elements all equal 0.25. With the covariates known, I then sampled complete data following from the data generating mechanism given by Equations 10-13 and the population values in Table 4.

For all simulations, the correct functional form for the $Y_i(z)$ was assumed to be known. I then randomly assigned half of the simulated units to treatment ($z_i = 1$) and half to control ($z_i = 0$) and calculated the observed data used in the simulation from $d_i^{obs} = z_i \times D_i(1) + (1 - z_i) \times D_i(0)$ and $y_i^{obs} = z_i \times Y_i(1) + (1 - z_i) \times Y_i(1)$.

Simulation 1: Evaluating TSLS & Conditional Mean Imputation Estimators

In the first simulation, I verified that TSLS and conditional mean imputation lead to biased and inconsistent estimates of a PCES across a variety of conditions. These conditions are reported in Table 2, and I also describe them here.

I used probability theory to previously show that TSLS will lead to inconsistent estimates of the intercept that will tend towards $\beta_{z,0} + \beta_{z,3}\sigma_{0,1}$, rather than $\beta_{z,0}$ in large samples. Thus, the asymptotic bias associated with a TSLS PCES estimator should only depend on $\theta_{z,3}$ and $\sigma_{0,1}$, and the bias should only affect the intercept value. On the other hand, I showed previously that all $\boldsymbol{\theta}$ estimates will be biased due to omitted variable bias. To verify this, I conducted simulations across four different specifications for Σ. In Condition A, all four residuals—$\epsilon_{0,i}$, $\epsilon_{1,i}$, $\epsilon_{2,i}$, and $\epsilon_{3,i}$—were assumed to be independent of one another, and I expect that under large samples, the intercept coefficient will be unbiased in this condition. In Condition B, I allow $\epsilon_{0,i}$ and $\epsilon_{1,i}$ to freely correlate ($\rho_{0,1} = 0.5$), and I also allow $\epsilon_{2,i}$ and $\epsilon_{3,i}$ to correlate ($\rho_{2,3} = 0.7$). Condition C frees the correlation between $\epsilon_{0,i}$ and $\epsilon_{2,i}$ as well as $\epsilon_{1,i}$ and $\epsilon_{3,i}$ ($\rho_{0,2} = \rho_{1,3} = 0.3$). Finally, all residuals are correlated in Condition D ($\rho_{0,3} = \rho_{1,2} = 0.2$).

In addition to assessing the bias properties of TSLS estimates when residuals are correlated, I also ensured that the bias of the TSLS $\boldsymbol{\theta}$ estimates do not depend on whether pre-treatment covariates were omitted. As discussed in Section 5, pre-treatment covariates are not confounders in randomized settings. Consequently, the predicted values in the second stage of TSLS should remain unbiased, even as efficiency is lost. To study this, all four conditions listed above were evaluated in the scenario where all six

predictors were observed and in the scenario where the two strongest predictors were omitted.

Finally, consistency implies that any bias should attenuate as the sample size increases. Thus, each simulation condition listed above was evaluated for sample size and was set to $N = 400$, and sample sizes were set to $N = 8{,}000$.

In summary, three conditions were fully crossed to evaluate the consistency properties of the TSLS and conditional mean imputation estimators: (1) the correlation among the residuals, (2) the number of omitted pre-treatment covariates, and (3) the sample size. For all simulation conditions, I conducted 10,000 replications.

Simulation 2: Imbalance Resulting from a Univariate Imputation Estimator

Simulation 2 examines the imbalance that results from conditional mean imputation that assumes all residuals are independent. Earlier I derived an estimate for the correction factor that must be applied for $\mathbb{E}\left[d_i^{imp} \middle| d_i^{miss}\right] = d_i^{miss}$. In this simulation, I study the bias and consistency properties of this estimate by comparing the difference in the estimated multiplier value to the true value that would be needed for $\mathbb{E}\left[d_i^{imp} \middle| d_i^{miss}\right] = d_i^{miss}$ under Condition D. I also study whether the bias of these multiplier estimates is sensitive to predictors being omitted.

Simulation 3: TSLS-Assisted SIMEX PCES Estimators

In the third simulation, I assess bias in the $\boldsymbol{\theta}$ SIMEX estimates when setting the residual correlations to Conditions B and D for small ($N = 400$) and large ($N = 8{,}000$) samples. I constructed $\boldsymbol{\theta}$ SIMEX estimates by subtracting SIMEX estimates of $\boldsymbol{\beta}_0$ from $\boldsymbol{\beta}_1$. As I previously discussed, the functional relationship for $g(\lambda)$ must be assumed to generate SIMEX estimates of the $\boldsymbol{\beta}_z$ parameters. Devanarayan & Stefanski (2002) analytically derived the $g(\lambda)$ relationship in the case of multiple regression with an interaction term; it follows the form $a + \frac{b}{c+\lambda}$, where a, b, and c are constants. Although the true functional form is known in the context of this simulation, I found that estimating the constants is highly unstable except in very large samples ($N > 10{,}000$) because of the vertical asymptote at $\lambda = -c$. In response, I assume that a quadratic function approximates $g(\lambda)$ sufficiently well, i.e. $g(\lambda) \approx \tilde{a} + \tilde{b}\lambda + \tilde{c}\lambda^2$ where $\tilde{a}$ is necessarily equal to the corresponding conditional mean β estimate.

Following the advice given by Devanarayan & Stefanski (2002) and Lockwood & McCaffrey (2015), I simulated measurement error along an equally spaced grid spanning $0 \leq \lambda \leq 2$ with $L = 20$ points along this grid. At each λ value, I conducted $B = 30$ independent simulations to complete the simulation phase of the SIMEX procedure. I observed that $\hat{\beta}_b(\lambda)$ estimates were more dispersed for larger λ values. To account for

this heteroskedasticity, I estimated the $\tilde{b}$ and $\tilde{c}$ constants by fitting a multiple regression model with precision weights.

Simulation Results

Simulation 1 Results

The bias of θ_0, θ_1, θ_2, and θ_3 TSLS estimates across the simulation conditions are listed in Table 2, and the sampling distributions across all replications are shown in Figures 10-13, respectively. I also include the bias from a univariate imputation estimator to serve as a point of reference. The sampling distribution for the θ_1, θ_2, and θ_3 TSLS estimates are all centered near zero across all conditions, suggesting these estimates are not biased. Moreover, the decreasing variance of the sampling distributions with increasing sample size suggests that residual correlations and omitting predictors does not negatively affect the consistency properties of the TSLS estimator.

In contrast, the TSLS estimates for θ_0 are biased whenever a non-zero correlation between $D_i(0)$ and $D_i(1)$ exists. This can be seen in Figure 10 whenever the sampling distribution of the bias estimates are not centered around zero outside of Condition A. Additionally, the dotted verticals in Figure 10 represent the expected bias (i.e. $\theta_3 \times \sigma_{0,1}$) for Conditions B-D. While the sampling distributions are centered around the expected bias value when there are no omitted predictors, the distributions are not centered around

this value when important predictors are omitted. I explain the cause of this unexpected result in the next section.

Table 4 reports the magnitude of the bias for $\boldsymbol{\theta}$ on a standardized metric under each simulation condition. These include the bias of the standardized coefficients for θ_1, θ_2, and θ_3, while the standardized metric for the θ_0 bias estimate is taken to be the ratio with the magnitude of the true ITT value (i.e. $\frac{\hat{\theta}_0-\theta_0}{|ITT|}$). MANOVA testing suggested that the magnitude of the bias for θ_0, θ_1, θ_3 are all zero across all conditions when $N = 8{,}000$. The bias also does not depend on residual correlations other than between $D_i(0)$ and $D_i(1)$ or on whether any pre-treatment predictors were omitted ($F_{9,479988}$= 1.15, p = 0.324). On the other hand, ANOVA testing did suggest that the θ_0 estimate differed when important predictors were omitted (t = -59.05, df = 159998, p<0.001), but the magnitude of the difference did not differ across conditions controlling for sample size ($F_{1,4}$= 0.689, p =0.607).

In summary, the data for Simulation 1 suggest that TSLS produces consistent estimates for θ_1, θ_2, and θ_3, but biased estimates for θ_0 whenever $\sigma_{0,1}$ is nonzero. I also find evidence that the bias approaches $\theta_3 \times \sigma_{0,1}$, but only when none of the predictors are omitted. I explain in the Discussion that this surprising result is likely because omitting predictors induces a larger residual correlation than the population $\sigma_{0,1}$ value specified when generating the data.

Additionally, the magnitude of the bias for the TSLS intercept estimates is quite large relative to the ITT, averaging over 40% of the ITT value in large samples with the bias increasing in small samples. Despite the large amount of bias in the TSLS estimates, conditional mean imputation leads to bias in the estimated intercepts of nearly 80%. This large degree of bias exists *even though* the covariates explained roughly 70% of the variance in the $D_i(0)$ and $D_i(1)$. Because bias in the intercept leads to bias in estimated treatments along the full joint distribution of $D_i(0)$ and $D_i(1)$, this implies that the PCES estimated with TSLS is highly sensitive to the residuals correlation among the intermediate variable. TSLS alone is, therefore, insufficient in addressing the file-matching problem to be a viable estimator in practice.

<Insert Figures 10-13 about here>

<Insert Table 4 about here>

Simulation 2 Results

Figures 14 and 15 display the bias in the estimated multiplier value compared to the true multiplier value that would need to be observed to ensure imputed values are balanced (i.e., $\mathbb{E}\left[d_i^{imp} \middle| d_i^{miss}\right] = d_i^{miss}$) for the treated and control arm, respectively. The

sampling distributions are all centered around zero, providing evidence that the multiplier estimates obtained from the observed data are unbiased across alternative conditions.

<Insert Figures 11-12 about here>

Simulation 3 Results

Figures 16-19 display the sampling distribution of the bias in the $\boldsymbol{\theta}$ estimates for small ($N = 400$) and large ($N = 8{,}000$) across conditions B and D. I include the corresponding sampling distributions for TSLS and conditional mean imputation estimates for reference.

Consistent with the results from Simulation 1, estimates from a conditional mean imputation model display the most bias. SIMEX estimators do attenuate the bias for the intercept estimates, but only slightly. However, the SIMEX estimates display more bias than the TSLS estimates in all conditions. I discuss the lackluster performance of the SIMEX estimator in the next section.

<Insert Figures 16-19 about here>

Discussion and Future Directions

Through a simulation study, two revealing findings emerged that require further explanation. First, the TSLS intercept estimates are even more biased than expected when pre-treatment covariates are omitted. Second, the proposed SIMEX did not fully attenuate the bias associated with the conditional mean imputation estimator.

I explain the first result by expounding on how omitted pre-treatment covariates induce a residual covariance value that is larger in magnitude than the population $\sigma_{0,1}$ value. To explain this, recall that the D potential outcomes were simulated after conditioning on *all six* covariates. When only two are observed, the component of the variance explained by the omitted covariates are confounded into the error structure, i.e.

$$D_i(z) = X_i\boldsymbol{\alpha}_z + \epsilon_{z,i} = X_i^{obs}\boldsymbol{\alpha}_z^{obs} + \left\{\epsilon_{z,i} + X_i^{omit}\boldsymbol{\alpha}_z^{omit}\right\} = X_i^{obs}\boldsymbol{\alpha}_{\boldsymbol{z}}^{\boldsymbol{obs}} + \epsilon_{z,i}^{omit} \tag{22}$$

where $\boldsymbol{\alpha}_z^{obs}$ are the population coefficients corresponding to the vector observed covariates, X_i^{obs}, and $\boldsymbol{\alpha}_{\boldsymbol{z}}^{omit}$ are the population coefficients corresponding to the vector of omitted covariates, X_i^{omit}. The covariance of the residuals in a model estimated by omitting covariates is then given by

$$\mathbb{C}\mathrm{ov}\left(\epsilon_{0,i}^{omit}, \epsilon_{1,i}^{omit}\right) = \sigma_{0,1} + \left(\boldsymbol{\alpha}_0^{omit}\right)'\mathbb{V}\mathrm{ar}\left(X_i^{omit}\right)\boldsymbol{\alpha}_0^{omit} \tag{23}$$

where $\mathbb{V}\mathrm{ar}\left(X_i^{omit}\right)$ is the variance-covariance matrix of the omitted pre-treatment predictors. Thus, omitting covariates results in residuals that no longer exhibit a

covariance exactly equal to $\sigma_{0,1}$. Instead, the shared variance from the omitted covariates induces an additional residual correlation.

The simulation study also revealed that the SIMEX estimator does not substantially attenuate the bias associated with the conditional mean imputation. There are two possible explanations for this finding. First, a quadratic function is a poor approximation for the true $g(\lambda)$. Alternatively, and perhaps compounding the first explanation, is that deviations in the corrected imputation from the true d_i^{miss} are not independently distributed, and the independence assumption is critical for consistent SIMEX estimates.

To identify the source SIMEX's bias was due to a poor approximating function, I assessed how the SIMEX estimator compares to a simulated situation specified so that the missing D potential outcome is fully explained by the covariates. Increasing amounts of measurement error were added to D. The data generating mechanism for the $Y(1)$ potential outcomes followed the form:

$$Y_i(1,\lambda) = \beta_{z,0} + \beta_{z,1}\left(D_i(0) + \lambda^{\frac{1}{2}} \times \mathrm{SEM_z} \times U_i\right) + \beta_{z,2}D_i(1) + \beta_{z,3}\left(\left(D_i(0) + \lambda^{\frac{1}{2}} \times \mathrm{SEM_z} \times U_i\right) \times D_i(1)\right) + \left(\epsilon_{z+2,i} +\right) \quad (24)$$

where $U_i \sim N(0,1)$ and the residuals set to Condition D. I calculated $\boldsymbol{\beta}$ estimates from data generated using Equation 29 using a large sample ($N = 1 \times 10^5$) for $L = 30{,}000$ different λ values spanning $-1 \leq \lambda \leq 6$. The results for the treatment arm are shown in

Figure 20 and indicated by the grey dot. The black line indicates the true $g(\lambda)$ function of the form $a + \frac{b}{c+\lambda}$, denoted $g_{TRUE}(\lambda)$. Thus, $g_{TRUE}(\lambda)$ represents the true functional relationship that would be observed if deviations in the corrected imputation values were independent.

Next, I compared the true function to corresponding results given the data generating mechanism and the SIMEX estimation procedure used in Simulation 3 using $N = 1 \times 10^5$ and $L = 30{,}000$ different λ values spanning $0 \leq \lambda \leq 6$. The corresponding $\boldsymbol{\beta}$ estimates are shown in blue. A best fit line for $g(\lambda)$ of the form $a + \frac{b}{c+\lambda}$ was fit to the data. The best fit line is illustrated by the blue line in Figure 20. The SIMEX estimate is represented by the blue dot, as it is the $g(-1)$ prediction. Given the discrepancies between the $g(-1)$ prediction and the corresponding prediction using $g^{TRUE}(\lambda)$, the results suggest that the bias of the SIMEX estimator is not the result of a poor extrapolation function. Thus, bias would still exist even if $g(\lambda)$ were correctly specified instead of using a quadratic approximation.

If the lackluster performance of the SIMEX estimator is not the result of using a quadratic approximation function, then it must result from a violation of the independence assumption. Specifically, in the simulation phase of the SIMEX procedure, independent measurement error was simulated and added to the corrected imputation value. The underlying assumption here is that the initial discrepancies between the

corrected imputations and the true d_i^{miss} are independent of the joint distribution of all potential outcomes $(D_i(0), D_i(1), Y_i(1), Y_i(0))$.[4] Table 5 reports the pairwise correlation of this discrepancy with each potential outcome. Indeed, some of the potential comes display high correlations with the discrepancy.

<Insert Table 5 about here>

A promising alternative to SIMEX for future studies may involve modeling the full joint distribution $(D_i(0), D_i(1), Y_i(0), Y_i(1))$ and leveraging the consistency properties of TSLS with multiply imputed chained equations (MICE). A MICE procedure would proceed iteratively by sampling the marginal distribution for each D and Y potential outcome given *all other* potential outcomes until convergence is reached. A TSLS-assisted MICE approach would sample from the asymptotic sampling distribution of $\left(\hat{\beta}_{z,1}^{TSLS}, \hat{\beta}_{z,2}^{TSLS}, \hat{\beta}_{z,3}^{TSLS}\right)$ when sampling the marginal distributions $Y_i(z)|(D_i(0), D_i(1), Y_i(z'))$.

Preliminary results suggest that a TSLS-assisted MICE estimator may significantly attenuate bias. Figure 21 shows a single chain of the intercept bias estimates

[4] To be completely accurate, the discrepancy should be independent of X_i as well. Given Assumption 5, however, this correlation with the X_i values should not result in bias given that the potential outcomes are observed.

using the modified MICE estimator with a large sample. The chain appears to converge to a bias value of approximately one-quarter the value of the bias resulting from using TSLS alone.

<Insert Figure 21 about here>

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444-455.

Bartolucci, F., & Grilli, L. (2011). Modeling partial compliance through copulas in a principal stratification framework. *Journal of the American Statistical Association*, *106*(494), 469-479.

Conlon, A. S. C., Taylor, J. M. G., & Elliott, M. R. (2017). Surrogacy assessment using principal stratification and a Gaussian copula model. *Statistical Methods in Medical Research*, *26*(1), 88-107.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, *89*(428), 1314-1328.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.

Devanarayan, V., & Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, *59*(3), 219-225.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 655-671.

Ding, P., & Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 757-777.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21-29.

Gilbert, P. B., & Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, *64*(4), 1146-1154.

Hastie, T., Tibshirani, Robert, & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed., Springer series in statistics). New York: Springer.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*(1), 69-88.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, H., & Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, *103*(481), 101-111.

Kemple, J. J., & Willner, C. J. (2008). *Career academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthood* (pp. 4-5). New York, NY: MDRC.

Little, R., & Rubin, Donald B. (2002). *Statistical analysis with missing data* (2nd ed., Wiley series in probability and statistics). Hoboken, N.J.: Wiley.

Lockwood, J. R., & McCaffrey, D. F. (2017). Simulation-Extrapolation with Latent Heteroskedastic Error Variance. *Psychometrika*, 1-20.

Morgan, S., & Winship, Christopher. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (Analytical methods for social research). New York, NY: Cambridge University Press.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, *5*(3), 215-244.

Reardon, S. F., & Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects?. *Sociological Methods & Research*, *42*(2), 143-163.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, Donald B. "Bayesian inference for causal effects: The role of randomization." *The Annals of statistics* (1978): 34-58.

Schwartz, S. L., Li, F., & Mealli, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, *106*(496), 1331-1344.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, *28*(4), 558-571.

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Appendix: Figures and Tables

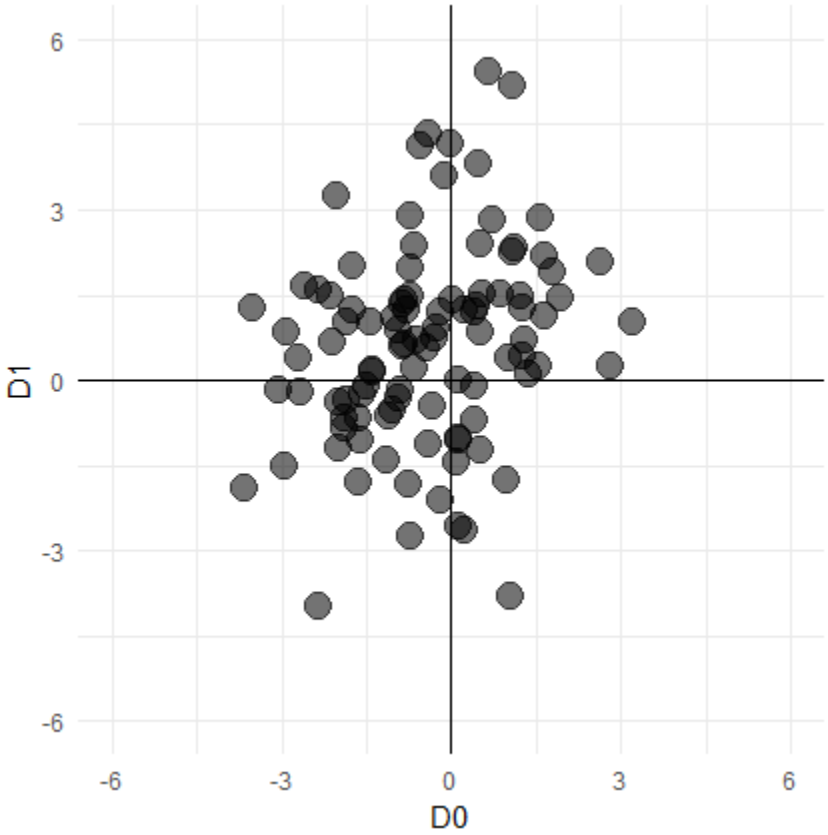

Figure 1: Hypothetical example of a "flat" PCES with no heterogeneity. Size of markers proportional to estimated treatment effect.

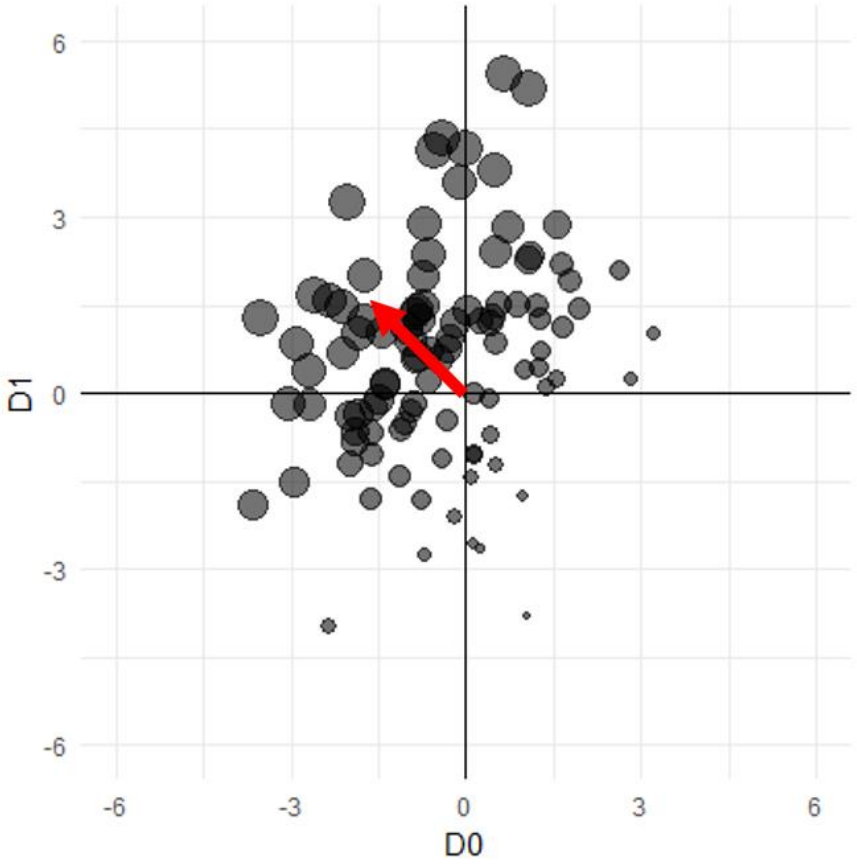

Figure 2: Hypothetical example of a PCES that would be consistent with Page's (2012) hypothesis. Largest treatment effects are found by those students who experienced the greatest gain in work world exposure. Marker size is proportional to treatment effect.

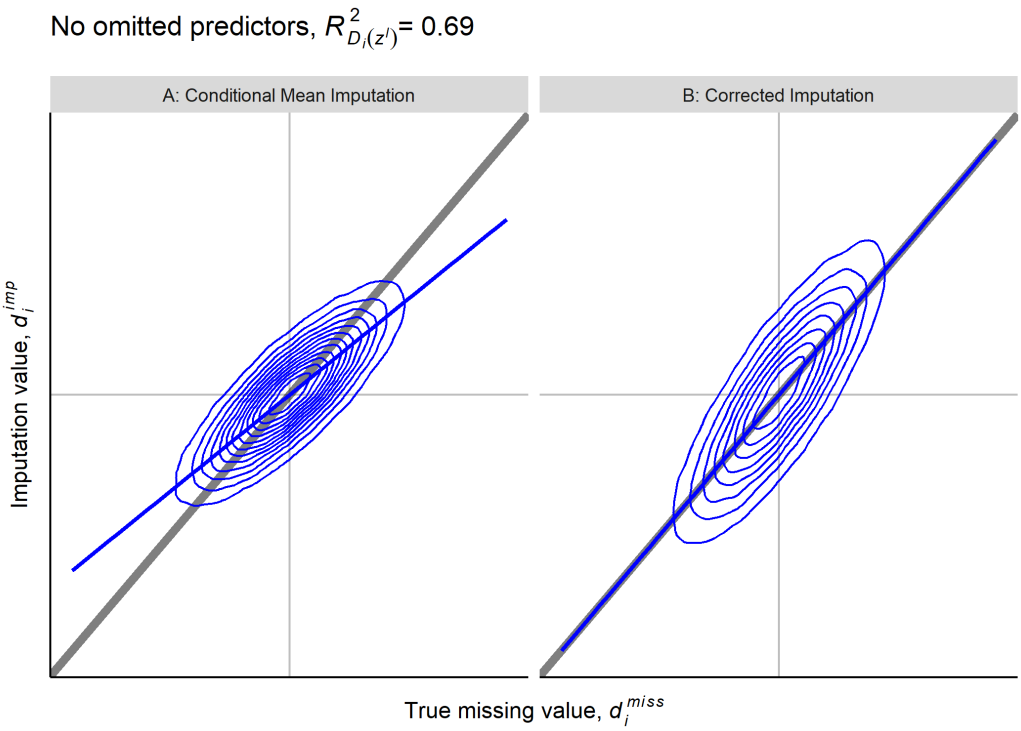

Figure 3: Heuristic of joint distributions of the true missing value (x-axis) and the predicted missing value (y-axis) using OLS with no omitted predictors. The left graph shows regression-to-the-mean of the uncorrected predicted values, while the right graph shows the restoration of balance.

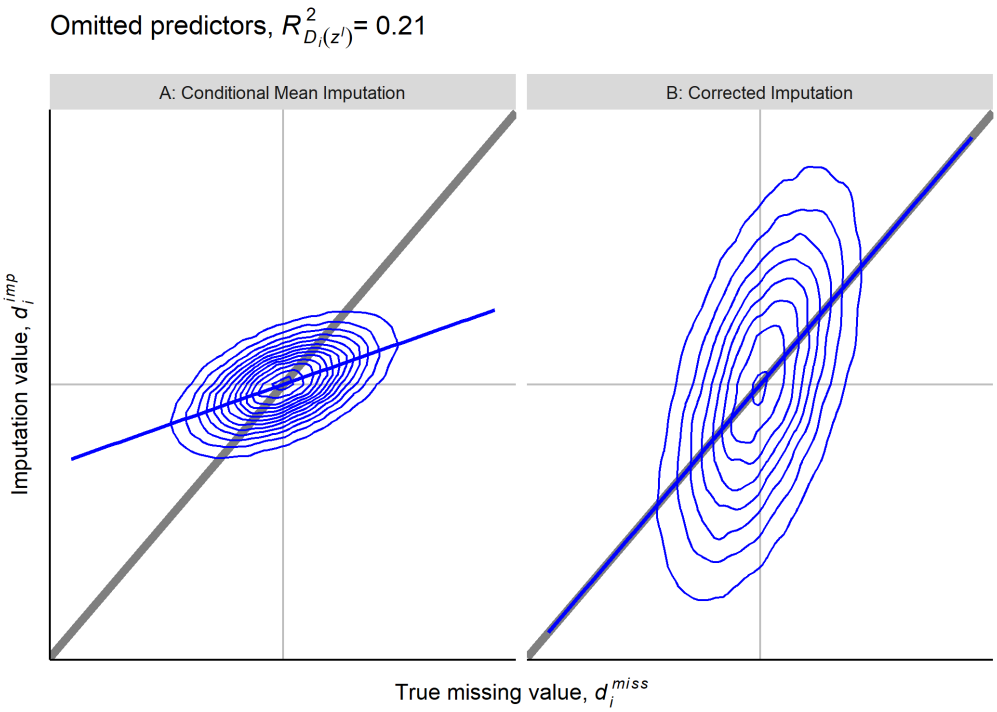

Figure 4: Heuristic of joint distributions of the true missing value (x-axis) and the predicted missing value (y-axis) using OLS with omitted predictors. Right graph shows that the correction can restore balance even when predictors are omitted.

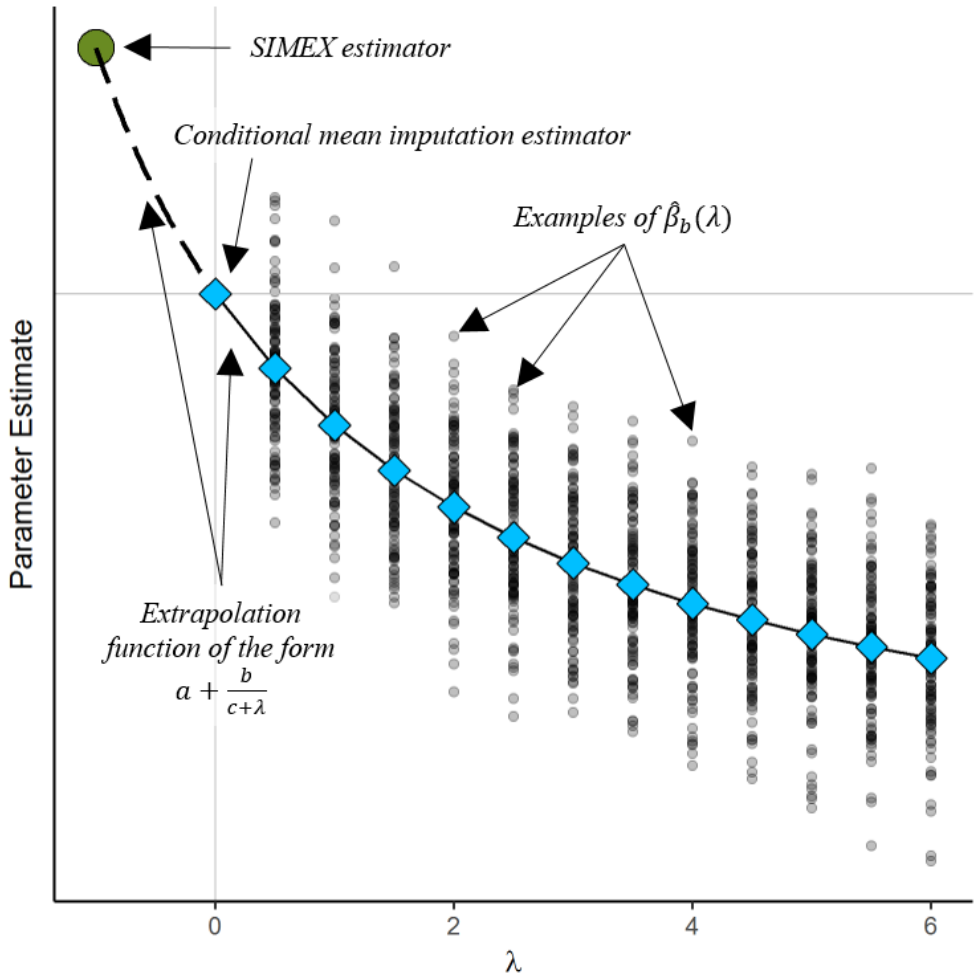

Figure 5: Heuristic of the SIMEX estimator obtained by fitting a function to simulated data with various amounts of measurement error, λ, and then extrapolating. Note that estimate using the raw data corresponds to $\lambda = 0$; the condition that would be observed in the absence of measurement error is represented by $\lambda = -1$. B = 100 simulations plotted at each λ value.

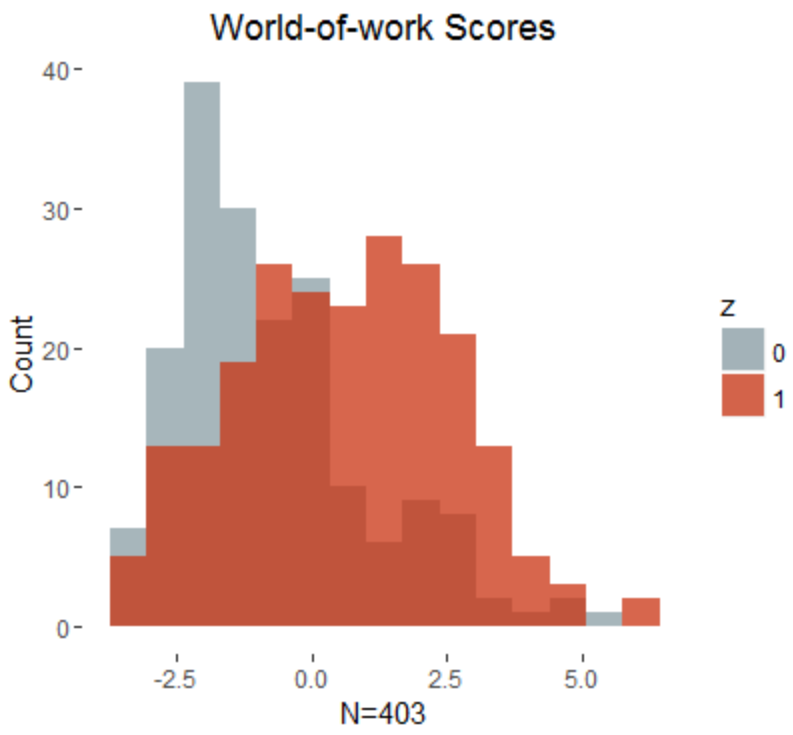

Figure 6: Comparison of world-of-work scores across treatment arms (N=403). Difference in skew suggests the presence of treatment effect heterogeneity on world-of-work scores.

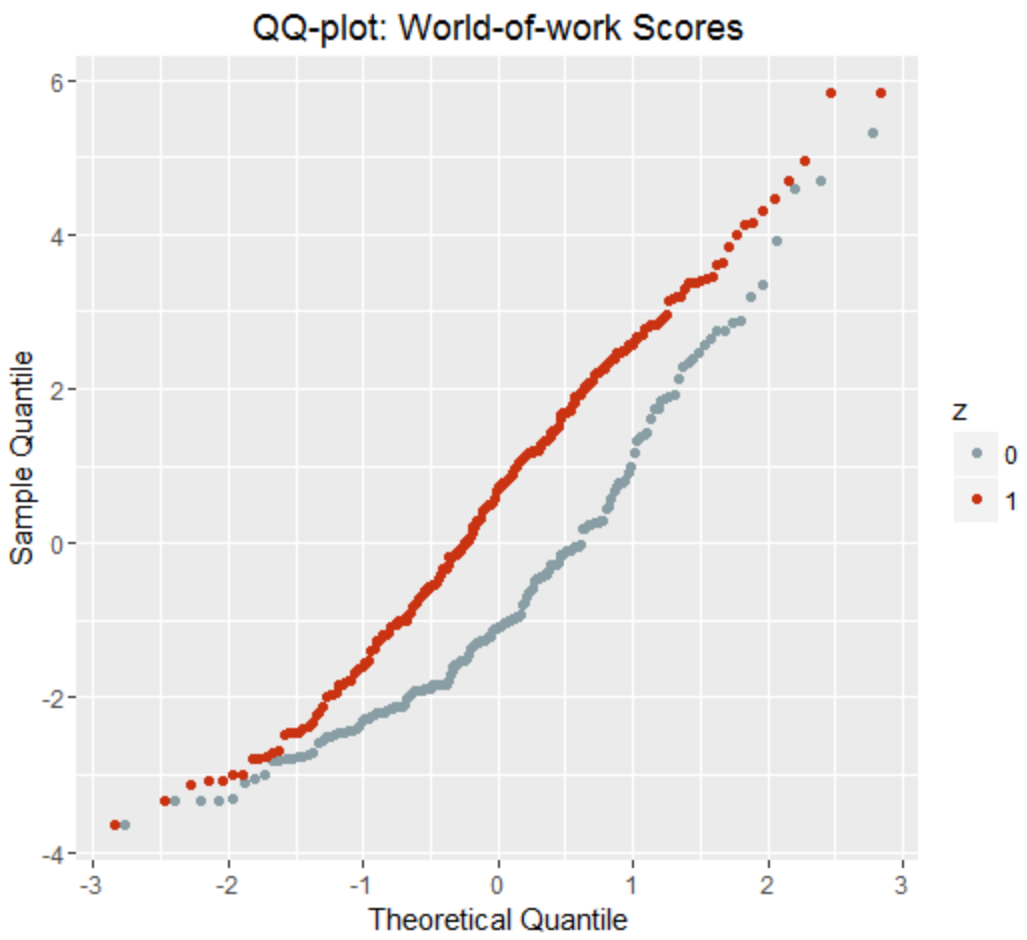

Figure 7: QQ-plots comparing distribution of world-of-work scores across treatment arms.

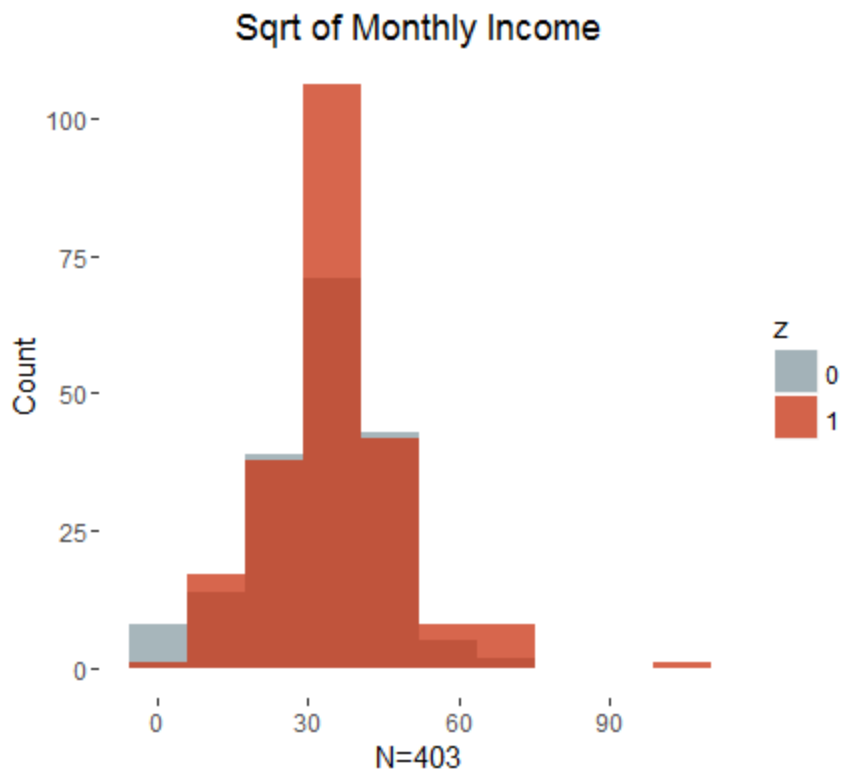

Figure 8: Comparison of wage distribution across treatment arms (N=403).

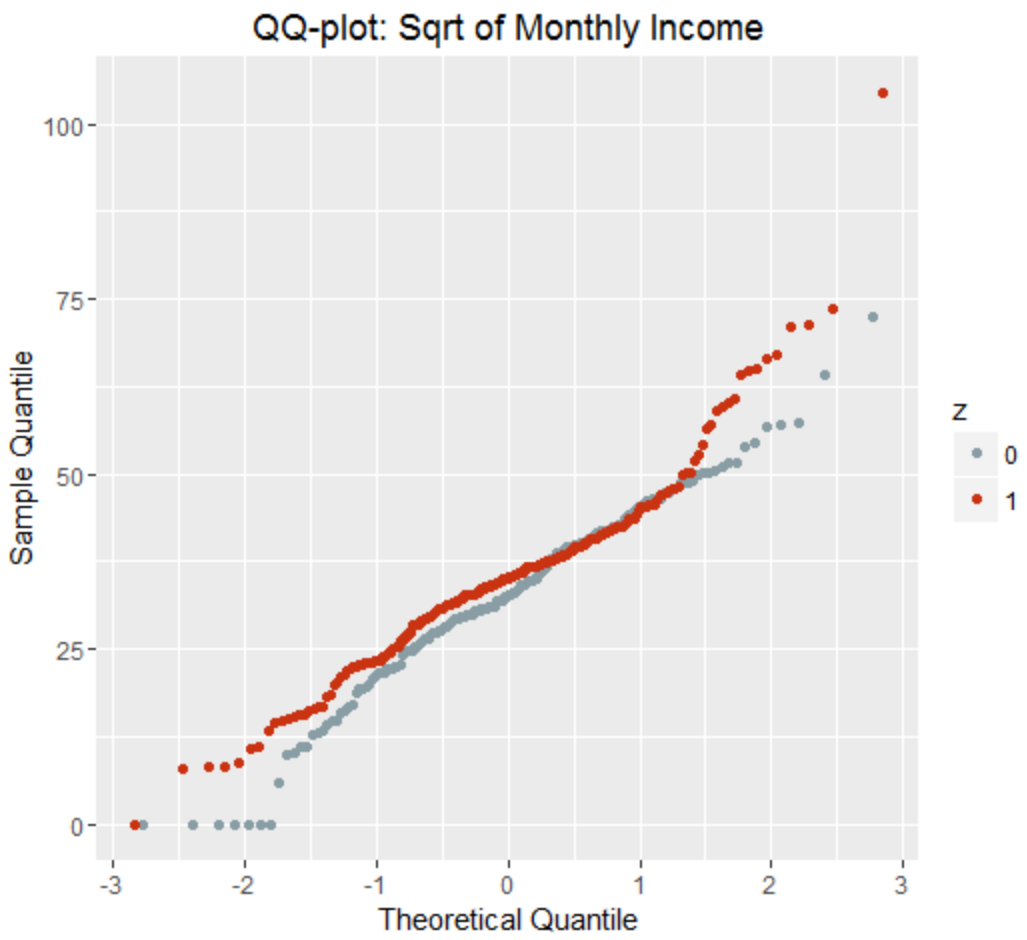

Figure 9: QQ-plot comparing distribution of wages across treatment arms. The difference in the distributions is largest near the tails, suggesting that the kertosis differs and that heterogeneity may exist.

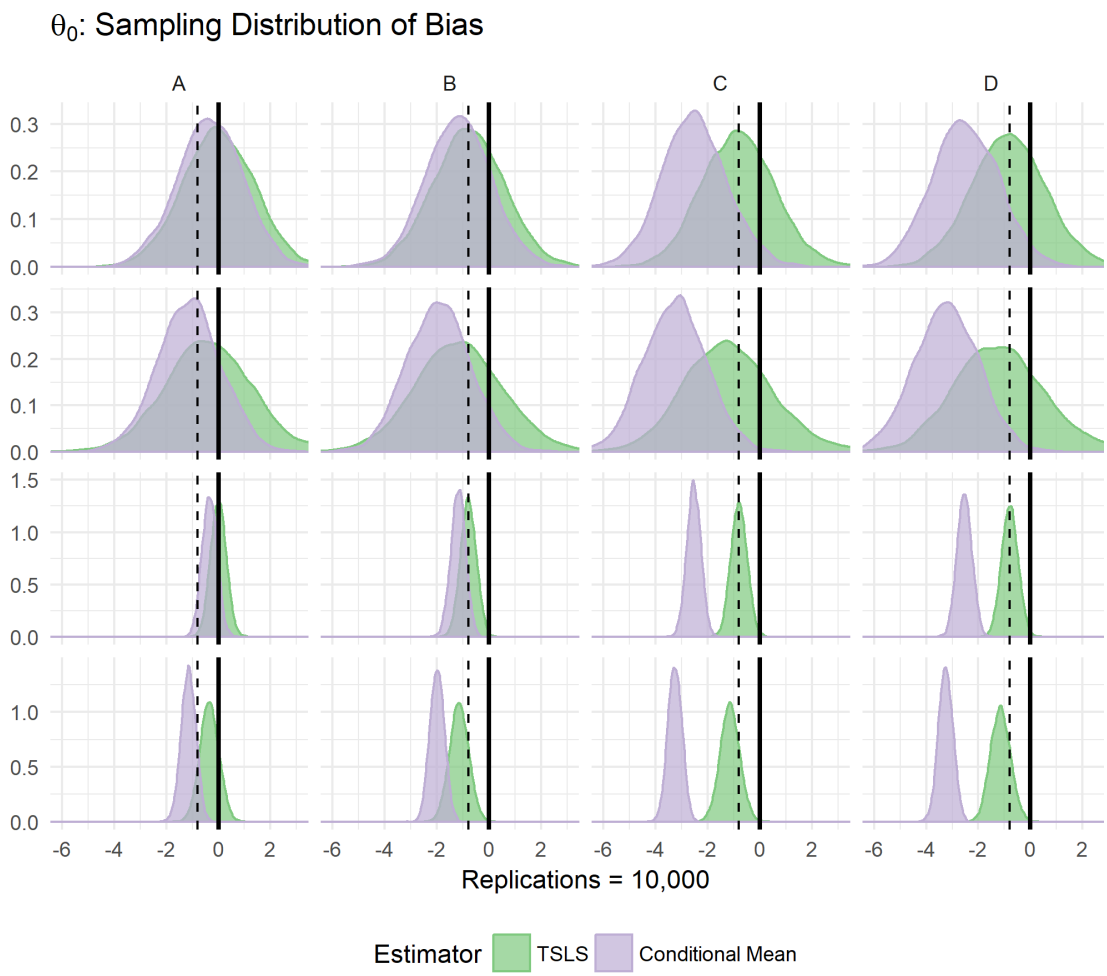

Figure 10: Sampling distribution of simulation results comparing the bias for θ_0 estimated with TSLS and univariate imputation. Dashed line represents the expected bias for the TSLS estimates in conditions 1-3.

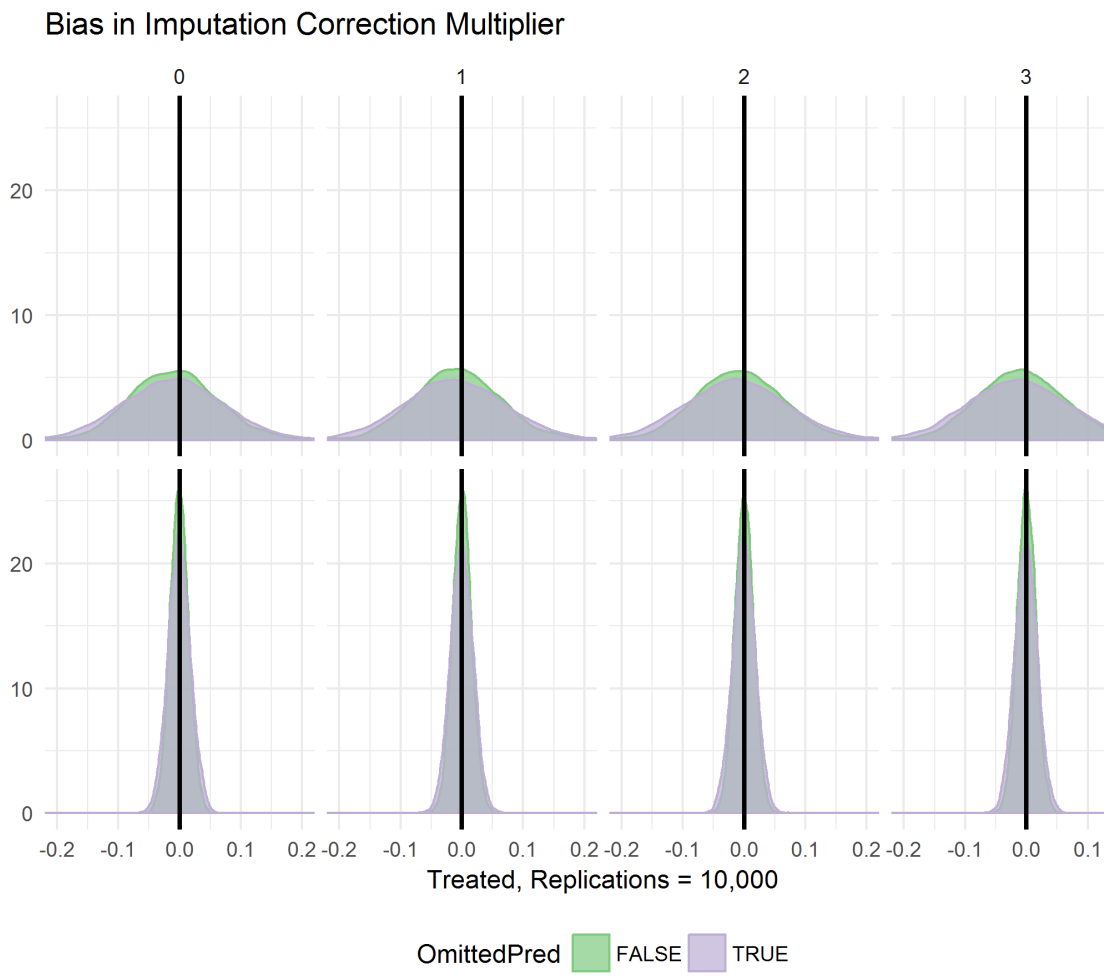

Figure 15: Sampling distribution showing bias in the estimated multiplier values compared to the true value that would be needed to restore balance given. Treated units only.

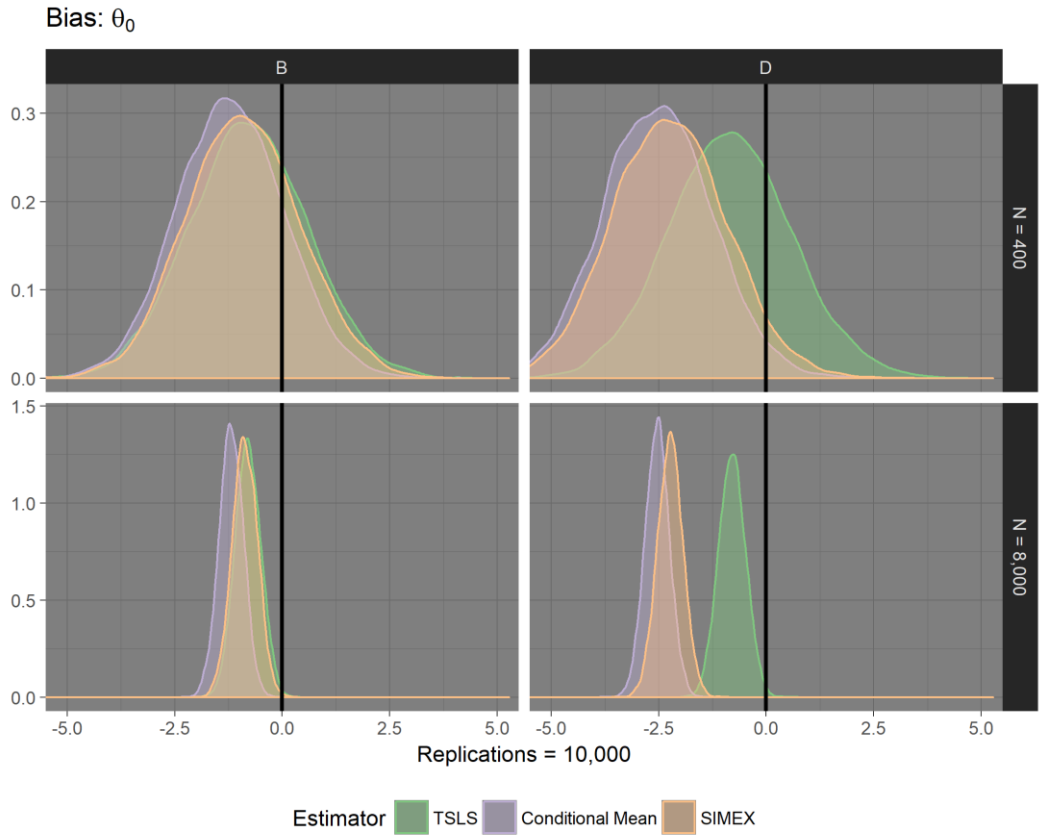

Figure 16: Comparison of sampling distributions of bias for θ_0 across the proposed SIMEX estimators. TSLS and conditional mean imputation included for reference.

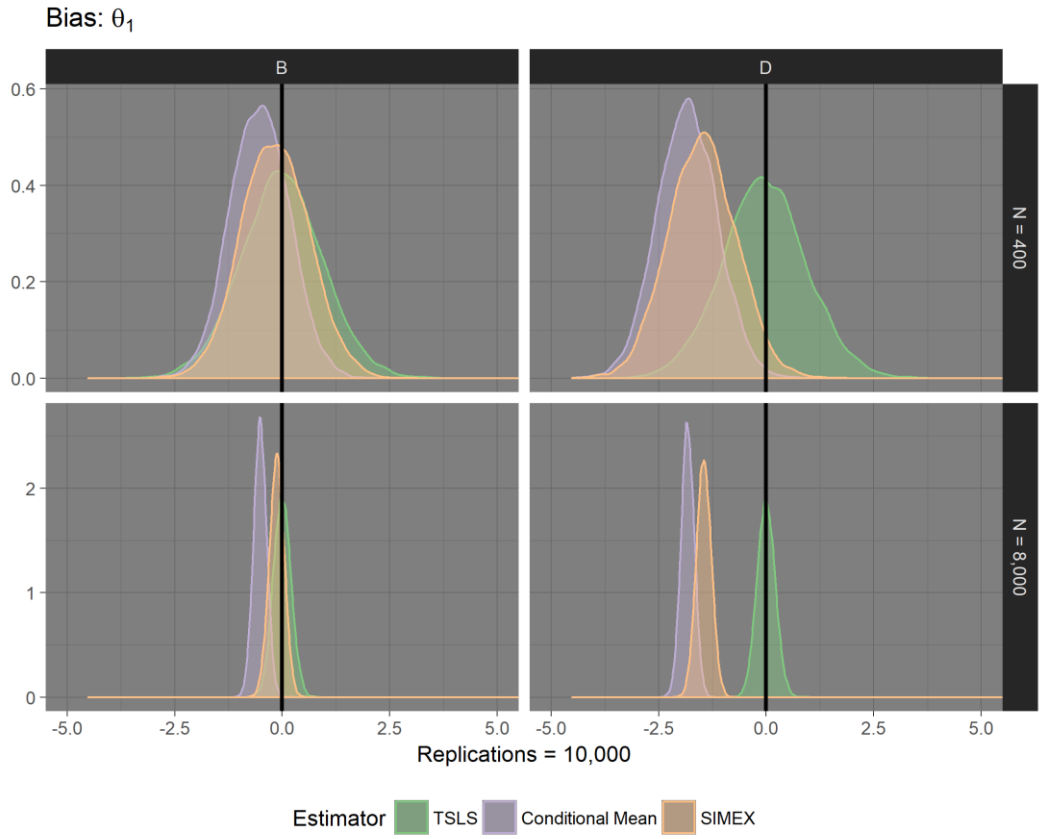

Figure 17: Comparison of sampling distributions of bias for θ_3 across the proposed SIMEX estimators. TSLS and conditional mean imputation included for reference.

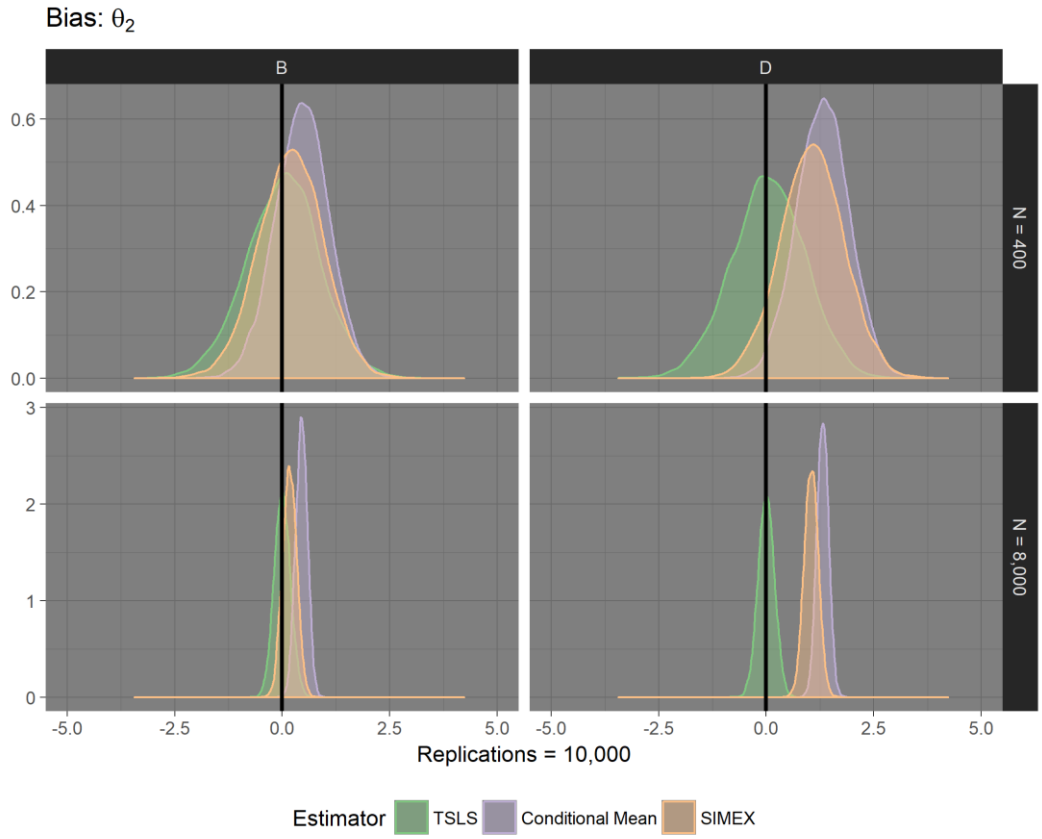

Figure 18: Comparison of sampling distributions of bias for θ_3 across the proposed SIMEX estimators. TSLS and conditional mean imputation included for reference.

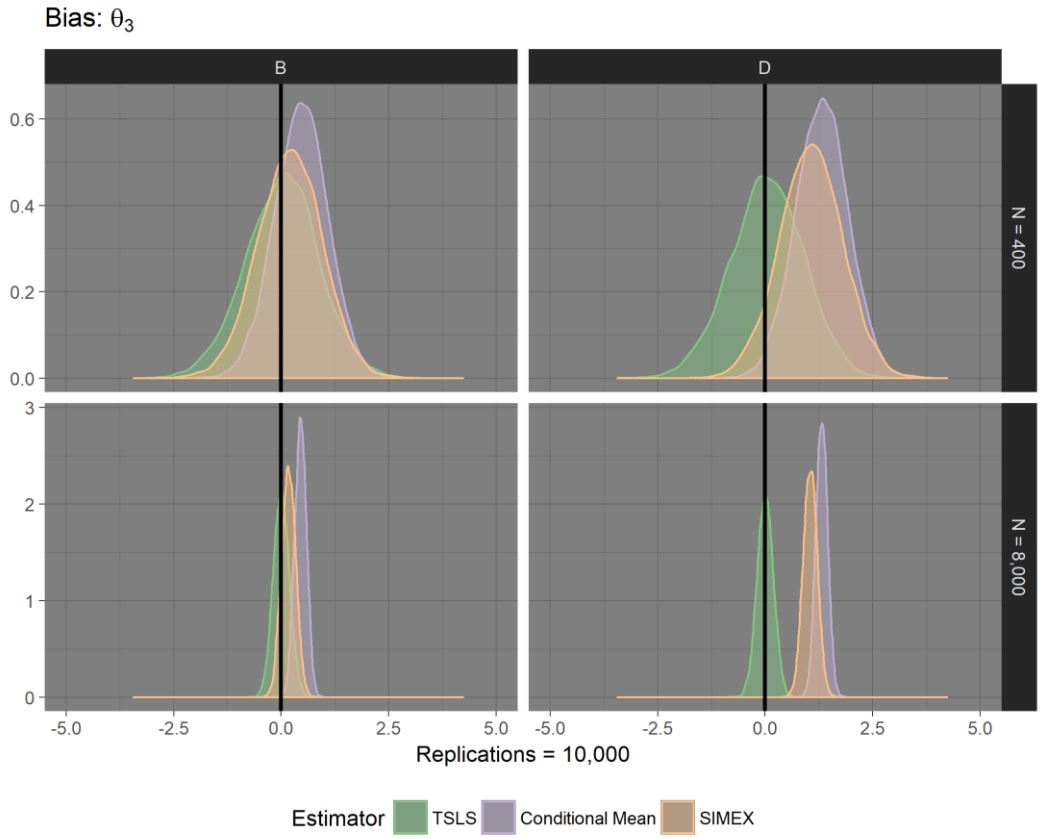

Figure 19: Comparison of sampling distributions of bias for θ_3 across the proposed SIMEX estimators. TSLS and conditional mean imputation included for reference.

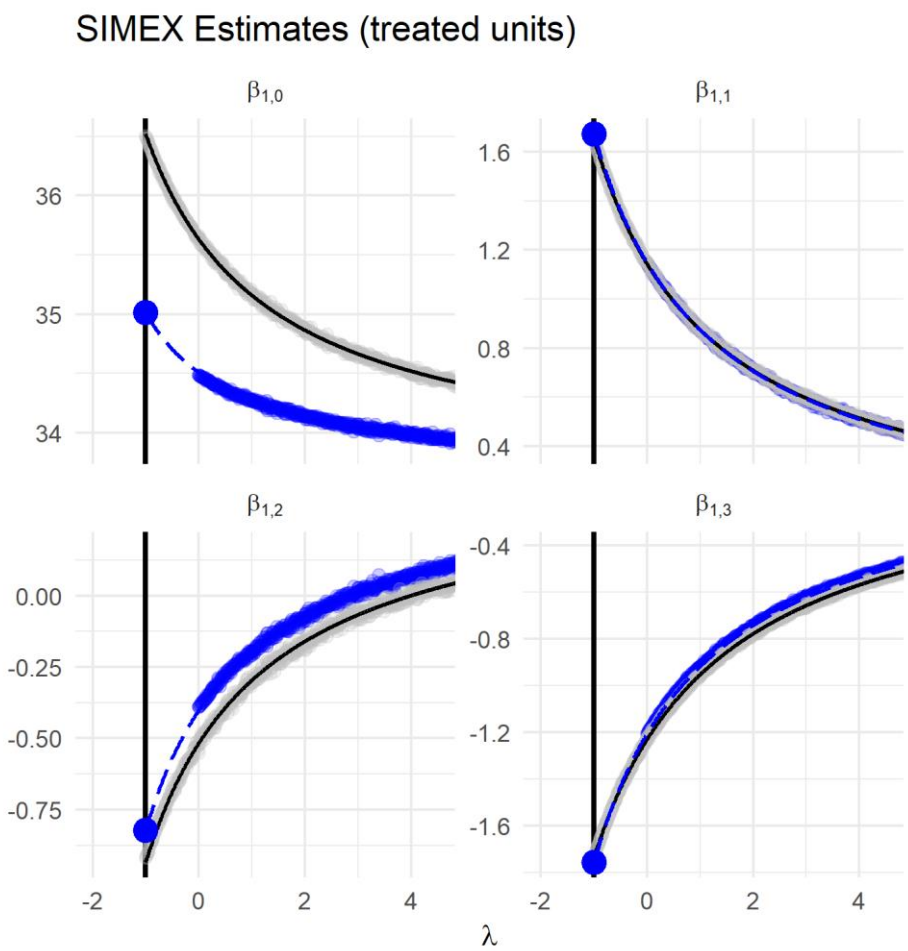

Figure 20: SIMEX estimates (blue) in versus true parameter estimates (black) of $\boldsymbol{\beta}_1$ in a very large sample (N = 100,000) with L = 30,000, B = 1. Residual correlations set to Condition B.

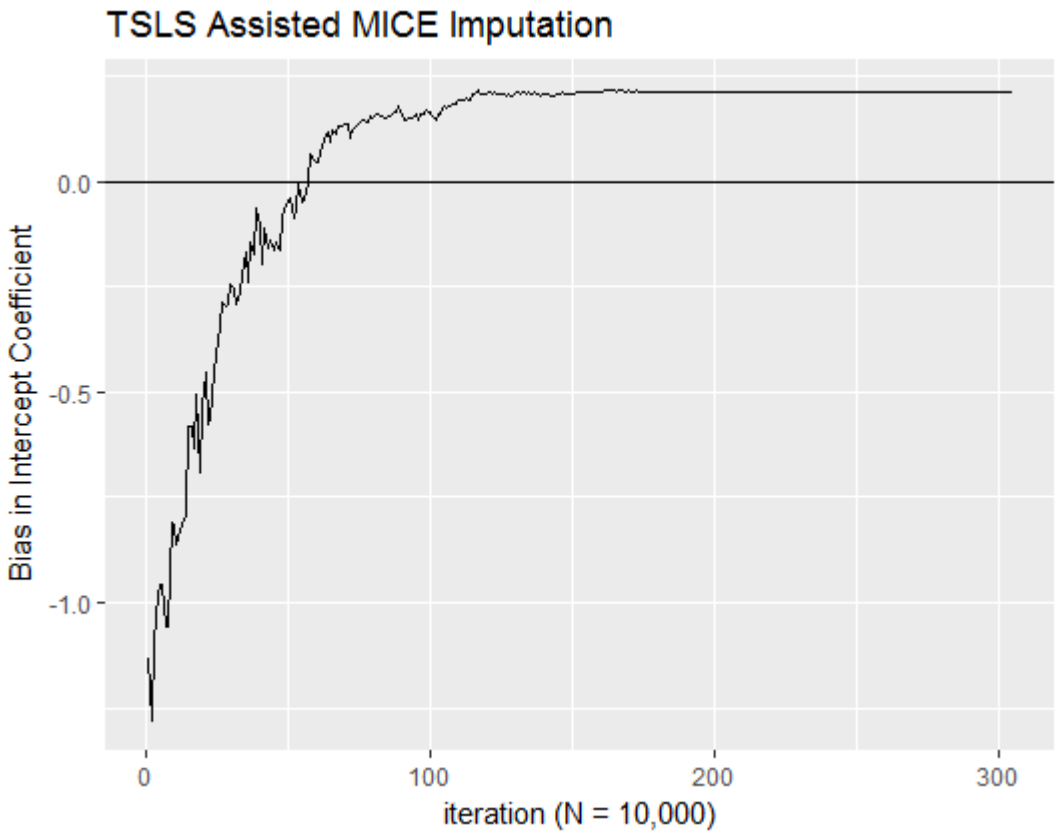

Figure 21: Bias in the chain of θ_0 estimates from a TSLS-assisted MICE estimator given a large sample (N=100,000). Single sample only, no replication. Residual correlations set to Condition D.

Table 1
Descriptive statistics for the endogenous variables.[1]

	World-of-work Scores (d^{obs})		Sqrt of Monthly Income (y^{obs})	
	Control	Treated	Control	Treated
N	182	221	182	221
Mean	-0.71	0.58	32.64	35.35
Std. Dev.	1.77	1.97	12.96	13.00
Skew	0.96	0.07	-0.31	0.91
Kertosis	0.57	-0.58	0.42	3.69

Note: [1]All calculated from an unweighted sample.

Table 2
Parameters and conditions for simulation study

Parameter(s)	Population Value(s)
	Fixed
$\boldsymbol{\alpha}_0$	$[-0.669, -0.624, -0.793, 0.341, 0.506, -0.383, -0.397]'$
$\boldsymbol{\alpha}_1$	$[0.546, -1.256, 0.772, 0.260, -0.348, -0.483, 0.740]'$
$\boldsymbol{\theta}$	$[3.409, 1.010, -0.919, -1.733]'$
$\boldsymbol{\beta}_0$	$[33.095, 0.643, 0, 0]'$
$\boldsymbol{\beta}_1$	$[36.503, 1.653, -0.919, -1.733]'$
σ_0^2	0.781
σ_1^2	1.074
σ_2^2	146.710
σ_3^2	85.762
	Alternative Conditions
Sample Size	{400, 8,000}
Number of X observed[1]	{4, 6}
Residual Correlation Matrix, P	$\left\{ \begin{array}{l} \mathrm{P_A} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\ \mathrm{P_B} = \begin{bmatrix} 1 & & & \\ \rho_{0,1} & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & \rho_{2,3} & 1 \end{bmatrix}, \\ \mathrm{P_C} = \begin{bmatrix} 1 & & & \\ \rho_{0,1} & 1 & & \\ \rho_{0,2} & 0 & 1 & \\ 0 & \rho_{1,3} & \rho_{2,3} & 1 \end{bmatrix}, \\ \mathrm{P_D} = \begin{bmatrix} 1 & & & \\ \rho_{0,1} & 1 & & \\ \rho_{0,2} & \rho_{1,2} & 1 & \\ \rho_{0,3} & \rho_{1,3} & \rho_{2,3} & 1 \end{bmatrix} \end{array} \right\}$ where $\rho_{0,1} = 0.5$, $\rho_{2,3} = 0.7$, $\rho_{0,2} = \rho_{1,3} = 0.3$, $\rho_{0,3} = \rho_{1,2} = 0.2$

Note: [1] adj-$R^2_{D(0)}$=0.40 and adj-$R^2_{D(1)}$=0.22 when only two predictors are observed, while adj-$R^2_{D(0)}$=0.75 and adj-$R^2_{D(1)} = 0.72$ when all six are observed.

Table 3
Data retrieved from Page (2012) to identify population PCES parameters for simulation.

		Midpoint for $D_i(0)$	Midpoint for $D_i(1)$	Principal Causal Effect
Class	$(D_i(0), D_i(1))$	Est.	Est.	Est. (S.E.)[1]
1	(low, low)	-2.69	-1.520	-1.48 (1.29)
2	(low, medium)	-2.69	0.722	3.67 (0.92)
3	(medium, medium)	-1.09	0.722	4.20 (1.05)
4	(low, high)	-2.69	2.530	10.33 (0.72)
5	(medium, high)	-1.09	2.530	5.65 (1.04)
6	(high, high)	0.90	2.530	-0.34 (2.01)

Note: [1]Calculated using the information provided by Table 1 in Page (2012)

Table 4
Bias (in standardized units) in principal causal effect surface parameters only 2 of the 4 predictors observed.[1] Replication = 10,000.

		θ_0		θ_1		θ_2		θ_3	
		Condition (A): Independent residuals							
	Omitted *X*	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***
TSLS	**False**	-0.118	-0.003	-0.003	0.000	0.002	0.000	0.004	0.000
	True	-0.313	-0.134	-0.007	0.000	0.008	0.000	0.012	0.001
Cond. Mean Imp.[2]	**False**	-0.259	-0.131	-0.051	-0.051	0.037	0.036	0.055	0.053
	True	-0.572	-0.431	-0.115	-0.114	-0.084	-0.085	0.122	0.120
		Condition (B): Correlated $D(0)$ & $D(1)$ + correlated $Y(0)$ & $Y(1)$ residuals							
		N = 400	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***
TSLS	**False**	-0.548	-0.418	-0.005	0.000	0.007	0.000	0.012	0.001
	True	-0.728	-0.604	-0.005	0.000	0.009	0.001	0.015	0.001
Cond. Mean Imp.	**False**	-0.751	-0.624	-0.081	-0.081	0.071	0.070	0.085	0.083
	True	-1.171	-1.049	-0.181	-0.182	0.150	0.150	0.193	0.192
		Condition (C): Condition (B) + Correlated $D(z)$ & $Y(z)$ residuals, $z = 0, 1$							
		N = 400	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***
TSLS	**False**	-0.544	-0.420	-0.007	-0.001	0.007	0.000	0.005	0.000
	True	-0.753	-0.600	-0.010	0.000	0.011	-0.001	0.014	0.000
Cond. Mean Imp.	**False**	-1.317	-1.328	-0.248	-0.284	0.205	0.205	0.083	0.083
	True	-1.955	-1.711	-0.377	-0.376	0.273	0.273	0.191	0.191
		Condition (D): Condition (C) + Correlated $D(z)$ & $Y(z')$ residuals, $z \neq z'$							
		N = 400	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***	***N = 400***	***N = 8000***
TSLS	**False**	-0.532	-0.418	-0.006	-0.001	0.008	0.000	0.005	0.001
	True	-0.746	-0.602	-0.011	0.000	0.012	0.000	0.017	0.000
Cond. Mean Imp.	**False**	-1.485	-1.329	-0.286	-0.285	0.205	0.206	0.085	0.083
	True	-1.930	-1.709	-0.379	-0.377	0.275	0.274	0.193	0.190

Notes: [1]The joint distributions of the simulated residuals follow a multivariate normal distribution. For the θ_0 value, standardization was conducted by dividing by the ITT effect.

Table 5
Correlation values between the corrected imputation discrepancy with the true missing value

	Treated $z_i = 0$	Control, $z_i = 1$
$D_i(0)$	-0.45	0.00
$D_i(1)$	0.00	-0.45
$Y_i(0)$	-1.45	0.00
$Y_i(1)$	-0.09	-0.04