JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2012)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2286

ESTIMATION OF TREATMENT EFFECTS WITHOUT AN EXCLUSION RESTRICTION: WITH AN APPLICATION TO THE ANALYSIS OF THE SCHOOL BREAKFAST PROGRAM

DANIEL L. MILLIMET[a,b]* AND RUSTY TCHERNIS[b,c,d]

[a] *Department of Economics, Southern Methodist University, Dallas, TX, USA*
[b] *IZA, Bonn, Germany*
[c] *Georgia State University, Atlanta, GA, USA*
[d] *NBER, Cambridge, MA, USA*

SUMMARY

The increase in childhood obesity has garnered the attention of many in policymaking circles. Consequently, school nutrition programs such as the School Breakfast Program (SBP) have come under scrutiny. The identification of the causal effects of such programs, however, is difficult owing to non-random selection into the program and the lack of exclusion restrictions. Here, we propose two new estimators aimed at addressing this situation. We compare our new estimators to existing approaches using simulated data. We show that while correlations might suggest that SBP causes childhood obesity, SBP is likely to reduce childhood obesity once selection is addressed.

Received 3 November 2009; Revised 12 February 2012


Supporting information may be found in the online version of this article.

1. INTRODUCTION

Empirical researchers in economics and other disciplines are often interested in the causal effect of a binary treatment on an outcome of interest. Often randomization is used to ensure comparability (in expectation) across the treatment and control groups. However, when randomization is not feasible – owing to ethical, budgetary, or political constraints – researchers must rely on non-experimental or observational data. With such data, non-random selection of subjects into the treatment group becomes a paramount concern.

When subjects self-select into the treatment group *only* on the basis of attributes observed by the researcher, there exist many statistical methods appropriate for the estimation of the causal effects of the treatment. The econometric literature on program evaluation has witnessed profound growth in this area over the past few decades.[1] However, when subjects self-select into the treatment group on the basis of attributes unobserved by the researcher, but correlated with the outcome of interest, the estimation of causal effects becomes difficult. The typical strategy is to rely on an instrumental variable (IV). However, a valid instrument is often unavailable. Moreover, even if one is available, it may identify an economically uninteresting parameter in the presence of heterogeneous treatment effects (Imbens and Angrist, 1994).

Here, we are interested in identifying the causal effect of participation in the School Breakfast Program (SBP) on childhood obesity. However, as discussed below, we believe that students in the treatment and control groups differ along important observed and unobserved dimensions. Moreover,

* Correspondence to: Daniel L. Millimet, Department of Economics, Southern Methodist University, Box 0496, Dallas, TX 75275-0496, USA. E-mail: millimet@smu.edu

[1] See Heckman *et al*. (1999), Imbens (2004), and Imbens and Wooldridge (2009) for excellent surveys. Busso *et al*. (2011) provide extensive evaluation of the finite-sample performance of many of these estimators.

we do not have access to a credible instrument. As a result, the usual approach for dealing with non-random selection into SBP – IV using an exclusion restriction – does not seem viable.

To proceed, we utilize three existing estimators that do not rely on an exclusion restriction for identification: the two-step estimator of Heckman's bivariate normal (BVN) selection model that relies on functional form assumptions for identification, an alternative control function (CF) approach outlined in Heckman *et al*. (1999) and Navarro (2008), and a parametric version of a recent IV estimator proposed in Klein and Vella (2009; hereafter KV) that exploits heteroskedasticity for identification. In addition, we also propose two new estimators for the analysis of binary treatments when selection into a treatment is on the basis of unobserved attributes, but one lacks an exclusion restriction.

The first estimator we propose is referred to as the *minimum-biased* (MB) estimator. This estimator entails minimizing the bias when estimating the effect of a treatment using an estimator that requires the conditional independence assumption (CIA), independence between treatment assignment and potential outcomes conditional on observed variables. This is accomplished by trimming the estimation sample to include only observations with a propensity score – the conditional probability of receiving the treatment given the observed variables – within a certain interval. The MB estimator has the advantage of being unbiased when the CIA holds, but minimizing the bias associated with estimators that require the CIA when this assumption fails (under certain conditions). However, it is important to stress that the MB estimator accomplishes this at the expense of changing the parameter being estimated. In this sense, our approach is similar to the strategy advocated in Crump *et al*. (2009) for dealing with limited overlap in the distributions of the observed variables in the treatment and control group.

The second estimator is referred to as the *bias-corrected* (BC) approach. While this estimator relies heavily on the BVN model to estimate the bias of estimators requiring the CIA when this assumption fails, it does not require specification of the functional form for the outcome of interest in the final step. Moreover, unlike the MB estimator, the BC estimator does not change the parameter being estimated.

Before applying these methods to estimate the causal effect of the SBP on child health, we first conduct an extensive Monte Carlo study. We do so not only using purely simulated data, but also using an Empirical Monte Carlo study design recently suggested in Huber *et al*. (2010). The Monte Carlo study complements recent study by Busso *et al*. (2011) which examine the finite-sample performance of many estimators requiring the CIA, as well as the simulations in Heckman *et al*. (1999).

Our Monte Carlo exercise yields several striking findings. First and foremost, researchers ought to be skeptical of estimates obtained using methods identified from functional form assumptions unless the correct functional form is known (e.g. Goldberger, 1983). Second, when the appropriate functional form is under-specified in that relevant regressors are omitted from the model, the MB estimator improves upon the performance of commonly used estimators that require the CIA, as well as those relying on functional form or heteroskedasticity for identification. Given that this is perhaps the most relevant case for applied researchers, this suggests the MB estimator should be incorporated into the practitioner's toolbox. Third, when the appropriate functional form is known (or is over-specified in that irrelevant regressors are also included in the model), our BC estimator, along with BVN, perform well among the estimators considered. Finally, the KV estimator exploiting heteroskedasticity performs relatively well when the error in the treatment equation is in fact heteroskedastic, but (not surprisingly) less well if this is not the case. The KV estimator is also more sensitive to over-specifying the model.

In terms of the application, our new estimators prove to be a nice complement to existing methods when the CIA fails, but one lacks an exclusion restriction. Specifically, we find a positive and statistically significant *association* between SBP and child weight when using estimators that require the CIA. The relationship remains positive, but becomes statistically insignificant, when we use our MB estimator. Finally, consistent with prior suggestive evidence in the literature, we find a negative and sometimes statistically significant *causal* effect of SBP participation on child weight using the BC, BVN,

DOI: 10.1002/jae

and KV estimators. This pattern of point estimates highlights the importance of controlling for non-random selection (on unobserved variables) in studies of the efficacy of the SBP. It also suggests there is little evidence of a causal nature that the SBP contributes to childhood obesity, and some evidence that it reduces the incidence of obesity.

The remainder of the paper is organized as follow. Section 2 briefly discusses the state of childhood obesity in the USA and the literature on school meal programs. Section 3 begins by providing a quick overview of the potential outcomes and corresponding treatment effects framework. Next, it details the bias of estimators that require conditional independence. Finally, we introduce our MB and BC estimators, and discuss the existing BVN, CF, and KV estimators. Section 4 presents the Monte Carlo study. Section 5 contains our analysis of the SBP. Section 6 concludes.

2. BACKGROUND

It is well known that childhood obesity has increased dramatically in the USA since the 1970s. Data from the National Health and Nutrition Examination Surveys (NHANES) covering 1976–1980 and 2003–2006 indicate that the prevalence of overweight preschool-aged children, aged 2–5 years, increased from 5.0% to 12.4%.[2] Among school-aged children, the prevalence has risen from 6.5% to 17.0% for those aged 6–11, and from 5.0% to 17.6% for those aged 12–19 years.[3]

This rise has severe, long-run implications given that roughly one-third of overweight preschool-aged children and one-half of overweight school-aged children become obese adults (Serdula *et al*., 1993). In turn, adult obesity is associated with numerous health and socio-economic problems. Trasande *et al*. (2009) estimate that childhood obesity resulted in US $237.6 million in hospitalization costs alone in 2005, up (in real terms) from $125.9 million in 2001. Finkelstein *et al*. (2009) estimate that the total medical costs from child and adult obesity exceeded $100 billion in 2006 and may have been as high as $147 billion in 2008.

Given this backdrop, policymakers in the USA have acted in a number of different directions, particularly within schools. These reforms culminated in November 2007 as the US Department of Health and Human Services launched the Childhood Overweight and Obesity Prevention Initiative. In addition to these recent policy developments, two long-standing federal programs affecting more than 30 million students on a typical school day have come under scrutiny: the SBP and the National School Lunch Program (NSLP).

The existing literature analyzing these programs relies on non-experimental data where the potential exists for non-random selection into treatment on the basis of student-specific unobserved attributes. Three noteworthy approaches have been employed in an attempt to circumvent this issue. First, Bhattacharya *et al*. (2006) employ a difference-in-difference (DD) strategy and compare the weight and nutritional intake of children during the summer to that during the school year across schools that do and do not offer breakfast. The authors find evidence that the SBP improves nutritional intake without increasing total calories consumed. Second, Schanzenbach (2009) utilizes a regression discontinuity (RD) approach that exploits the sharp income cut-off for eligibility for reduced-price meals to assess the impact of the NSLP. She finds that NSLP participation increases the probability of being obese due to the additional calories provided by school lunches (see also Campbell *et al*., 2011). Finally, Millimet *et al*. (2010; hereafter MTH) apply the methodology developed in Altonji *et al*. (2005) to assess the sensitivity of the effects of participation in the SBP and NSLP estimated under exogeneity to

[2] Overweight is defined as an age- and gender-specific body mass index (BMI) greater than the 95th percentile based on growth charts from the Centers for Disease Control (CDC). These charts define the 95th percentile based on the relevant distribution devised in the 2000 growth charts obtained using national survey data spanning the period 1963–1994. See http://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf.

[3] See http://www.cdc.gov/obesity/childhood/index.html.

DOI: 10.1002/jae

deviations from this assumption. The authors find suggestive evidence of a beneficial (harmful) effect of SBP (NSLP).

Although informative, each of these approaches has limitations. The DD strategy assumes that schools do not decide to participate in the SBP on the basis of the weight trajectories of their students, and that children self-selecting into schools that participate in the SBP do not have different trends in weight and nutrition relative to students attending schools not offering breakfast. In addition, the treatment is school-level breakfast availability, not student-level SBP participation. The RD approach potentially confounds the effects of participation in the SBP and NSLP (as well as other transfer programs) since common eligibility criteria are used. Moreover, the RD approach estimates the local average treatment effect (LATE) when the treatment effect is heterogeneous.[4] Finally, the Altonji *et al*. (2005) approach is merely suggestive, and fails to provide point estimates of the causal effects of the programs except for binary outcomes under functional form assumptions.

In this study, we assess the causal effect of the SBP on child obesity, as well as propose new techniques applicable to the general evaluation of binary treatments. We focus on the SBP for two reasons. First, MTH and Schanzenbach (2009) find little evidence that NSLP participation is non-random conditional on observed variables. Second, MTH finds evidence of *positive selection* into the SBP: children with steeper weight trajectories are more likely to participate conditional on observed variables. Moreover, addressing non-random selection into the SBP is vital not only for obtaining consistent estimates of the causal effect of SBP participation, but also for estimating the causal effect of NSLP participation.[5] MTH conclude that *if* there is positive selection into SBP participation on child weight gains, *then* the SBP (NSLP) lowers (raises) child weight once this non-random selection is addressed. Because the underlying selection into SBP is crucial for consistently estimating the causal effects of both programs, and the analysis in MTH is only suggestive of a beneficial impact of SBP participation once selection is addressed, obtaining point estimates of the causal effect of SBP participation utilizing different techniques is vital.

3. THE EVALUATION PROBLEM

3.1. Setup

Consider a random sample of N individuals from a large population indexed by $i = 1, \ldots, N$. Utilizing the potential outcomes framework (see, for example, Neyman, 1923; Fisher, 1935; Roy, 1951; Rubin, 1974), let $Y_i(T)$ denote the potential outcome of individual i under treatment T, $T \in \mathcal{T}$. Here, we consider only the case of binary treatments: $\mathcal{T} = \{0, 1\}$. The causal effect of the treatment ($T = 1$) relative to the control ($T = 0$) is defined as the difference between the corresponding potential outcomes. Formally, $\tau_i = Y_i(1) - Y_i(0)$.

In the evaluation literature, several population parameters are of potential interest. The most commonly used include the ATE, the ATT, and the ATU. These are defined as

$$\tau_{\text{ATE}} = \text{E}[\tau_i] = \text{E}[Y_i(1) - Y_i(0)] \tag{1}$$

$$\tau_{\text{ATT}} = \text{E}[\tau_i | T = 1] = \text{E}[Y_i(1) - Y_i(0) | T = 1] \tag{2}$$

[4] While the LATE is the parameter of interest if one wishes to estimate the impact of marginal expansions in eligibility for free or reduced-price meals, the LATE may not be very useful in estimating the impact of making school meals universally available at subsidized prices. Moreover, if there is only partial compliance at the discontinuity, then the parameter being estimated represents a different LATE. See Lee and Lemieux (2010) for a general discussion.

[5] This is true despite MTH concluding that there is no selection into NSLP on the basis of student-level unobservables. The reason is that if SBP is erroneously treated as exogenous, then the bias due to correlation with the error term spills over and also biases the coefficient on NSLP participation, since SBP and NSLP participation are highly correlated.

DOI: 10.1002/jae

$$\tau_{\text{ATU}} = \text{E}[\tau_i|T=0] = \text{E}[Y_i(1) - Y_i(0)|T=0] \tag{3}$$

In general, the parameters in ((1))–((3)) may vary with a vector of covariates, X. As a result, each of the parameters may be defined conditional on a particular value of X as follows:

$$\tau_{\text{ATE}}[X] = \text{E}[\tau_i|X] = \text{E}[Y_i(1) - Y_i(0)|X] \tag{4}$$

$$\tau_{\text{ATT}}[X] = \text{E}[\tau_i|X, T=1] = \text{E}[Y_i(1) - Y_i(0)|X, T=1] \tag{5}$$

$$\tau_{\text{ATU}}[X] = \text{E}[\tau_i|X, T=0] = \text{E}[Y_i(1) - Y_i(0)|X, T=0] \tag{6}$$

The parameters in (1)–(3) are obtained by taking the expectation of the corresponding parameter in (4)–(6) over the distribution of X in the relevant population (the unconditional distribution of X for the ATE, and distribution of X conditional on $T=1$ and $T=0$ for the ATT and ATU, respectively).

For each individual, we observe the triple $\{Y_i, T_i, X_i\}$, where Y_i is the observed outcome, T_i is a binary indicator of the treatment received, and X_i is a vector of covariates. The only requirement of the covariates included in X_i is that they are predetermined (that is, they are unaffected by T_i) and do not perfectly predict treatment assignment (although this latter requirement is not necessary for all the estimators discussed here). The relationship between the potential and observed outcomes is given by $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$, which makes clear that only one potential outcome is observed for any individual.

The methods utilized by researchers to circumvent this missing data problem are classified into two groups: selection on observed (SOO) variables estimators and selection on unobserved (SOU) variables estimators. The distinction lies in whether a method consistently estimates the causal effect of the treatment in the presence of unobserved attributes of subjects that are correlated with both treatment assignment and the outcome of interest conditional on the set of observed variables. Assuming a lack of such unobserved variables is referred to as the *conditional independence* (CIA) or *unconfoundedness* assumption (Rubin, 1974; Heckman and Robb, 1985). Formally, under the CIA, treatment assignment is said to be independent of potential outcomes conditional on the set of covariates, X. As a result, selection into treatment is random conditional on X and the average effect of the treatment can be obtained by comparing outcomes of individuals in different treatment states with identical values of the covariates. To 'solve' the dimensionality problem that is likely to arise if X is a lengthy vector, Rosenbaum and Rubin (1983) propose using the propensity score, $P(X_i) = Pr(T_i = 1|X_i)$, instead of X as the conditioning variable.[6]

If the CIA fails to hold, then consistent estimation of the causal effect requires a SOU estimation technique. The difficulty in this case is that obtaining a consistent *point estimate* of a measure of the treatment effect typically requires an exclusion restriction (i.e. an observed variable that impacts treatment assignment, but not the outcome of interest conditional on treatment assignment).

Unfortunately, valid exclusion restrictions, as usually conceived, are not available in many situations. At that point, researchers have limited options: (i) abandon point estimation, instead focusing on bounding treatment effects; (ii) rely on functional form assumptions or a small number of observations with propensity scores drawn from the tails of the distribution for identification; or (iii) identify the causal effect from higher moments of the observed variables. Here, we propose two new options in addition to applying previous estimators utilizing functional form assumptions and

[6] Reliance on the propensity score does not technically solve the dimensionality problem as it simply replaces one dimensionality problem with another: instead of matching observations on the basis of X, now one must aggregate X into a scalar measure of the probability of receiving treatment. However, the latter is more easily accomplished using a flexible parametric model for treatment assignment.

DOI: 10.1002/jae

heteroskedasticity for identification. Our first estimator, referred to as the *minimum-biased* estimator, entails the utilization of a SOO estimator, but trims the estimation sample so as to minimize the bias arising from the failure of the CIA. The second estimator, referred to as the *bias-corrected* estimator, removes the bias entirely from the SOO estimator.

To proceed, we first derive the bias under certain assumptions (namely, joint normality) of estimators that require the CIA when, in fact, the CIA fails. We then derive our MB and BC estimators under these same assumptions. Finally, we extend our estimators to the case where joint normality does not hold.

3.2. Bias when the CIA Fails

Given knowledge of the propensity score, or an estimate thereof, and sufficient overlap between the distributions of the propensity score across the $T=1$ and $T=0$ groups (typically referred to as the *common support* condition; see Dehejia and Wahba, 1999; Smith and Todd, 2005; Khan and Tamer, 2010), the parameters discussed above can be estimated in a number of ways under the CIA (see footnote 1). Regardless of which technique is employed, each will be biased if the CIA fails to hold.

Black and Smith (2004) and Heckman and Navarro-Lozano (2004) consider the bias when estimating the ATT under the CIA and the assumption is incorrect. The bias of the ATT at some value of the propensity score, *P(X)*, is given by

$$B_{\text{ATT}}[P(X)] = \hat{\tau}_{\text{ATT}}[P(X)] - \tau_{\text{ATT}}[P(X)] = \text{E}[Y(0)|T=1, P(X)] - \text{E}[Y(0)|T=0, P(X)] \tag{7}$$

where $\hat{\tau}_{\text{ATT}}$ refers to some propensity score-based estimator of the ATT requiring the CIA.

To better understand the behavior of the bias, Black and Smith (2004) and Heckman and Navarro-Lozano (2004) make the following two assumptions:

A1. Potential outcomes and latent treatment assignment are additively separable in observed and unobserved variables:

$$\begin{aligned} Y(0) &= g_0(X) + \varepsilon_0 \\ Y(1) &= g_1(X) + \varepsilon_1 \\ T^* &= h(X) - u \\ T &= \begin{cases} 1 & \text{if } T^* > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

A2. $\varepsilon_0, \varepsilon_1, u \sim N_3(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \rho_{01} & \rho_{0u} \\ & \sigma_1^2 & \rho_{1u} \\ & & 1 \end{bmatrix}$$

Given A1 and A2, (7) simplifies to

$$B_{\text{ATT}}[P(X)] = -\rho_{0u}\sigma_0 \frac{\phi(h(X))}{\Phi(h(X))[1-\Phi(h(X))]} \tag{8}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution function, respectively. As noted in Black and Smith (2004), $B_{\text{ATT}}[P(X)]$ is minimized when $h(X)=0$, which implies that $P(X)=0.5$. Thus the authors recommend that researchers estimate τ_{ATT} using the 'thick support' region of the propensity score (e.g. $P(X) \in (0.33, 0.67)$) as a sensitivity analysis.

DOI: 10.1002/jae

Prior to continuing, it is important to note that if the ATT varies with X (and, hence, $P(X)$), then using only observations on the thick support estimates a *different* parameter from the population ATT given in (2). Indeed, the procedure suggested in Black and Smith (2004) accomplishes the following. It searches over the parameters defined in (5) to find the value of $P(X)$ for which the $\tau_{\text{ATT}}[P(X)]$ can be estimated with the least bias. Stated differently, when unconfoundedness fails, τ_{ATT}, the population ATT, cannot be estimated in an unbiased manner using estimators that rely on this assumption. Rather than invoking different assumptions to identify the population ATT (e.g. those utilized by SOU estimators), the Black and Smith (2004) approach identifies the *parameter that can be estimated with the smallest bias under unconfoundedness*.[7] Whether or not the parameter being estimated with the least bias, $\hat{\tau}_{\text{ATT}} = \text{E}[\text{E}[\tau_i|P(X), T = 1]]$, where the outer expectation is over $X|0.33 < P(X) < 0.67$ and $T = 1$, is an interesting economic parameter is a different question. The key point, however, is that when restricting the estimation sample to observations with propensity scores contained in a subset of the unit interval, the parameter being estimated will generally differ from the population ATT unless the treatment effect does not vary with X (i.e. $\text{E}[\tau_i|P(X), T = 1] = \text{E}[\tau_i|T = 1]$).

With this point in mind, consider the ATE and the ATU. For the ATU, it is trivial to show that

$$B_{\text{ATU}}[P(X)] = \text{E}[\varepsilon_1|T = 1, P(X)\,] - \text{E}[\varepsilon_1|T = 0, P(X)] = -\rho_{1u}\sigma_1 \frac{\phi(h(X))}{\Phi(h(X))[1 - \Phi(h(X))]} \tag{9}$$

which is also minimized when $h(X) = 0$, or $P(X) = 0.5$. However, it is useful to note that

$$\begin{aligned} B_{\text{ATU}}[P(X)] &= \text{E}[\delta + \varepsilon_0|T = 1, P(X)] - \text{E}[\delta + \varepsilon_0|T = 0, P(X)\,] \\ &= B^{\text{ATT}}[P(X)] + \text{E}[\delta|T = 1, P(X)] - \text{E}[\delta|T = 0, P(X)] \end{aligned}$$

where $\delta = \varepsilon_1 - \varepsilon_0$ is the unobserved, individual-specific gain from treatment. Heckman *et al*. (2006) refer to selection into treatment on the basis of δ as *essential heterogeneity*. When such heterogeneity exists and is correlated with treatment assignment, the magnitude of the bias of the ATU may either be larger or smaller than the corresponding bias of the ATT. If we add the following assumption:

A3. Non-negative selection into the treatment on individual-specific, unobserved gains:

$$\text{E}[\delta|T = 1, P(X)] \geqslant \text{E}[\delta|T = 0, P(X)]$$

then $|B_{\text{ATU}}[P(X)]| \geqslant |B_{\text{ATT}}[P(X)]|$ for all $P(X)$.[8]

Now consider the ATE. Utilizing the fact that $\tau_{\text{ATE}}[P(X)] = P(X)\tau_{\text{ATT}}[P(X)] + [1 - P(X)]\tau_{\text{ATU}}[P(X)]$, and rewriting $Y(1) = g_1(X) + (\delta + \varepsilon_0)$, the bias for the ATE is given by

$$\begin{aligned} B_{\text{ATE}}[P(X)] &= -\rho_{0u}\sigma_0\left\{\frac{\phi(h(X))}{\Phi(h(X))[1 - \Phi(h(X))]}\right\} + [1 - P(X)]\left\{-\rho_{\delta u}\sigma_\delta \frac{\phi(h(X))}{\Phi(h(X))[1 - \Phi(h(X))]}\right\} \\ &= -\{\rho_{0u}\sigma_0 + [1 - P(X)]\rho_{\delta u}\sigma_\delta\}\left\{\frac{\phi(h(X))}{\Phi(h(X))[1 - \Phi(h(X))]}\right\} \end{aligned} \tag{10}$$

where $\rho_{\delta u}$ is the correlation between δ and u and σ_δ is the standard deviation of δ. Note that (10) is equivalent to the expression derived in Heckman and Navarro-Lozano (2004).

[7] In this sense, the approach advocated in Black and Smith (2004) is similar to the strategy advocated in Crump *et al*. (2009) for dealing with limited overlap in the distributions of the observed variables in the treatment and control group. Crump *et al*. (2009) advocate restricting the sample to only those observations with a propensity score within a subset of the unit interval chosen to minimize the variance of the resulting average treatment effect estimator.

[8] See Heckman and Vytlacil (2005) for further discussion on this assumption.

DOI: 10.1002/jae

Equation (10) leads to three salient points. First, under A1–A3, $|B_{\text{ATU}}[P(X)]| \geqslant |B_{\text{ATE}}[P(X)]| \geqslant |B_{\text{ATT}}[P(X)]|$. Second, the value of $P(X)$ that minimizes the bias of the ATE, referred to as the *bias-minimizing propensity score* (BMPS) and denoted $P^*(X)$, is not fixed; rather, it depends on the values of $\rho_{0u}\sigma_0$ and $\rho_{\delta u}\sigma_\delta$. In particular, the bias of the ATE is minimized when $h(X)=0$ only in the case where $\rho_{\delta u}=0$ (i.e. no essential heterogeneity). Third, there are two special cases for which the bias disappears:

i. no selection on unobserved variables impacting outcomes in the untreated state: if $\rho_{0u}\sigma_0=0$, then $lim_{P(X)\to 1}B_{\text{ATE}}[P(X)]=0$;
ii. offsetting selection on ε_0 and δ: if $\rho_{0u}\sigma_0=-\rho_{\delta u}\sigma_\delta$, then $lim_{P(X)\to 0}B_{\text{ATE}}[P(X)]=0$.

4. ESTIMATION

4.1. The Minimum-Biased Approach

In light of the preceding discussion, we propose a new estimation approach when the CIA is not likely to hold, but one lacks a valid exclusion restriction. We couch our technique within the normalized inverse probability weighted (IPW) estimator of Hirano and Imbens (2001), given by

$$\hat{\tau}_{\text{IPW, ATE}} = \left[\sum_{i=1}^{N}\frac{Y_iT_i}{\hat{P}(X_i)}\Big/\sum_{i=1}^{N}\frac{T_i}{\hat{P}(X_i)}\right] - \left[\sum_{i=1}^{N}\frac{Y_i(1-T_i)}{1-\hat{P}(X_i)}\Big/\sum_{i=1}^{N}\frac{(1-T_i)}{1-\hat{P}(X_i)}\right] \tag{11}$$

Millimet and Tchernis (2009) and Busso *et al.* (2011) provide evidence of the superiority of the normalized estimator in practical settings.

Under the CIA, the IPW estimator in (11) provides an unbiased estimate of τ_{ATE}. When this assumption fails, the bias is given by (10). Rather than abandon this estimator, however, we propose to minimize the bias by estimating (11) using only observations with a propensity score in a neighborhood around the BMPS, P^*.[9] Formally, we propose the following MB estimator of the ATE:

$$\tau_{\text{MB, ATE}}[P^*] = \left[\sum_{i\in\Omega}\frac{Y_iT_i}{\hat{P}(X_i)}\Big/\sum_{i\in\Omega}\frac{T_i}{\hat{P}(X_i)}\right] - \left[\sum_{i\in\Omega}\frac{Y_i(1-T_i)}{1-\hat{P}(X_i)}\Big/\sum_{i\in\Omega}\frac{(1-T_i)}{1-\hat{P}(X_i)}\right] \tag{12}$$

where $\Omega = \left\{i\middle|\hat{P}(X_i)\in C(P^*)\right\}$ and $C(P)$ denotes a neighborhood around P. In the estimation below, we define $C(P^*)$ as $C(P^*)=\{\hat{P}(X_i)\big|\hat{P}(X_i)\in(\underline{P},\bar{P})\}$, where $\underline{P}=max\{0.02, P^*-\alpha_\theta\}$, $\bar{P}=min\{0.98, P^*+\alpha_\theta\}$, and $\alpha_\theta>0$ is the smallest value such that at least θ percent of both the treatment and control groups are contained in Ω. Below, we set $\theta=0.05$ and 0.25. For example, $\alpha_{0.05}$ is the smallest value such that 5% of the treatment group and 5% of the control group have a propensity score in the interval $(\underline{P},\bar{P})$. Thus smaller values of θ should reduce the bias at the expense of higher variance. Note that we trim observations with propensity scores above (below) 0.98 (0.02), regardless of the value of θ, to prevent any single observation from receiving too large a weight.

As defined above, the set Ω is unknown since, in general, P^* is unknown. To estimate the set Ω, we propose to estimate P^* assuming A1, A2, and functional forms for $g_0(X)$, $g_1(X)$, and $h(X)$ using Heckman's BVN selection model. Specifically, assuming $g_t(X)=X\beta_t$, $t=0,1$, and $h(X)=X\gamma$, then

$$y_i = X_i\beta_0 + X_iT_i(\beta_1-\beta_0) + \beta_{\lambda 0}(1-T_i)\left[\frac{\phi(X_i\gamma)}{1-\Phi(X_i\gamma)}\right] + \beta_{\lambda 1}T_i\left[\frac{-\phi(X_i\gamma)}{\Phi(X_i\gamma)}\right] + \eta_i \tag{13}$$

[9] For simplicity, we suppress the notation denoting the fact that the BMPS depends on X.

DOI: 10.1002/jae

where $\varphi(\cdot)/\Phi(\cdot)$ is the inverse Mills' ratio, η is a well-behaved error term, $\beta_{\lambda 0} = \rho_{0u}\sigma_0$, and $\beta_{\lambda 1} = \rho_{0u}\sigma_0 + \rho_{\delta u}\sigma_\delta$. Thus a two-step approach including OLS estimation of (13) after replacing γ with an estimate obtained from a first-stage probit model yields consistent estimates of $\rho_{0u}\sigma_0$ and $\rho_{\delta u}\sigma_\delta$.[10] With these estimates, one can use (10) to obtain an estimate of P^*.[11]

Our proposed estimator immediately raises a question: If one is willing to maintain the assumptions underlying the BVN selection model, why not just use the OLS estimates of (13) to estimate the ATE? With this approach, the estimator is given by

$$\hat{\tau}_{\text{BVN,ATE}} = \bar{X}\left(\hat{\beta}_1 - \hat{\beta}_0\right) \tag{14}$$

However, the MB estimator offers some advantages. To see this, note that data-generating processes (DGPs) fall into one of four cases depending on whether the CIA holds or does not hold and whether the assumptions of the BVN model hold or do not hold. Figure 1 summarizes the relative performance of IPW, MB, and BVN in each case. When the CIA holds and the functional form assumptions of BVN hold, all three estimators are consistent; MB is inefficient relative to IPW and it is unclear a priori how BVN and IPW compare. IPW and MB continue to be consistent (with MB being less efficient) if the CIA holds, but the BVN assumptions do not. In this case, BVN is no longer consistent. If the CIA fails to hold and the functional form assumptions are correct, then BVN is consistent while IPW and MB are not; MB will have a smaller bias than IPW. Finally, if the CIA and the functional form assumptions fail to hold, all three estimators are inconsistent and the relative biases are unknown. Thus the MB estimator offers the advantage of a robustness check on IPW: if the CIA holds, IPW is consistent and more efficient than MB, yet MB minimizes the bias when the CIA fails (and the functional form assumptions are valid). While BVN is consistent when the CIA fails and the functional form assumptions hold, it is potentially the worst of the three estimators in the other three cases.

In addition, when one is interested in the ATT rather than the ATE, MB can be utilized without estimation of (13) since P^* is known to be one-half. Moreover, it is trivial to show that $P^* = 0.5$ for the ATT and ATU under a wider class of models than joint normality. For example, consider the case where ε_0, ε_1, and u are drawn from a finite mixture of trivariate normal distributions. Because the bias of the ATT (or ATU) is minimized by minimizing the bias for each component obtaining draws from a particular trivariate normal distribution *and* the BMPS is one-half within each component, the bias of the ATT (or ATU) is minimized at $P^* = 0.5$. Furthermore, because a mixture of a sufficient number of trivariate normal distributions can approximate almost any joint distribution, this implies that joint normality is not needed to conclude that one-half is the BMPS for the ATT (or ATU). Thus, when the CIA holds, MB provides a consistent, but inefficient, alternative to IPW. However, when the CIA fails, MB has a smaller bias than IPW. In this case, our estimator – couched in the IPW estimator of the ATT – is given by

$$\hat{\tau}_{\text{MB,ATT}}[0.5] = \sum_{i\in\Omega} Y_i T_i - \left[\sum_{i\in\Omega} \frac{Y_i(1-T_i)\hat{P}(X_i)}{1-\hat{P}(X_i)} \Big/ \sum_{i\in\Omega} \frac{(1-T_i)\hat{P}(X_i)}{1-\hat{P}(X_i)}\right] \tag{15}$$

[10] Note that throughout the paper, when we refer to the BVN estimator in practice, we are referring to estimates obtained using this two-step estimation procedure, as in the simulations in Heckman *et al*. (1999). We utilize the two-step estimator, rather than Full Information Maximum Likelihood (FIML), for several reasons. First, the FIML estimator may have convergence problems in the absence of an exclusion restriction. Second, there is evidence that the two-step estimator is more robust to deviations from the assumed error distribution. Third, the two-step estimator remains consistent even in the face of classical measurement error in the outcome; FIML does not. These final two advantages are particularly important in our application. See Puhani (2000), Leung and Yu (2000), and Bushway *et al*. (2007) for further discussion.

[11] To estimate P^*, we conduct a grid search over 1000 equally spaced values of $h(\cdot)$ from -5 to 5. If P^* is above 0.98, we truncate it to 0.98; if P^* is below 0.02, we truncate it to 0.02.

DOI: 10.1002/jae

		BVN functional form correctly specified	
		YES	**NO**
CIA holds	**YES**	IPW, MB, BVN, BC consistent	IPW, MB consistent BVN, BC inconsistent
	NO	IPW inconsistent MB inconsistent, but less than IPW BVN, BC consistent	IPW, MB, BVN, BC inconsistent

Figure 1. Performance of IPW, MB, and BVN estimators of the ATE under different scenarios

For comparison, under the functional form assumptions for $g_0(X)$, $g_1(X)$, and $h(X)$ discussed above, OLS estimation of (13) also produces a consistent estimate of the ATT. This estimator is given by

$$\hat{\tau}_{\text{BVN, ATT}} = \bar{X}_1\left(\hat{\beta}_1 - \hat{\beta}_0\right) + \hat{\beta}_{\lambda 1}\overline{\left[\frac{-\phi(X_i\hat{\gamma})}{\Phi(X_i\hat{\gamma})}\right]}_1 \tag{16}$$

where $\bar{X}_1$ and $\overline{\left[\frac{-\phi(X_i\hat{\gamma})}{\Phi(X_i\hat{\gamma})}\right]}_1$ are the sample means of X and the selection correction term, respectively, in the treatment group.

4.2. The Bias-Corrected Approach

While (12) and (15) provide a minimum-biased estimator of the ATE and ATT, respectively, estimation of the error correlation structure using (13) immediately gives rise to the possibility of a bias-corrected version of each estimator. Specifically, after estimating (13), estimates of the *bias* of the MB estimator of the ATE and ATT are given by

$$\widehat{B_{\text{ATE}}[P^*]} = -[\widehat{\rho_{0u}\sigma_0} + (1-P^*)\widehat{\rho_{\delta u}\sigma_\delta}]\left[\frac{\phi\left(\Phi^{-1}(P^*)\right)}{P^*(1-P^*)}\right]$$

$$\widehat{B_{\text{ATT}}[0.5]} \approx -1.6 * \widehat{\rho_{0u}\sigma_0}\,.$$

The minimum bias bias-corrected estimator, denoted MB-BC, for the ATE is then given by

$$\hat{\tau}_{\text{MB-BC,ATE}}[P^*] = \hat{\tau}_{\text{MB,ATE}}[P^*] - \widehat{B_{\text{ATE}}[P^*]} \tag{17}$$

where the corresponding estimators for the ATT and ATU follow.

MB-BC has the same interpretation as the MB estimator; specifically, with heterogeneous treatment effects, the parameter being estimated has changed. However, one can avoid this by considering a bias-corrected estimator of the *unconditional* average treatment effect. To proceed, estimate MB-BC conditional on the propensity score, $P(X)$, and then estimate the (unconditional) average treatment effect by taking the expectation of this over the distribution of X in the population (or sub-population of treated). Formally, our bias-corrected estimators, denoted by BC, for the ATE is given by

DOI: 10.1002/jae

$$\hat{\tau}_{\text{BC,ATE}} = \hat{\tau}_{\text{IPW,ATE}} - \sum_i \widehat{B_{\text{ATE}}\left[\hat{P}(X_i)\right]} \tag{18}$$

where again the corresponding estimators for the ATT and ATU follow.

Before continuing, note that while there appears to be little difference between BC and BVN, since BC utilizes BVN to estimate the correlation parameters, BVN relies exclusively on the functional form specification of the outcome equation. In contrast, BC does not, once the correlation parameters have been estimated. As demonstrated below, this leads to improvement in practice.

4.3. Deviations from Normality

The assumption of joint normality may be unrealistic in many applications. To relax this assumption, we follow Lee (1984), who utilizes the fact that under certain assumptions the (unknown) joint density of ε_j^* and u, $f_j\left(\varepsilon_j^*, u\right)$, $j=0,1$, may be written as a bivariate Edgeworth series of distributions where $\varepsilon_j = \sigma_j \varepsilon_j^*$ and ε_j^* has unit variance. Formally:

$$f_j\left(\varepsilon_j^*, u\right) = \phi_2\left(\varepsilon_j^*, u\right) + \sum_{r+s\geq 3} (-1)^{r+s} A_{rs} \frac{1}{r!s!} \frac{\partial^{r+s}\phi_2\left(\varepsilon_j^*, u\right)}{\partial u^r \partial \varepsilon_j^{*s}}, \quad j=0,1 \tag{19}$$

where φ_2 is the bivariate standard normal density and A_{rs} are the cumulants (or semi-invariants) of ε_j^* and u. The cumulants are functions of moments of the distribution of ε_j^* and u; cumulants for $r+s>2$ are zero under bivariate normality and non-zero otherwise (Mardia, 1970; Lee, 1984). As in Lee (1984), we consider the case where $r+s \in \{3,4\}$, referred to as a Type AA surface in Mardia (1970), and $u \sim N(0,1)$. Thus, for simplicity, cumulants for $r+s>4$ are set to zero.

In light of this, the bias of the ATT is now given by

$$\begin{aligned} B_{\text{ATT}}[P(X)] &= \text{E}[\varepsilon_0 | T=1, P(X)] - \text{E}[\varepsilon_0 | T=0, P(X)\,] \\ &= -\left\{ \rho_{0u}\sigma_0 + \kappa_{12}\sigma_0 \frac{h(X)}{2} + \kappa_{13}\sigma_0 \frac{\left[h(X)^2 - 1\right]}{6} \right\} \left\{ \frac{\phi(h(X))}{\Phi(h(X))[1-\Phi(h(X))]} \right\} \end{aligned} \tag{20}$$

where κ_{ij} are the cumulants. For the ATU, we have by symmetry

$$\begin{aligned} B_{\text{ATU}}[P(X)] &= \text{E}[\varepsilon_1 | T=1, P(X)] - \text{E}[\varepsilon_1 | T=0, P(X)\,] \\ &= -\left\{ \rho_{1u}\sigma_1 + \kappa'_{12}\sigma_1 \frac{h(X)}{2} + \kappa'_{13}\sigma_1 \frac{\left[h(X)^2 - 1\right]}{6} \right\} \left\{ \frac{\phi(h(X))}{\Phi(h(X))[1-\Phi(h(X))]} \right\} \end{aligned} \tag{21}$$

For the ATE, the bias is

$$B_{\text{ATE}}[P(X)] = -\left\{ \begin{array}{c} \rho_{0u}\sigma_0 + \kappa_{12}\sigma_0 \dfrac{h(X)}{2} + \kappa_{13}\sigma_0 \dfrac{\left[h(X)^2 - 1\right]}{6} \\ +[1-P(X)]\left\{ \rho_{\delta u}\sigma_\delta + \kappa''_{12}\sigma_\delta \dfrac{h(X)}{2} + \kappa''_{13}\sigma_\delta \dfrac{\left[h(X)^2 - 1\right]}{6} \right\} \end{array} \right\} \left\{ \frac{\phi(h(X))}{\Phi(h(X))[1-\Phi(h(X))]} \right\} \tag{22}$$

DOI: 10.1002/jae

Minimizing the bias of the *three* parameters – as P^* is no longer always one-half for the ATT or ATU – requires knowledge or estimates of several additional parameters. However, these are estimable from an altered version of the BVN selection model under A1:

$$\begin{aligned} y_i = {} & X_i\beta_0 + X_iT_i(\beta_1 - \beta_0) + \beta_{\lambda 01}(1 - T_i)\left[\frac{\phi(X_i\gamma)}{1 - \Phi(X_i\gamma)}\right] + \beta_{\lambda 02}(1 - T_i)\left[\frac{X_i\gamma}{2}\frac{\phi(X_i\gamma)}{1 - \Phi(X_i\gamma)}\right] \\ & + \beta_{\lambda 03}(1 - T_i)\left[\frac{(X_i\gamma)^2 - 1}{6}\frac{\phi(X_i\gamma)}{1 - \Phi(X_i\gamma)}\right] + \beta_{\lambda 11}T_i\left[\frac{-\phi(X_i\gamma)}{\Phi(X_i\gamma)}\right] + \beta_{\lambda 12}T_i\left[\frac{-X_i\gamma}{2}\frac{\phi(X_i\gamma)}{\Phi(X_i\gamma)}\right] \\ & + \beta_{\lambda 13}T_i\left[\frac{1 - (X_i\gamma)^2}{6}\frac{\phi(X_i\gamma)}{\Phi(X_i\gamma)}\right] + \eta_i \end{aligned} \tag{23}$$

where

$$\begin{aligned} \beta_{\lambda 01} &= \rho_{0u}\sigma_0; \qquad & \beta_{\lambda 11} &= \rho_{0u}\sigma_0 + \rho_{\delta u}\sigma_\delta \\ \beta_{\lambda 02} &= \kappa_{12}\sigma_0; \qquad & \beta_{\lambda 12} &= \kappa_{12}\sigma_0 + \kappa'_{12}\sigma_\delta \\ \beta_{\lambda 03} &= \kappa_{13}\sigma_0; \qquad & \beta_{\lambda 13} &= \kappa_{13}\sigma_0 + \kappa'_{13}\sigma_\delta \end{aligned}$$

Upon estimation of ((23)), the BMPS is found by minimizing the bias in ((22)). The remainder of the estimation algorithm is unchanged. We denote the resulting estimators MB-EE, MB-BC-EE, and BC-EE (where EE denotes use of the Edgeworth expansion).

4.4. Other Estimators

For comparison, we also consider two estimators from the recent literature: an alternative CF approach and the IV estimator of Klein and Vella (2009).

4.4.1. Control Function Approach

The CF approach is outlined nicely in Heckman *et al*. (1999) and Navarro (2008). The idea, of which the BVN model is a special case, is to devise a function such that treatment assignment is no longer correlated with the error term in the outcome equation upon its inclusion. To proceed, rewrite outcomes as

$$Y_i(t) = \alpha_t + g_t(X_i) + \mathrm{E}[\varepsilon_t | X_i, T_i = t] + \eta_{it}, \quad t = 0, 1 \tag{24}$$

where we now explicitly include an intercept, α_t, and X contains only the vector of covariates. The terms $\mathrm{E}[\varepsilon_0|X, T=0] = \mathrm{E}[\varepsilon_0|u > \Phi^{-1}[P(X)]]$ and $\mathrm{E}[\varepsilon_1|X, T=1] = \mathrm{E}[\varepsilon_1|u < \Phi^{-1}[P(X)]]$ satisfy the requirements of a control function as η_{it} is now a well-behaved error term.

Approximating $\mathrm{E}[\varepsilon_t|X, T=t]$ with a polynomial (inclusive of an intercept term) in *P(X)* yields

$$Y_i(t) = (\alpha_t + \pi_{t0}) + g_t(X_i) + \sum\nolimits_{s=1}^{S} \pi_{ts}P(X_i)^s + \eta_{it}, t = 0, 1 \tag{25}$$

where S is the order of the polynomial. The following equation is then estimable via OLS:

$$\begin{aligned} y_i &= (\alpha_0 + \pi_{00})(1 - T_i) + (\alpha_1 + \pi_{10})T_i + X_i\beta_0 + X_iT_i(\beta_1 - \beta_0) \\ &\quad + \textstyle\sum_{s=1}^{S}\pi_{0s}(1 - T_i)P(X_i)^s + \sum_{s=1}^{S}\pi_{1s}T_iP(X)^s + \eta_i \\ &\equiv \tilde{\alpha}_0(1 - T)_i + \tilde{\alpha}_1T_i + X_i\beta_0 + X_iT_i(\beta_1 - \beta_0) \\ &\quad + \textstyle\sum_{s=1}^{S}\pi_{0s}(1 - T_i)P(X_i)^s + \sum_{s=1}^{S}\pi_{1s}T_iP(X)^s + \eta_i \end{aligned} \tag{26}$$

In the absence of an exclusion restriction, identification rests on the nonlinearity of the propensity score.

DOI: 10.1002/jae

However, estimation of any of the average treatment effect parameters requires decomposing the intercepts into α_t and π_{t0}, $t=0,1$. This is achieved, assuming *P(X)* has full support, by noting that

$$\lim_{P\to 0} \mathrm{E}[\varepsilon_0|X,T=0] = \mathrm{E}[\varepsilon_0|u > \Phi^{-1}[P(X)]] = 0$$
$$\lim_{P\to 1} \mathrm{E}\left[\varepsilon_1|X,T=1\right] = \mathrm{E}\left[\varepsilon_1\middle|u < \Phi^{-1}[P(X)]\right] = 0$$

In other words, the selection problem disappears as the propensity score moves to either extreme since the unobserved variables, u, no longer influence treatment choice. As such, the control function becomes zero and the intercepts from the potential outcome equations are identified. Specifically:

$$\begin{aligned} \pi_{00} = 0 \quad &\Leftrightarrow \alpha_0 = \tilde{\alpha}_0 \\ \pi_{10} = -\textstyle\sum_{s=1}^{S}\pi_{1s} \quad &\Leftrightarrow \alpha_1 = \tilde{\alpha}_1 + \textstyle\sum_{s=1}^{S}\pi_{1s} \end{aligned}$$

Finally, the ATE and ATT are given by

$$\hat{\tau}_{\mathrm{CF,ATE}} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{X}\left(\hat{\beta}_1 - \hat{\beta}_0\right) \tag{27}$$

$$\hat{\tau}_{\mathrm{CF,ATT}} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{X}_1\left(\hat{\beta}_1 - \hat{\beta}_0\right) + \mathrm{E}\left[\varepsilon_1 - \varepsilon_0\middle|\widehat{T_i = 1}\right] \tag{28}$$

where

$$\mathrm{E}[\widehat{\varepsilon_0|T_i} = 1] = -\left[\textstyle\sum_{s=1}^{S}\hat{\pi}_{0s}\overline{P(X)_0^s}\right]\left(\dfrac{1-\overline{P(X)}}{\overline{P(X)}}\right)$$
$$\mathrm{E}[\widehat{\varepsilon_1|T_i} = 1] = -\textstyle\sum_{s=1}^{S}\hat{\pi}_{1s} + \sum_{s=1}^{S}\hat{\pi}_{1s}\overline{P(X)_1^s}$$

and the first equality follows from the fact that $\mathrm{E}[\varepsilon_0] = \mathrm{E}[\varepsilon_0|T=0](1-\overline{P(X)}) + \mathrm{E}[\varepsilon_0|T=1]\overline{P(X)} = 0$, $\overline{P(X)}$ is the mean propensity score, and $\overline{P(X)}_t$, $t=0,1$, is the mean propensity score in group t.

4.4.2. Klein and Vella (2009) Estimator

In contrast to the CF approach, which overcomes the identification problem using observations at the extremes of the support of the propensity score, KV propose an IV estimator that arguably utilizes more information from the middle of the distribution. Our parametric implementation relies on a similar functional form assumption to the BVN estimator in the absence of heteroskedasticity, but effectively induces a valid exclusion restriction in the presence of heteroskedasticity.[12] To proceed, suppose that latent treatment assignment is now given by

$$T^* = X\gamma - u^*$$

where $u^* = S(X)u$ and u is drawn from a standard normal density. In this case, the probability of receiving the treatment conditional on X is given by

$$\Pr\left(T = 1|X\right) = \Phi\left(\frac{X}{S(X)}\gamma\right) \tag{29}$$

Assuming $S(X) = \exp(X\delta)$, the parameters of (29) are estimable by maximum likelihood (ML), with the log-likelihood function given by

[12] Note that Klein and Vella (2009) also consider a semiparametric IV estimator. Our analysis should not be viewed as an evaluation of this estimator.

DOI: 10.1002/jae

$$ln\mathcal{L} = \sum_i \left[ln\Phi\left(\frac{X\gamma}{\exp(X\delta)}\right)\right]^{T_i} \left\{ ln\left[1 - \Phi\left(\frac{X\gamma}{\exp(X\delta)}\right)\right]\right\}^{1-T_i} \tag{30}$$

where the element of δ corresponding to the intercept is normalized to zero for identification.

The ML estimates are then used to obtain the predicted probability of treatment, $P(\hat{X})$, which may be used as an instrument for T in equation ((13)) excluding the selection correction terms.[13] Note that even if $S(X) = 1$, $P(\hat{X})$ remains a valid instrument since it is nonlinear in X. However, since the nonlinearity arises mostly in the tails, identification typically relies on extreme observations (as in the BVN and CF approaches).[14] On the other hand, if $S(X) \neq 1$, then the KV approach effectively induces a valid exclusion restriction as $Z \equiv X/S(X)$ is frequently linearly independent of X (Klein and Vella, 2009).

5. MONTE CARLO STUDY

5.1. Simulated Data

5.1.1. Setup

To assess the performance of the various estimators, we simulate data using two primary designs for the DGP. The first design imposes the constant treatment effect setup (i.e. $\tau_i = \tau$ for all i). The second design gives rise to heterogeneous treatment effects, where the heterogeneity is due to essential heterogeneity (i.e. τ_i varies across i, but this variation arises due to differences in $\varepsilon_{1i} - \varepsilon_{0i}$). In part motivated by our application, we do not consider the case where the treatment effect is heterogeneous on the basis of observed variables; the vectors β_0 and β_1 are identical except for the intercept.

We simulate 250 datasets, each with 5000 observations, containing

$$\begin{aligned} h(X) &= 0.5(x_1 - x_2) + 0.5\left(x_1^2 - x_2^2\right) + 2x_1x_2 \\ T^* &= 0.5 + h(X) - u \\ T &= \begin{cases} 1 & \text{if } T^* > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where $x_1, x_2 \overset{\text{i.i.d.}}{\sim} U(-1, 1)$ and u is defined below.[15]

In the constant treatment effect setup, $\varepsilon = \varepsilon_0 = \varepsilon_1$ and potential outcomes are given by

$$\begin{aligned} Y(0) &= g_0(X) + \varepsilon_0 = h(X) + \varepsilon \\ Y(1) &= g_1(X) + \varepsilon_0 = 1 + h(X) + \varepsilon \end{aligned}$$

which implies that $\tau_i = 1$ for all i and $\tau_{\text{ATE}} = \tau_{\text{ATT}} = \tau_{\text{ATU}} = 1$. In the second design, where treatment effects vary due to essential heterogeneity, potential outcomes are given by

$$\begin{aligned} Y(0) &= g_0(X) + \varepsilon_0 = h(X) + \varepsilon_0 \\ Y(1) &= g_1(X) + \varepsilon_1 = 1 + h(X) + \varepsilon_1 \end{aligned}$$

which implies that $\tau_i = 1 + \varepsilon_{1i} - \varepsilon_{0i} = 1 + \delta_i$ and $\tau_{\text{ATE}} = 1 + \text{E}[\delta_i]$, $\tau_{\text{ATT}} = 1 + \text{E}[\delta_i | T_i = 1]$, and $\tau_{\text{ATU}} = 1 + \text{E}[\delta_i | T_i = 0]$.

[13] Interactions between $P(\hat{X})$ and X may also serve as instruments given inclusion of interactions between T and X in (13).

[14] Mroz (1999) analyzes this case and finds that while an IV strategy using the predicted probability obtained from a homoskedastic probit performs reasonably well in terms of bias, it does very poorly in terms of mean squared error.

[15] We also simulated 50 datasets with 250,000 observations each to assess large sample performance. Results are available in Appendix B (supporting information).

DOI: 10.1002/jae

Within each of the two DGP designs, we consider four general error structures. First, the errors are multivariate normal, $\varepsilon_0, \varepsilon_1, u \sim N_3(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho_{01} & \rho_{0u} \\ & 1 & \rho_{1u} \\ & & 1 \end{bmatrix}$$

and $\rho_{01} = 1$ in the common effect setup, implying $\rho_{\delta u} = 0$ and $\rho_{0u} = \rho_{1u}$. In the heterogeneous effect setup, $\rho_{01} = 0.5$ implying $\rho_{\delta u} = \rho_{1u} - \rho_{0u}$. Second, the errors are drawn from an asymmetric and non-normal multivariate distribution with the desired correlation matrix using the method in Headrick and Sawilowsky (1999). Specifically, ε_0 and ε_1 are mean zero with unit variance and skewness and kurtosis close to a χ^2_1 distribution (skewness is roughly $\sqrt{8}$ and kurtosis is around 15); u continues to have a standard normal distribution. The third and fourth error structures are identical to the preceding cases, except that we introduce heteroskedasticity in the treatment assignment equation: $u^* = (1 + 0.45(x_1 + x_2))u$ and $T^* = 0.5 + h(X) - u^*$.

In the constant treatment effect design, we consider three sets of values for Σ: $\rho_{0u} = 0$, -0.25, and -0.50, where the CIA holds in the first case and larger values of ρ_{0u} (in absolute value) correspond to greater selection on unobserved variables.[16] In the heterogeneous treatment effect design, we consider three sets of values for Σ: $\rho_{0u} = \rho_{\delta u} = 0$, $\rho_{0u} = -0.20$ and $\rho_{\delta u} = -0.10$, and $\rho_{0u} = -0.40$ and $\rho_{\delta u} = -0.10$, where the CIA again holds in the first case and larger values of ρ_{0u} (in absolute value) correspond to greater selection on unobserved variables. Figure 2 summarizes the various DGPs.

Two final comments are warranted. First, we assume in all cases that the researcher knows that the treatment effect does not vary on the basis of observed variables. As such, we exclude interactions between T and the set of covariates in X in (13) and (23). In addition, at this juncture (but not later), we assume the researcher knows the correct specification for the covariates; hence *h(x)* is correctly specified, as is the form of the heteroskedasticity, *S(X)*. Second, as identification in the CF approach is aided by the propensity score having full support, we verify that this holds within each of the DGPs.[17]

5.1.2. Results

Correct specification. Results for the ATE and ATT from the first DGP design – constant treatment effect – are presented in Table I. Figures represent root mean squared errors (RMSE).[18] Panel A (panel B) displays the results when the errors are normal (asymmetric and non-normal). The left (right) set of columns corresponds to the case where the error in the treatment assignment is homoskedastic (heteroskedastic). Within each column, the shaded number represents the smallest RMSE.

Columns (1) and (4) correspond to the simplest case: the treatment effect is constant and the CIA holds. Three findings emerge. First, consonant with our expectations, IPW performs best for the ATE with or without homoskedastic errors (column (1)). Our MB estimator with a large radius ($\theta = 0.25$), while outperforming the remaining estimators, has an RMSE roughly one-third higher than IPW. The additional inefficiency from using MB-EE is not nearly as severe as using a smaller radius ($\theta = 0.05$). Second, the IPW and MB estimator with a large radius ($\theta = 0.25$) achieve almost identical performance for the ATT when errors are homoskedastic; our MB estimator with a large radius

[16] Given the DGP for T^*, $\rho_{0u} < 0$ indicates positive selection (i.e. unobserved variables associated with better outcomes are also associated with a higher probability of receiving the treatment).

[17] Formally this is not required since we use a cubic polynomial of the propensity score ($S = 3$). Nonetheless, at least within the population defined by the DGPs (but not necessarily any small sample), the propensity has full support.

[18] RMSE is given by $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\tau} - \tau)^2}$, where M is the number of simulations and τ is the population value of the parameter being estimated.

DOI: 10.1002/jae

	Homoskedastic errors		Heteroskedastic errors	
	Constant treatment effect ($\rho_{01} = 0$)	**Heterogeneous treatment effects ($\rho_{01} = 0.5$)**	**Constant treatment effect ($\rho_{01} = 0$)**	**Heterogeneous treatment effects ($\rho_{01} = 0.5$)**
Normal distribution	$\rho_0\sigma_0 = 0$ $\rho_0\sigma_0 = -0.25$ $\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$, $\rho_\delta\sigma_\delta = 0$ $\rho_0\sigma_0 = -0.2$, $\rho_\delta\sigma_\delta = -0.1$ $\rho_0\sigma_0 = -0.4$, $\rho_\delta\sigma_\delta = -0.1$	$\rho_0\sigma_0 = 0$ $\rho_0\sigma_0 = -0.25$ $\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$, $\rho_\delta\sigma_\delta = 0$ $\rho_0\sigma_0 = -0.2$, $\rho_\delta\sigma_\delta = -0.1$ $\rho_0\sigma_0 = -0.4$, $\rho_\delta\sigma_\delta = -0.1$
Asymmetric non-normal distribution	$\rho_0\sigma_0 = 0$ $\rho_0\sigma_0 = -0.25$ $\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$, $\rho_\delta\sigma_\delta = 0$ $\rho_0\sigma_0 = -0.2$, $\rho_\delta\sigma_\delta = -0.1$ $\rho_0\sigma_0 = -0.4$, $\rho_\delta\sigma_\delta = -0.1$	$\rho_0\sigma_0 = 0$ $\rho_0\sigma_0 = -0.25$ $\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$, $\rho_\delta\sigma_\delta = 0$ $\rho_0\sigma_0 = -0.2$, $\rho_\delta\sigma_\delta = -0.1$ $\rho_0\sigma_0 = -0.4$, $\rho_\delta\sigma_\delta = -0.1$

Figure 2. Monte Carlo designs. Within each DGP, we estimate four specifications depending on whether the observables included in the estimation procedure are under-, correctly, or over-specified. See text for further details

($\theta = 0.25$) attains an RMSE about 25% larger than IPW with heteroskedastic errors (column (4)). Thus the efficiency loss associated with MB is diminished when P^* does not have to be estimated.

Third, even when normality holds, MB-BC, BC, KV, BVN, and CF have higher RMSEs than IPW and MB for both the ATE and ATT. This is not surprising given the efficiency loss from not utilizing the CIA when this holds in the DGP. Finally, MB-BC, BC, BVN, and MB outperform their counterparts (i.e. the -EE estimators) allowing for non-normality even when the errors are non-normal. Thus the flexibility offered by these estimators is more than offset by the imprecision resulting from the increased parametrization of the model.

Columns (2), (3), (5), and (6) introduce selection on unobserved variables into the DGP, with (3) and (6) containing the largest amount of selection on unobserved variables. Four salient conclusions emerge. First, MB yields improvements over IPW in most cases. In particular, MB outperforms IPW in nearly every case when the errors are normal, as well as when the errors are non-normal and the focus is on the ATE. Moreover, the smaller radius ($\theta = 0.05$) is often preferred. When the errors are non-normal and the focus is on the ATT, MB-EE with a large radius ($\theta = 0.25$) tends to outperform MB and IPW. Thus the results confirm Figure 1 as MB is less efficient than IPW if the CIA holds, but offers an advantage if it does not. Moreover, the advantage of MB over IPW when the CIA fails also extends in practice to (at least some) situations where the functional form assumptions of the BVN model do not hold.

Second, when the errors are homoskedastic, BVN, MB-BC with a large radius ($\theta = 0.25$), and BC perform comparably and best overall. The only exception is with severe selection on unobserved variables and non-normal errors, where KV performs best (panel B, columns (3) and (6)). CF and the remaining estimators based on non-normality perform noticeably worse. Third, when the errors are heteroskedastic, KV performs best in all cases but one (panel A, column (5)).

Fourth and finally, while KV, BVN, MB-BC with a large radius ($\theta = 0.25$), and BC outperform the rest of the estimators when the CIA fails, it is vital to emphasize that none compares favorably to traditional IV (although they potentially estimate a parameter that may be of greater interest than that estimated by IV). While not shown in the table, if we alter the DGP design to include a strong, valid instrument in the treatment assignment equation, the RMSE of two-stage least squares (TSLS) is approximately 0.06 in each case – about 20% higher than IPW when the CIA holds.[19]

[19] Treatment assignment is simulated as $T^* = 0.5 + h(X) + z - u$, where $z \sim N(0, 1)$.

DOI: 10.1002/jae

Table I. Monte Carlo results: estimates in the common effect model ($\tau_i = 1$)

	Homoskedastic error in treatment equation						Heteroskedastic error in treatment equation					
	ATE			ATT			ATE			ATT		
	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$
	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
(A) *Normally distributed errors*												
τ_{IPW}	**0.045**	0.450	0.892	0.053	0.457	0.901	**0.047**	0.422	0.849	**0.049**	0.424	0.850
$\tau_{MB,\ 0.05}$	0.111	0.402	0.783	0.103	0.410	0.805	0.130	0.421	0.807	0.110	0.433	0.820
$\tau_{MB,\ 0.25}$	0.062	0.405	0.806	**0.051**	0.409	0.805	0.063	0.409	0.821	0.063	0.429	0.824
$\tau_{MB,\ EE,\ 0.05}$	0.128	0.470	0.917	0.125	0.476	0.919	0.156	0.385	0.805	0.172	0.415	0.822
$\tau_{MB,\ EE,\ 0.25}$	0.066	0.441	0.869	0.065	0.452	0.875	0.072	0.417	0.831	0.067	0.420	0.840
τ_{KV}	0.365	0.358	0.359	0.365	0.358	0.359	0.250	**0.254**	**0.249**	0.250	0.254	**0.249**
τ_{CF}	1.049	1.017	0.979	0.980	0.954	0.920	0.550	0.546	0.551	0.524	0.519	0.528
τ_{BVN}	0.285	**0.280**	**0.275**	0.210	**0.226**	0.283	0.256	0.276	0.281	0.193	**0.252**	0.357
τ_{BC}	0.290	0.287	0.282	0.287	0.281	0.275	0.270	0.284	0.277	0.277	0.291	0.276
$\tau_{MB\text{-}BC,\ 0.05}$	0.291	0.300	0.286	0.304	0.295	0.285	0.269	0.286	0.298	0.265	0.303	0.308
$\tau_{MB\text{-}BC,\ 0.25}$	0.279	0.285	0.279	0.279	0.272	**0.267**	0.249	0.273	0.295	0.261	0.288	0.293
$\tau_{BVN,\ EE}$	1.613	1.587	1.564	1.055	1.054	1.073	0.877	0.879	0.901	0.578	0.578	0.617
$\tau_{BC,\ EE}$	1.619	1.595	1.570	1.447	1.442	1.426	0.898	0.903	0.936	0.824	0.817	0.841
$\tau_{MB\text{-}BC,\ EE,\ 0.05}$	1.434	1.411	1.493	1.396	1.373	1.359	0.793	0.829	0.904	0.768	0.751	0.797
$\tau_{MB\text{-}BC,\ EE,\ 0.25}$	1.404	1.373	1.465	1.366	1.341	1.325	0.776	0.800	0.900	0.733	0.731	0.785
(B) *Asymmetric, non-normally distributed errors*												
τ_{IPW}	**0.049**	0.379	0.625	**0.055**	0.347	0.487	**0.049**	0.371	0.642	**0.051**	0.343	0.515
$\tau_{MB,\ 0.05}$	0.118	0.346	0.436	0.103	0.417	0.782	0.129	0.323	0.379	0.105	0.434	0.800
$\tau_{MB,\ 0.25}$	0.064	0.348	0.467	0.056	0.392	0.720	0.065	0.347	0.467	0.063	0.415	0.738
$\tau_{MB,\ EE,\ 0.05}$	0.142	0.408	0.775	0.128	0.334	0.497	0.174	0.351	0.606	0.161	0.327	0.447
$\tau_{MB,\ EE,\ 0.25}$	0.074	0.370	0.638	0.072	0.337	0.486	0.073	0.356	0.565	0.070	0.330	0.484
τ_{KV}	0.372	0.392	**0.403**	0.372	0.392	**0.403**	0.286	**0.273**	**0.267**	0.286	**0.273**	**0.267**
τ_{CF}	1.063	1.053	1.076	1.001	0.995	0.989	0.601	0.571	0.544	0.563	0.533	0.519

DOI: 10.1002/jae

Table I. (*Continued*)

	Homoskedastic error in treatment equation						Heteroskedastic error in treatment equation					
	ATE			ATT			ATE			ATT		
	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.25$	$\rho_0\sigma_0 = -0.50$
	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$	$\rho_{01} = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
τ_{BVN}	0.290	0.323	0.420	0.215	**0.316**	0.616	0.254	0.297	0.398	0.191	0.309	0.603
τ_{BC}	0.294	0.325	0.417	0.295	0.344	0.638	0.265	0.300	0.414	0.270	0.330	0.643
$\tau_{MB\text{-}BC,\ 0.05}$	0.309	0.323	0.425	0.303	0.391	0.899	0.279	0.292	0.388	0.273	0.396	0.895
$\tau_{MB\text{-}BC,\ 0.25}$	0.285	**0.319**	0.421	0.286	0.369	0.838	0.250	0.288	0.419	0.257	0.377	0.837
$\tau_{BVN,\ EE}$	1.623	1.651	1.771	1.075	1.077	1.153	0.915	0.921	0.928	0.606	0.583	0.616
$\tau_{BC,\ EE}$	1.634	1.661	1.781	1.483	1.485	1.576	0.936	0.950	0.956	0.859	0.829	0.840
$\tau_{MB\text{-}BC,\ EE,\ 0.05}$	1.371	1.501	1.524	1.379	1.400	1.502	0.768	0.928	1.041	0.864	0.751	0.845
$\tau_{MB\text{-}BC,\ EE,\ 0.25}$	1.335	1.479	1.501	1.361	1.387	1.434	0.741	0.898	1.010	0.842	0.745	0.764

Note: Numbers reflect the root mean squared error based on 250 simulated datasets with 5000 observations. IPW, inverse propensity score weighting estimator; MB, minimum-biased estimator using a cut-off level (α) chosen to retain 5% or 25% of the treatment and control groups; KV, Klein and Vella (2009) estimator; BVN, Heckman bivariate normal selection model; MB-BC, bias-corrected estimator using the same cut-off levels (α). EE implies estimator relies on the Edgeworth expansion formula. Bold font indicates best performance within each column. See text for further details.

DOI: 10.1002/jae

Next, we turn to the results for the ATE and ATT from the second DGP design. Now, the treatment effect varies across observations due to essential heterogeneity. Results are presented in Table II.

Columns (1) and (4) again correspond to the case where the CIA holds. The qualitative findings are unchanged from the corresponding columns in Table I. Specifically, IPW continues to perform best for both ATE and ATT. MB with a large radius ($\theta = 0.25$) outperforms the remainder of the estimators, but performs worse than IPW due to a loss in efficiency. Moreover, as before, the efficiency loss – relative to MB with a large radius ($\theta = 0.25$) – from using MB-EE with a large radius is less severe than the loss from using a small radius ($\theta = 0.05$).

Similarly, the majority of the findings from Table I continue to hold when the CIA no longer holds (columns (2), (3), (5), and (6)). First and foremost, MB, particularly with a small radius ($\theta = 0.05$), continues to outperform IPW in nearly all cases except when the errors are non-normal and the focus is on the ATT. In this case, MB-EE with a large radius ($\theta = 0.25$) perhaps modestly outperforms IPW. Second, when the errors are homoskedastic, BVN, MB-BC with a large radius ($\theta = 0.25$), and BC perform comparably and best overall. Third, when the errors are heteroskedastic, KV performs best when the errors are normal, but not necessarily when the errors are non-normal (panel B), unlike in the constant treatment effect case. In the non-normal case, BVN performs best when the selection on unobserved variables is not too severe (panel II, columns (2) and (5)). Fourth, CF and the remaining estimators based on non-normality perform much worse than the other estimators. Finally, IV continues to perform substantially better, yielding an RMSE around 0.07–0.08.[20]

While the Monte Carlo results are only illustrative, if we take a step back two conclusions emerge when the functional forms are correctly specified; see Figure 3 for a summary. First, when the CIA holds, IPW is preferred, but MB and MB-EE with a large radius outperform the remaining estimators. As expected, both are inefficient relative to IPW. However, MB and MB-EE with a large radius are more robust to a failure of the CIA than IPW. Thus MB and MB-EE succeed in providing researchers with an alternative estimator that can serve as a useful robustness check to the usual IPW estimator when the CIA is questionable. Second, if the researcher is confident that the CIA does not hold, then KV, BVN, MB-BC, and BC are preferred among the estimators considered here (but may perform worse than traditional IV; see also Heckman *et al*., 1999).[21] In particular, BVN tends to be preferred when errors are homoskedastic, as well as when the errors are heteroskedastic, non-normal, and the treatment effect is heterogeneous. Otherwise, KV is preferred.

Under- and over-specifying the model. The preceding results assume that the proper specification for $h(\mathrm{X})$ and $S(X)$ is known. Since this is rarely true in practice, we relax this assumption. The results are relegated to Appendix A (supporting information). The DGP designs correspond exactly to Tables I and II. However, we now consider three additional specifications for $h(\mathrm{X})$ and $S(\mathrm{X})$ in the estimation. Specification (1) includes only linear terms for x_1 and x_2 (hence x_1^2, x_2^2, and $x_1 x_2$ are omitted) in the probit and outcome models. Specification (2) is the correct specification reported in Tables I and II. Specification (3) is identical to specification (2) but adds cubic terms for x_1 and x_2 as well as the complete set of interaction terms between the linear and quadratic terms. Specification (4) is identical to specification (3) but incorporates an additional, completely irrelevant regressor, $x_3 \sim U(-1, 1)$, and includes x_3, x_3^2 , x_3^3, and the full set of interaction terms in the set of covariates. In specifications (1)–(3), we model $S(X) = S(x_1, x_2)$; in specification (4), we model $S(X) = S(x_1, x_2, x_3)$.

Specification (1) is *under-specified*; Specifications (3) and (4) are *over-specified*. While the under-specified case is arguably the most important in practice, we also examine the over-specified cases

[20] Note, with essential heterogeneity, IV is no longer consistent for the ATE or ATT, but rather estimates the LATE. Hence the RMSE is larger than in the constant treatment effect case. See Heckman *et al*. (1999).

[21] As stated previously, traditional IV estimates the LATE. Thus the relative performance of traditional IV in terms of estimating the ATE or ATT depends on the extent of essential heterogeneity.

DOI: 10.1002/jae

Table II. Monte Carlo results: heterogeneous effect model ($\tau_i = 1 + \delta_i$)

	Homoskedastic error in treatment equation						Heteroskedastic error in treatment equation					
	ATE			ATT			ATE			ATT		
	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.20$	$\rho_0\sigma_0 = -0.40$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.20$	$\rho_0\sigma_0 = -0.40$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.20$	$\rho_0\sigma_0 = -0.40$	$\rho_0\sigma_0 = 0$	$\rho_0\sigma_0 = -0.20$	$\rho_0\sigma_0 = -0.40$
	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$	$\rho_{01} = 0.50$
	$\rho_\delta\sigma_\delta = 0$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = 0$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = 0$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = 0$	$\rho_\delta\sigma_\delta = -0.10$	$\rho_\delta\sigma_\delta = -0.10$
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
(A) *Normally distributed errors*												
τ_{IPW}	**0.043**	0.417	0.769	**0.049**	0.367	0.721	**0.044**	0.396	0.738	**0.048**	0.348	0.687
$\tau_{MB,\ 0.05}$	0.106	0.378	0.715	0.105	0.330	0.649	0.125	0.382	0.725	0.100	0.352	0.662
$\tau_{MB,\ 0.25}$	0.059	0.387	0.711	0.052	0.323	0.643	0.062	0.390	0.726	0.062	0.344	0.662
$\tau_{MB,\ EE,\ 0.05}$	0.119	0.456	0.797	0.116	0.355	0.695	0.157	0.382	0.706	0.154	0.307	0.648
$\tau_{MB,\ EE,\ 0.25}$	0.065	0.415	0.755	0.065	0.323	0.663	0.066	0.389	0.722	0.068	0.306	0.647
τ_{KV}	0.375	0.363	0.373	0.375	0.365	0.373	0.240	**0.241**	**0.244**	0.240	**0.246**	**0.249**
τ_{CF}	1.052	1.007	0.961	0.984	0.944	0.903	0.536	0.533	0.520	0.508	0.516	0.511
τ_{BVN}	0.289	0.291	**0.283**	0.215	**0.240**	**0.281**	0.271	0.282	0.295	0.204	0.258	0.345
τ_{BC}	0.296	0.295	0.287	0.298	0.298	0.289	0.281	0.286	0.289	0.287	0.297	0.296
$\tau_{MB\text{-}BC,\ 0.05}$	0.295	0.312	0.304	0.304	0.316	0.297	0.284	0.281	0.306	0.290	0.313	0.322
$\tau_{MB\text{-}BC,\ 0.25}$	0.281	**0.286**	0.285	0.289	0.295	0.282	0.266	0.272	0.301	0.278	0.295	0.311
$\tau_{BVN,\ EE}$	1.718	1.630	1.531	1.123	1.081	1.048	0.845	0.784	0.778	0.550	0.532	0.542
$\tau_{BC,\ EE}$	1.729	1.636	1.536	1.548	1.469	1.397	0.867	0.806	0.804	0.779	0.743	0.734
$\tau_{MB\text{-}BC,\ EE,0.05}$	1.387	1.358	1.425	1.447	1.399	1.331	0.769	0.770	0.802	0.695	0.690	0.758
$\tau_{MB\text{-}BC,\ EE,0.25}$	1.359	1.323	1.406	1.434	1.373	1.299	0.753	0.755	0.790	0.668	0.672	0.741
(B) *Asymmetric, non-normally distributed errors*												
τ_{IPW}	**0.042**	0.380	0.606	**0.047**	0.287	0.446	**0.045**	0.366	0.606	**0.047**	0.288	0.465
$\tau_{MB,\ 0.05}$	0.112	0.315	0.326	0.111	0.384	0.592	0.132	0.284	**0.239**	0.118	0.402	0.722
$\tau_{MB,\ 0.25}$	0.061	0.316	0.402	0.058	0.347	0.522	0.064	0.309	0.401	0.063	0.362	0.662
$\tau_{MB,\ EE,\ 0.05}$	0.126	0.487	0.778	0.116	0.307	0.368	0.167	0.400	0.563	0.160	0.285	0.476
$\tau_{MB,\ EE,\ 0.25}$	0.065	0.395	0.628	0.063	0.290	0.347	0.072	0.357	0.539	0.071	0.271	0.463
τ_{KV}	0.356	0.386	0.420	0.356	0.400	0.438	0.274	0.290	0.305	0.274	0.302	**0.316**
τ_{CF}	0.959	0.991	1.061	0.904	0.916	0.965	0.594	0.638	0.639	0.561	0.581	0.573

DOI: 10.1002/jae

Table II. (*Continued*)

	Homoskedastic error in treatment equation						Heteroskedastic error in treatment equation					
	ATE			ATT			ATE			ATT		
	$\rho_0\sigma_0=0$	$\rho_0\sigma_0=-0.20$	$\rho_0\sigma_0=-0.40$	$\rho_0\sigma_0=0$	$\rho_0\sigma_0=-0.20$	$\rho_0\sigma_0=-0.40$	$\rho_0\sigma_0=0$	$\rho_0\sigma_0=-0.20$	$\rho_0\sigma_0=-0.40$	$\rho_0\sigma_0=0$	$\rho_0\sigma_0=-0.20$	$\rho_0\sigma_0=-0.40$
	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$	$\rho_{01}=0.50$
	$\rho_\delta\sigma_\delta=0$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=0$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=0$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=0$	$\rho_\delta\sigma_\delta=-0.10$	$\rho_\delta\sigma_\delta=-0.10$
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
τ_{BVN}	0.269	**0.294**	**0.291**	0.201	**0.212**	0.315	0.267	**0.261**	0.258	0.204	**0.235**	0.376
τ_{BC}	0.274	0.295	0.293	0.272	0.269	**0.311**	0.280	0.271	0.268	0.286	0.278	0.378
$\tau_{MB\text{-}BC,\ 0.05}$	0.274	0.312	0.329	0.288	0.299	0.427	0.288	0.295	0.327	0.300	0.342	0.609
$\tau_{MB\text{-}BC,\ 0.25}$	0.258	0.296	0.310	0.271	0.273	0.367	0.261	0.264	0.295	0.270	0.306	0.553
$\tau_{BVN,\ EE}$	1.506	1.648	1.705	0.993	1.077	1.107	0.942	1.018	0.991	0.626	0.653	0.649
$\tau_{BC,\ EE}$	1.512	1.656	1.713	1.359	1.469	1.506	0.963	1.041	1.013	0.881	0.911	0.891
$\tau_{MB\text{-}BC,\ EE,\ 0.05}$	1.326	1.615	1.661	1.296	1.414	1.438	0.837	0.989	1.168	0.874	0.846	0.918
$\tau_{MB\text{-}BC,\ EE,\ 0.25}$	1.302	1.620	1.716	1.287	1.394	1.399	0.814	0.963	1.148	0.847	0.825	0.849

Note: See Table I.

DOI: 10.1002/jae

	Homoskedastic errors		Heteroskedastic errors	
	ATE	**ATT**	**ATE**	**ATT**
I. CIA holds				
Normal distribution	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest
Asymmetric non-normal distribution	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest	IPW best, MB ($\theta = 0.25$) inefficient, but outperforms the rest
II. CIA does not hold				
Normal distribution	BVN best; MB outperforms IPW	BVN best; MB outperforms IPW	KV best; MB and MB-EE outperform IPW	KV best; MB ($\theta = 0.25$) and MB-EE outperform IPW
Asymmetric non-normal distribution	MB-BC ($\theta = 0.25$), BVN, or KV best depending on the nature of the treatment effect (homogeneous or heterogeneous) and extent of selection on unobservable variables; MB outperforms IPW	BC ($\theta = 0.25$), BVN, or KV best depending on the nature of the treatment effect (homogeneous or heterogeneous) and extent of selection on unobservable variables; MB-EE outperforms IPW	BC ($\theta = 0.25$), BVN, or KV best depending on the nature of the treatment effect (homogeneous or heterogeneous); MB ($\theta = 0.05$) outperforms IPW	BVN or KV best depending on the nature of the treatment effect (homogeneous or heterogeneous) and extent of selection on unobservable variables, MB-EE outperforms IPW

Figure 3. Monte Carlo Results. Summary based on Tables I–IV.

since Millimet and Tchernis (2009) find that over-specifying the propensity score model when the CIA holds is warranted. We explore whether this conclusion extends to the current situation.

In the interest of brevity, we briefly highlight a few key findings. First, when the CIA holds, IPW has the smallest RMSE as long as the model is not under-specified; over-specifying the model is inconsequential and in some cases reduces the RMSE. The benefit from over-specifying the model is consistent with the prior literature documenting that using an estimated propensity score is preferable even when the true propensity score is known (Hirano *et al*., 2003; Morgan and Harding, 2006; Millimet and Tchernis, 2009; see also the discussion in Heckman and Navarro-Lozano, 2004, and Heckman *et al*., 1997, 1998). Abadie and Imbens (2009) also find that the asymptotic variance of propensity score matching estimators is smaller when the propensity score is estimated (as opposed to known). Similarly, the performances of MB and MB-EE are not hindered and sometimes helped when the model is over-specified relative to using the correct specification. Over-specifying the model does impose some cost on the remaining estimators in most cases.

Second, when the CIA holds with the correct specification but the model is under-specified in practice (which thereby invalidates the CIA), MB or MB-EE with a small radius ($\theta = 0.05$) noticeably outperforms IPW as well as the remaining estimators; MB-BC, BC, KV, CF, and BVN perform extremely poorly if the model is under-specified. The non-normal (EE) counterparts to MB-BC, BC, and BVN perform even worse, producing RMSEs close to 200 and upwards. Clearly, the added parametrization of the BVN model resulting from the Edgeworth series is highly sensitive to misspecification of the functional form. Nonetheless, the superior performance of MB and MB-EE is striking. When the researcher is confident in the absence of salient unobserved variables, but unsure of the correct functional form for the propensity score model, MB and MB-EE provide a valuable robustness check.

DOI: 10.1002/jae

Third, when the CIA fails to hold and the model is under-specified, MB and MB-EE with a small radius ($\theta = 0.05$) continue to perform much better than the remaining estimators. In particular, when the errors are homoskedastic, MB-EE tends to be preferred. When the errors are heteroskedastic, MB (MB-EE) is preferred when estimating the ATE (ATT). One consistent exception is when estimating the ATE with non-normal and heteroskedastic errors. In this case, IPW performs best, although none fares well.

Finally, when the CIA fails to hold and the model is over-specified, BVN or MB-BC with a large radius ($\theta = 0.25$) has the smallest RMSE if the errors are normal. This is true even if the errors are heteroskedastic, where KV outperforms BVN if the model is correctly specified but not if the model is over-specified. With non-normal and homoskedastic errors, MB performs best when estimating the ATE, while BVN performs best in most cases when estimating the ATT (KV is best for the ATT when the extent of selection on unobserved variables is very strong). For the ATE, BVN suffers from over-fitting – relative to the correct specification – leading to the superior performance by MB. Lastly, with non-normal and heteroskedastic errors, BVN and KV are predominantly the top two performers. The relative ranking between the two depends on the nature of the treatment effect (homogeneous vs. heterogeneous) and the extent of selection on unobserved variables.

Large-sample results. The simulations in Appendix A are replicated in Appendix B (supporting information), except that the results are based on 50 replications of 250,000 observations. With the increasing reliance of large-scale census datasets, such sample sizes are frequently encountered in practice. Again, in the interest of brevity, we briefly highlight a few key findings. First, there are a few instances where MB with a large radius ($\theta = 0.25$) outperforms IPW even when the CIA holds and the model is correctly or over-specified. Thus, MB achieves a small reduction in finite sample bias that outweighs the efficiency loss of MB when the sample size is quite large. Second, and importantly, there are several cases where MB-BC or BC performs best when the CIA fails, outperforming BVN despite the reliance of these estimators on BVN to estimate the correlation parameters. For example, with normal, heteroskedastic errors and focusing on the ATE, BC has a RMSE about 40% smaller than BVN and 10% smaller than KV when the model is over-specified. As such, our MB-BC and BC estimators offer a useful alternative to practitioners analyzing larger samples. However, the RMSE remains substantially larger than that for a traditional IV estimator, which falls to about 0.01 (0.03–0.06) without (with) essential heterogeneity.

5.2. Empirical Monte Carlo

5.2.1. Setup

Because the insights gained from any Monte Carlo design may be limited to the particular DGPs considered, we also undertake a so-called Empirical Monte Carlo study as developed in Huber *et al.* (2010). The authors 'suggest a different approach of conducting simulations . . .using the real data to simulate realistic "placebo treatments" among the non-treated' (p. 3). While we defer discussion of the data until the next section, we implement this approach in two ways. The first ensures the CIA holds, while the second does not.

To implement the first approach, we use the following algorithm:

1. Using the full sample from the actual data (discussed in Section 6.2), estimate the propensity score using the heteroskedastic probit model in (30).
2. Retaining only the sample of non-treated, simulate treatment assignment as

$$\tilde{T}_i = \mathrm{I}\left(\frac{X\hat{\gamma}}{exp\left(X\hat{\delta}\right)} + 0.32 + \varsigma_i > 0\right)$$

where $\varsigma_i \sim N(0, 1)$ and $\hat{\gamma}$ and $\hat{\delta}$ are the ML estimates from step 1. The constant, 0.32, is chosen to ensure that $\tilde{T}_i = 1$ for roughly 30% of the observations. The CIA holds given the independence of ς.

DOI: 10.1002/jae

3. Apply the various estimators using the non-treated sample, including the observed outcome, y, along with the simulated treatment assignment. The true value of τ_i is zero for all i by construction; thus this setup imposes a constant treatment effect.
4. Repeat steps 1–3.

To assess the performance of the various estimators when the CIA does not hold, we amend step 2 in the above algorithm. It now becomes

2′. Retaining only the sample of non-treated, simulate treatment assignment as

$$\tilde{T}_i = \mathrm{I}\left(\frac{X\hat{\gamma}}{exp\left(X\hat{\delta}\right)} + 0.2 + \varsigma_i > 0\right)$$

where $\varsigma_i \sim N(\tilde{y}_i, 1)$, $\tilde{y}_i = \left(y_i - \mu_y\right)/\sigma_y$, μ_y and σ_y is the mean and standard deviation of the outcome, and $\hat{\gamma}$ and $\hat{\delta}$ are the ML estimates from step 1. The constant, 0.2, is chosen to ensure that $\tilde{T}_i = 1$ for approximately 30% of the observations. The CIA does not hold since treatment assignment is correlated with y_0.

5.2.2. Results

Results from the Empirical Monte Carlo are presented in Tables III (CIA holds) and IV (CIA does not hold). In each table, we perform the Monte Carlo study using eight outcomes: BMI growth, overweight (1 = yes), obese (1 = yes), and underweight (1 = yes) for third- and fifth-grade students (discussed in Section 6.2).

The results in Table III accord perfectly with the previous Monte Carlo results presented in Tables I and II. Specifically, IPW is always the preferred estimator, with MB with a large radius ($\theta = 0.25$) faring slightly worse; MB-EE with a large radius ($\theta = 0.25$) does only slightly worse than MB. As before, the efficiency loss from using a smaller bandwidth is more costly than the efficiency loss from the estimation of the additional parameters required by MB-EE. After these estimators, KV and BVN have the next smallest RMSE, with the order varying, followed by the MB-BC and BC estimators. CF, BVN-EE, MB-BC-EE, and BC-EE have the highest RMSE.

The results in Table IV heavily favor KV, consistent with the results in Table I for heteroskedastic, normal or non-normal errors with a constant treatment effect. For all the third-grade outcomes and three of the four fifth-grade outcomes, KV achieves the lowest RMSE. The relative performance of the remaining estimators is not fixed. Lastly, MB, particularly with a small radius ($\theta = 0.05$), outperforms IPW in every case. When the outcome is underweight status, the improvement in RMSE is quite meaningful, falling by between roughly 50% and 70%. This continues to confirm the usefulness of our MB estimator when one is unsure if the CIA holds.

5.3. Discussion

In light of the Monte Carlo results, researchers are presented with a new set of tools for the estimation of the causal impacts of a treatment. If one believes that the CIA holds, but the proper functional form is unknown, then it is best to use IPW and err on the side of over-specifying the propensity score model. Whereas the usual practice in empirical research amounts to finding the most parsimonious specification that succeeds in balancing the covariates according to some metric (see, for example, Morgan and Harding 2006), the findings here and elsewhere suggest that a more saturated model is preferable.

DOI: 10.1002/jae

Table III. Empirical Monte Carlo results: selection on observable variables only

	ATE				ATT			
	BMI growth	Overweight	Obese	Underweight	BMI growth	Overweight	Obese	Underweight
(A) *Third grade*								
τ_{IPW}	**0.003**	**0.016**	**0.013**	**0.011**	**0.003**	**0.017**	**0.014**	**0.010**
$\tau_{MB,\ 0.05}$	0.008	0.045	0.033	0.026	0.007	0.041	0.033	0.025
$\tau_{MB,\ 0.25}$	0.004	0.024	0.017	0.014	0.004	0.022	0.018	0.013
$\tau_{MB,\ EE,\ 0.05}$	0.008	0.047	0.036	0.028	0.007	0.040	0.032	0.021
$\tau_{MB,\ EE,\ 0.25}$	0.004	0.023	0.019	0.015	0.004	0.022	0.019	0.013
τ_{KV}	0.013	0.098	0.051	0.073	0.013	0.098	0.051	0.073
τ_{CF}	0.037	0.320	0.326	0.270	0.052	0.571	0.618	0.315
τ_{BVN}	0.019	0.127	0.092	0.067	0.012	0.078	0.060	0.037
τ_{BC}	0.018	0.125	0.089	0.066	0.026	0.178	0.138	0.079
$\tau_{MB\text{-}BC,\ 0.05}$	0.020	0.128	0.086	0.072	0.026	0.177	0.140	0.082
$\tau_{MB\text{-}BC,\ 0.25}$	0.018	0.122	0.082	0.068	0.025	0.171	0.132	0.079
$\tau_{BVN,\ EE}$	0.092	0.670	0.759	0.468	0.052	0.433	0.517	0.243
$\tau_{BC,\ EE}$	0.092	0.672	0.761	0.470	0.130	1.066	1.280	0.596
$\tau_{MB\text{-}BC,\ EE,0.05}$	0.119	0.966	1.108	0.431	0.087	0.754	0.782	0.422
$\tau_{MB\text{-}BC,\ EE,0.25}$	0.117	0.957	1.100	0.421	0.085	0.756	0.781	0.420
(B) *Fifth grade*								
τ_{IPW}	**0.005**	**0.020**	**0.015**	**0.013**	**0.005**	**0.020**	**0.016**	**0.010**
$\tau_{MB,\ 0.05}$	0.013	0.050	0.041	0.035	0.012	0.047	0.044	0.026
$\tau_{MB,\ 0.25}$	0.007	0.025	0.022	0.017	0.006	0.024	0.021	0.013
$\tau_{MB,\ EE,\ 0.05}$	0.012	0.056	0.039	0.032	0.011	0.049	0.036	0.021
$\tau_{MB,\ EE,\ 0.25}$	0.006	0.029	0.023	0.017	0.006	0.026	0.021	0.012
τ_{KV}	0.057	0.092	0.084	0.137	0.057	0.092	0.084	0.137
τ_{CF}	0.046	0.222	0.216	0.680	0.064	0.287	0.451	0.773
τ_{BVN}	0.035	0.101	0.091	0.119	0.020	0.061	0.058	0.065
τ_{BC}	0.035	0.098	0.089	0.122	0.045	0.135	0.132	0.143
$\tau_{MB\text{-}BC,\ 0.05}$	0.038	0.102	0.092	0.124	0.046	0.140	0.139	0.141
$\tau_{MB\text{-}BC,\ 0.25}$	0.035	0.095	0.082	0.118	0.044	0.137	0.130	0.138
$\tau_{BVN,\ EE}$	0.099	0.539	0.488	1.033	0.055	0.304	0.351	0.535
$\tau_{BC,\ EE}$	0.099	0.539	0.489	1.037	0.137	0.770	0.888	1.318
$\tau_{MB\text{-}BC,\ EE,\ 0.05}$	0.164	0.773	0.921	0.875	0.106	0.512	0.602	0.928
$\tau_{MB\text{-}BC,\ EE,\ 0.25}$	0.161	0.759	0.913	0.866	0.104	0.509	0.603	0.930

Note: Numbers reflect the root mean squared error based on 250 simulations of placebo treatments for roughly 30% of the control group used in Table VI. See Table I and text for further details.

Our MB and MB-EE estimators, however, provide a useful robustness check since parametric models are probably often under-specified in practice. Although inefficient when the model is correctly specified or over-specified, they outperform IPW when the model is under-specified. The improved performance is achieved through a restriction of the sample to observations with a propensity score lying in the subset of the unit interval where the average treatment effects can be estimated with smaller bias. The bias reduction more than offsets the diminished sample size, resulting in a lower RMSE. As noted previously, this is similar to the solution to limited overlap advocated in Crump *et al*. (2009).

If one does not believe that the CIA holds and the proper functional form and error structure are unknown, then KV and BVN perform (relatively) well in general if the model is correctly or over-specified; MB-BC and BC perform relatively well in large samples in such situations. However, as these estimators are highly sensitive to misspecification of the functional form, our MB and MB-EE estimators continue to achieve the lowest RMSE when the model is under-specified. As this is probably the case most often confronted by applied researchers, this suggests that MB and MB-EE ought to become part of the applied researcher's toolkit in the absence of traditional exclusion restrictions.

That said, it is worth re-emphasizing two salient points before turning to the application. First, the MB estimators (most likely) alter the interpretation of the parameter being estimated. As a result, they

DOI: 10.1002/jae

Table IV. Empirical Monte Carlo results: selection on observable and unobservable variables

	ATE				ATT			
	BMI Growth	Overweight	Obese	Underweight	BMI Growth	Overweight	Obese	Underweight
(A) *Third grade*								
IPW	0.111	0.697	0.626	0.477	0.103	0.666	0.533	0.338
MB, 0.05	0.097	0.671	0.465	0.155	0.094	0.650	0.466	0.176
MB, 0.25	0.099	0.681	0.502	0.222	0.100	0.654	0.473	0.214
MB, EE, 0.05	0.114	0.711	0.702	0.793	0.098	0.587	0.379	0.162
MB, EE, 0.25	0.110	0.711	0.680	0.663	0.099	0.634	0.462	0.210
KV	**0.037**	**0.237**	**0.085**	**0.041**	**0.037**	**0.237**	**0.085**	**0.041**
CF	0.111	1.004	0.693	0.509	0.133	1.298	1.139	0.867
BVN	0.054	0.661	0.456	0.685	0.090	0.735	0.671	0.129
BC	0.054	0.655	0.449	0.670	0.074	0.835	0.880	0.273
MB-BC, 0.05	0.061	0.668	0.465	0.602	0.067	0.812	0.797	0.395
MB-BC, 0.25	0.063	0.677	0.502	0.537	0.072	0.816	0.804	0.358
BVN, EE	0.109	1.256	0.864	2.024	0.097	1.056	0.824	1.090
BC, EE	0.109	1.256	0.864	2.031	0.147	1.827	1.461	2.675
MB-BC, EE,0.05	0.124	1.084	0.722	0.856	0.111	1.378	1.035	1.748
MB-BC, EE, 0.25	0.124	1.072	0.723	0.902	0.112	1.419	1.108	1.796
(B) *Fifth grade*								
τ_{IPW}	0.156	0.696	0.652	0.480	0.145	0.676	0.576	0.341
$\tau_{MB,\ 0.05}$	0.137	0.634	0.544	0.139	0.135	0.671	0.522	0.150
$\tau_{MB,\ 0.25}$	0.139	0.665	0.573	0.206	0.140	0.671	0.533	0.196
$\tau_{MB,\ EE,\ 0.05}$	0.151	0.674	0.644	0.757	0.138	0.644	0.439	0.168
$\tau_{MB,\ EE,\ 0.25}$	0.149	0.688	0.653	0.650	0.139	0.665	0.516	0.205
τ_{KV}	0.105	**0.307**	**0.177**	**0.111**	**0.105**	**0.307**	**0.177**	0.111
τ_{CF}	0.102	0.449	0.521	0.544	0.125	0.619	0.855	0.729
τ_{BVN}	0.090	0.465	0.533	0.834	0.132	0.611	0.697	**0.105**
τ_{BC}	**0.089**	0.457	0.523	0.823	0.118	0.545	0.876	0.381
$\tau_{MB\text{-}BC,\ 0.05}$	0.099	0.469	0.544	0.743	0.109	0.546	0.807	0.525
$\tau_{MB\text{-}BC,\ 0.25}$	0.102	0.501	0.573	0.678	0.113	0.545	0.819	0.479
$\tau_{BVN,\ EE}$	0.139	0.719	0.556	2.352	0.112	0.640	0.615	1.149
$\tau_{BC,EE}$	0.139	0.719	0.556	2.356	0.183	1.007	0.881	2.951
$\tau_{MB\text{-}BC,EE,0.05}$	0.147	0.741	0.661	1.334	0.135	0.748	0.651	1.904
$\tau_{MB\text{-}BC,EE,0.25}$	0.147	0.732	0.679	1.415	0.135	0.755	0.700	1.946

Note: Numbers reflect the root mean squared error based on 250 simulations of placebo treatments for roughly 30% of the control group used in Table VI. See Tables I and III and text for further details.

may estimate a parameter considered to be uninteresting. Thus researchers ought to pay attention to the value of P^* as well as the attributes of observations with propensity scores close to this value. Second, none of the estimators considered here matches the performance of a traditional IV estimator, although IV may also change the interpretation of the parameter being estimated.

6. APPLICATION

6.1. Background

The SBP is a federally funded program, overseen by the US Department of Agriculture (USDA), but administered by state education agencies.[22] The SBP was established in 1966 by the Child Nutrition

[22] See Roy *et al.* (2012) for further detail on the SBP.

DOI: 10.1002/jae

Act, and made permanent in 1975. Participation by schools – both public and private – is voluntary (unless mandated by the state). In 1970, roughly 0.5 million students were served on an average school day. This figure increased to 4.0 million in 1990, 7.5 million in 2000, and 11.1 million in 2009 (9.1 million of which were free or reduced-price meals).[23] The program cost the federal government $2.6 billion in 2009.

If schools do participate, they are reimbursed a fixed amount per breakfast served.[24] However, meals must meet federal nutrition guidelines to qualify for reimbursement. During the period covered by the data, meals were governed by the 1995 'School Meals Initiative for Healthy Children' (SMI). Under the SMI no more than 30% of the breakfast's calories can be derived from fat, and less than 10% from saturated fat. Breakfasts also must provide one-quarter of the Recommended Dietary Allowance for protein, calcium, iron, vitamin A and vitamin C, and contain an age-appropriate level of calories. These guidelines were revised as part of the Healthy, Hunger-Free Kids Act of 2010 to meet the recommendations of the 2005 Dietary Guidelines for Americans. The implementation date for these changes is the 2012–2013 school year.

States are required to monitor local school food authorities through reviews conducted at least once every five years. In turn, the USDA monitors state compliance with this review requirement. The USDA has also begun to provide regional and local training to ensure adequate monitoring, as well as training in the preparation of healthy meals and dissemination of information for children related to the importance of a healthy diet. Results from the School Nutrition Dietary Assessment III (SNDA-III) collected in 2005 indicate improved compliance with the guidelines relative to the 1998–99 school year.[25]

6.2. Data

To analyze the impact of SBP participation on child health, we utilize data from the *Early Childhood Longitudinal Study – Kindergarten Class of 1998–99* (ECLS-K). We measure participation in the SBP in spring first grade; we ignore kindergarten participation due to many children attending half-day programs. Our outcomes of interest are measures of child health in spring third and fifth grade or the change from fall first grade to spring third and fifth grade. As such, we are analyzing more of the long-run relationship between child health and SBP participation. We utilize eight measures of child health: (i) growth rate in BMI (i.e. change in log BMI) from first grade to spring third and fifth grade; (ii) indicators for overweight status in spring third and fifth grade; (iii) indicators for obesity status in spring third and fifth grade; and (iv) indicators for underweight status in spring third and fifth grade. We define overweight (obesity) as having a BMI above the 85th (95th) percentile and underweight as having a BMI below the 20th percentile.[26]

We include the underweight indicator as the original focus of the SBP (and the NSLP) was on providing all children with a minimum level of nutrition (Guthrie *et al*., 2009). Only recently, with the rise in childhood obesity, has concern shifted to the upper end of the weight distribution, with these

[23] Students residing in households with family incomes at or below 130% of the federal poverty line are eligible for free meals, while those in households with family incomes between 130% and 185% of the federal poverty line are entitled to reduced price meals. In addition, children from households that receive aid through food stamps, Temporary Assistance for Needy Families, or the Food Distribution Program on Indian Reservations are automatically eligible for free meals.

[24] For the 2010–2011 school year, reimbursement rates are $1.48 per free meal, $1.18 per reduced-price breakfast, and $0.26 per full-priced breakfast. Schools establish their own prices for full-price meals, but prices for reduced-price meals are capped.

[25] See http://www.mathematica-mpr.com/nutrition/schoolmealsstudy.asp.

[26] Percentiles are obtained using the *-zanthro-* command in Stata. The CDC defines underweight as being below the fifth percentile. However, there are very few children below the fifth percentile in the data; thus we use a higher cut-off. In terms of trends, using the official CDC definition of underweight, the incidence of underweight children between the ages of 6 and 11 has declined from 5.3% in 1971–1974 to 2.7% in 2003–2006 (http://www.cdc.gov/nchs/data/hestat/underweight_children.pdf).

DOI: 10.1002/jae

programs trying to help children maintain a healthy weight (as opposed to being overweight). Thus it is important to examine whether the SBP meets its original objective of bringing underweight children into a healthy weight range.

The following covariates are included in X: child's race (white, black, or Hispanic), age, gender, child's birth weight, household socioeconomic status (SES), mother's employment status (full-time, part-time, or not working), mother's education (less than high school, high school or GED, vocational or training, four-year college degree, or advanced degree), number of children's books at home, mother's age at first birth, an indicator if the child's mother received WIC (Women, Infants, and Children) benefits prior to kindergarten, region (northeast, midwest, or south), city type (urban or suburban), and an indicator if the household never worried about running out of food in the past year. We also include quadratic and interaction terms involving the continuous variables, as well as interactions of race with age, gender, and SES. For the KV estimator, we model the error variance as a function of age, SES status, a dummy for residing in the south, and an urban dummy. All variables come from the fall or spring kindergarten wave.

Children with missing data for gender and age are dropped from our sample. We also restrict the sample to public schools. Missing values for the remaining control variables are imputed and imputation dummies are added to the control set. The final sample contains 9952 students when analyzing third-grade outcomes, of whom 3071 participate in the SBP. The sample size falls to 7876 when we analyze fifth-grade outcomes.[27] Table V provides summary statistics.

Prior to continuing, it is worth recalling from above that prior evidence in MTH indicates that children on steeper weight trajectories from birth through kindergarten entry are more likely to participate in the SBP. This suggests that unobserved variables associated with higher weight are likely correlated with treatment assignment, invalidating the CIA. Examples of such variables may include family background attributes such as the extent of parental oversight in the home or neighborhood attributes such as the level of crime and quality of recreational amenities. In addition, since schools are not required to participate in the SBP (and only 87% of schools participating in the NSLP did in 2009–10), school-level unobserved variables may also invalidate the CIA.[28] Examples may include the frequency and quality of physical education or the weight status of peers. Consequently, we expect IPW to be biased upward for BMI, overweight status, and obesity status; and biased down for underweight status.

6.3. Results

6.3.1. Baseline

The results are presented in Tables VI and VII. 90% confidence intervals based on the percentile method are obtained by bootstrap using 250 repetitions. [29] The ATT allows one to assess the expected effect of the program on current participants, and thus is relevant as an evaluation of the current program. The ATE allows one to assess the expected effect of current programs if near-universal participation, which is the goal of many, is achieved.[30] Moreover, the ATU, which is also relevant for assessing the effects of program expansion, may be deduced from the ATE and ATT and is therefore omitted.

Turning to the results in Tables VI and VII, IPW yields a positive and statistically significant association between SBP and BMI growth, overweight status, and obesity status. There is also a negative and statistically significant association between SBP participation and underweight status. In terms of magnitudes, the results indicate that SBP participation in first grade is associated with roughly a 1% (2%) increase in BMI growth between first and third (fifth) grade, with the associations being stronger

[27] See Appendix C (supporting information) for more details on the data creation.
[28] See http://frac.org/wp-content/uploads/2010/07/us.pdf.
[29] Note that when obtaining confidence intervals for MB and MB-EE we re-estimate $P^*(X)$ within each bootstrap repetition.
[30] See, for example, http://frac.org/federal-foodnutrition-programs/school-breakfast-and-lunch/outreach/.

DOI: 10.1002/jae

Table V. Summary statistics

Variable	Full sample		SBP participants		SBP non-participants	
	Mean	SD	Mean	SD	Mean	SD
SBP participation (1 = yes)	0.309	0.462	1	0	0	0
Third-grade child weight						
BMI growth rate	0.104	0.090	0.116	0.095	0.099	0.087
Overweight (1 = yes)	0.365	0.482	0.421	0.494	0.341	0.474
Obese (1 = yes)	0.197	0.398	0.238	0.426	0.179	0.383
Under (1 = yes)	0.087	0.281	0.068	0.252	0.095	0.293
Fifth-grade child weight						
BMI growth rate	0.198	0.122	0.221	0.124	0.188	0.119
Overweight (1 = yes)	0.413	0.492	0.491	0.500	0.380	0.485
Obese (1 = yes)	0.230	0.421	0.293	0.455	0.203	0.402
Under (1 = yes)	0.085	0.279	0.062	0.241	0.095	0.294
Controls						
Age (in months)	109.448	4.334	109.578	4.468	109.390	4.271
Gender (1 = boy)	0.513	0.500	0.508	0.500	0.515	0.500
White (1 = yes)	0.554	0.497	0.344	0.475	0.648	0.478
Black (1 = yes)	0.137	0.344	0.269	0.444	0.078	0.268
Hispanic (1 = yes)	0.191	0.393	0.265	0.442	0.157	0.364
Child's birth weight (ounces)	118.324	23.358	116.669	24.470	119.063	22.807
Child's birth weight (1 = missing)	0.080	0.271	0.120	0.325	0.062	0.241
Central city (1 = yes)	0.352	0.478	0.412	0.492	0.325	0.469
Urban fringe and large town (1 = yes)	0.397	0.489	0.267	0.442	0.455	0.498
Northeast (1 = yes)	0.178	0.383	0.111	0.314	0.208	0.406
Midwest (1 = yes)	0.247	0.431	0.197	0.398	0.269	0.444
South (1 = yes)	0.348	0.477	0.480	0.500	0.290	0.454
Mother's age at first birth (years)	23.567	5.090	21.290	4.232	24.584	5.111
Mother's age at first birth (1 = missing)	0.117	0.321	0.164	0.370	0.095	0.294
WIC benefits prior to kindergarten (1 = yes)	0.460	0.498	0.733	0.443	0.338	0.473
WIC benefits prior to kindergarten (1 = missing)	0.031	0.173	0.046	0.209	0.024	0.154
Mother's education = less than high school (1 = yes)	0.148	0.355	0.275	0.446	0.091	0.288
Mother's education = high school (1 = yes)	0.312	0.463	0.376	0.484	0.284	0.451
Mother's education = some college (1 = yes)	0.307	0.461	0.243	0.429	0.336	0.472
Mother's education = bachelor's degree (1 = yes)	0.136	0.343	0.045	0.206	0.177	0.382
Mother's education = advanced college degree (1 = yes)	0.065	0.247	0.016	0.125	0.087	0.282
Mother employed full-time during kindergarten (1 = yes)	0.403	0.491	0.401	0.490	0.404	0.491
Mother employed part-time during kindergarten (1 = yes)	0.189	0.391	0.131	0.337	0.214	0.410
Mother not working during kindergarten (1 = yes)	0.288	0.453	0.309	0.462	0.278	0.448
SES index	−0.055	0.771	−0.490	0.678	0.139	0.729
SES index (1 = missing)	0.016	0.126	0.026	0.160	0.012	0.107
Never worried about running out of food in household (1 = yes)	0.819	0.385	0.696	0.460	0.874	0.331
Never worried about running out of food in household (1 = missing)	0.040	0.197	0.064	0.244	0.030	0.171
Number of children's books in household	72.250	56.149	51.441	48.070	81.537	57.003
Number of children's books in household (1 = missing)	0.109	0.311	0.143	0.350	0.093	0.291

Note: $N = 9952$ for the full sample for third-grade outcomes; of this, 3071 are SBP participants and 6881 are non-participants. $N = 7876$ for the full sample for third-grade outcomes; of this, 2374 are SBP participants and 5502 are non-participants. Data are from the ECLS-K. BMI growth rate calculated using baseline data from first grade.

when focusing on the ATE rather than ATT. Similarly, first-grade SBP participation is associated with a 3.4 (5.9) percentage point increase in obesity in third (fifth) grade when focusing on the ATE; and 3.3 (5.2) percentage point increase when focusing on the ATT. The association with overweight status tends to be even larger; a 5.0 (8.4) percentage point increase in third (fifth) grade when focusing on the ATE and a 3.6 (5.1) percentage point increase when focusing on the ATT. Finally, first-grade SBP participation is associated with a 2.6 (3.0) percentage point *decrease* in underweight status in third (fifth) grade when focusing on the ATE; and a 1.7 (1.5) percentage point decrease when focusing on the ATT.

DOI: 10.1002/jae

Table VI. Effect of SBP participation: ATE

	Third-grade outcome				Fifth-grade outcome			
	BMI Growth	Overweight	Obese	Underweight	BMI Growth	Overweight	Obese	Underweight
τ_{IPW}	0.009	0.050	0.034	−0.026	0.024	0.084	0.059	−0.030
	[0.005, 0.015]	[0.022, 0.079]	[0.013, 0.055]	[−0.039, −0.012]	[0.018, 0.030]	[0.053, 0.113]	[0.033, 0.085]	[−0.042, −0.018]
$\tau_{MB,\ 0.05}$	0.015	0.015	0.028	−0.020	0.018	0.060	0.057	−0.014
	[−0.008, 0.021]	[−0.054, 0.088]	[−0.040, 0.094]	[−0.063, 0.014]	[−0.005, 0.033]	[−0.045, 0.123]	[−0.059, 0.117]	[−0.081, 0.035]
$\tau_{MB,\ 0.25}$	0.005	0.027	0.014	−0.018	0.012	0.053	0.033	−0.005
	[−0.002, 0.013]	[−0.006, 0.057]	[−0.004, 0.052]	[−0.044, −0.001]	[0.002, 0.021]	[−0.005, 0.072]	[−0.006, 0.065]	[−0.050, 0.008]
$\tau_{MB\text{-}EE,\ 0.05}$	0.014	0.090	0.056	−0.042	0.048	0.180	0.094	−0.073
	[0.009, 0.029]	[0.036, 0.170]	[0.009, 0.103]	[−0.082, 0.002]	[0.004, 0.068]	[0.051, 0.248]	[0.024, 0.176]	[−0.103, −0.025]
$\tau_{MB\text{-}EE,\ 0.25}$	0.013	0.062	0.034	−0.035	0.031	0.115	0.052	−0.048
	[0.005, 0.020]	[0.015, 0.106]	[0.009, 0.063]	[−0.053, −0.010]	[0.007, 0.041]	[0.027, 0.158]	[0.022, 0.100]	[−0.061, −0.013]
τ_{KV}	−0.008	−0.087	−0.180	−0.028	0.021	−0.090	−0.018	−0.043
	[−0.035, 0.026]	[−0.232, 0.124]	[−0.306, 0.008]	[−0.126, 0.075]	[−0.033, 0.070]	[−0.240, 0.153]	[−0.170, 0.180]	[−0.173, 0.072]
τ_{BVN}	−0.017	−0.161	−0.169	−0.013	−0.031	−0.248	−0.090	0.027
	[−0.046, 0.015]	[−0.318, −0.005]	[−0.278, −0.029]	[−0.098, 0.082]	[−0.073, 0.013]	[−0.402, −0.017]	[−0.230, 0.088]	[−0.087, 0.135]
τ_{BC}	−0.018	−0.167	−0.180	−0.016	−0.030	−0.256	−0.090	0.024
	[−0.048, 0.015]	[−0.331, −0.001]	[−0.291, −0.042]	[−0.104, 0.083]	[−0.075, 0.016]	[−0.415, −0.013]	[−0.233, 0.098]	[−0.096, 0.137]
$\tau_{MB\text{-}BC,\ 0.05}$	−0.007	−0.169	−0.159	−0.020	−0.021	−0.219	−0.070	0.017
	[−0.053, 0.019]	[−0.327, 0.037]	[−0.302, 0.003]	[−0.106, 0.097]	[−0.083, 0.019]	[−0.446, 0.022]	[−0.257, 0.100]	[−0.113, 0.148]
$\tau_{MB\text{-}BC,\ 0.25}$	−0.017	−0.157	−0.173	−0.018	−0.026	−0.225	−0.094	0.026
	[−0.047, 0.016]	[−0.325, 0.027]	[−0.276, −0.011]	[−0.099, 0.082]	[−0.076, 0.014]	[−0.405, 0.023]	[−0.245, 0.075]	[−0.088, 0.135]
P^*	0.672	0.538	0.462	0.868	0.754	0.593	0.498	0.874
	[0.124, 0.957]	[0.293, 0.939]	[0.263, 0.838]	[0.035, 0.944]	[0.395, 0.980]	[0.384, 0.959]	[0.114, 0.863]	[0.044, 0.972]
P^*_{EE}	0.033	0.059	0.866	0.020	0.020	0.026	0.857	0.020
	[0.020, 0.979]	[0.020, 0.974]	[0.020, 0.980]	[0.020, 0.955]	[0.020, 0.980]	[0.020, 0.980]	[0.020, 0.980]	[0.020, 0.975]

Note: Treatment is defined as participation in SBP in first grade. BMI growth measured from fall first grade. Overweight (obese) is defined as BMI above the 85th (95th) percentile. Underweight is defined as BMI below the 20th percentile. Covariates include all variables from Table V entered linearly, as well as squared and interaction terms for the continuous covariates. 90% empirical confidence intervals in brackets are obtained using 250 bootstrap repetitions. IPW, inverse propensity score weighting estimator; MB, minimum-biased estimator using $\theta = 0.05$ or 0.25; KV, Klein and Vella (2009) estimator; BVN, Heckman bivariate normal selection model; MB-BC, bias corrected estimator using $\theta = 0.05$ or 0.25; and, P^* (P^*_{EE}) is the bias-minimizing propensity score.

DOI: 10.1002/jae

Table VII. Effect of SBP participation: ATT

	Third-grade outcome				Fifth-grade outcome			
	BMI Growth	Overweight	Obese	Underweight	BMI Growth	Overweight	Obese	Underweight
τ_{IPW}	0.006	0.036	0.033	−0.017	0.015	0.051	0.052	−0.015
	[0.001, 0.012]	[0.012, 0.061]	[0.013, 0.055]	[−0.029, −0.006]	[0.009, 0.021]	[0.025, 0.080]	[0.032, 0.076]	[−0.029, −0.001]
$\tau_{MB,\ 0.05}$	0.000	0.035	0.000	−0.006	0.010	0.070	0.068	−0.015
	[−0.012, 0.012]	[−0.036, 0.101]	[−0.044, 0.077]	[−0.046, 0.024]	[−0.017, 0.027]	[−0.066, 0.097]	[−0.050, 0.100]	[−0.053, 0.034]
$\tau_{MB,\ 0.25}$	0.005	0.023	0.018	−0.015	0.010	0.025	0.033	−0.006
	[−0.001, 0.011]	[−0.005, 0.056]	[−0.008, 0.047]	[−0.032, −0.001]	[0.003, 0.019]	[0.000, 0.071]	[0.011, 0.073]	[−0.027, 0.011]
$\tau_{MB\text{-}EE,\ 0.05}$	0.009	0.049	0.060	−0.034	0.018	0.055	0.090	−0.017
	[0.003, 0.024]	[−0.005, 0.106]	[0.000, 0.096]	[−0.052, 0.006]	[0.000, 0.038]	[−0.003, 0.123]	[0.036, 0.144]	[−0.048, 0.011]
$\tau_{MB\text{-}EE,\ 0.25}$	0.005	0.035	0.034	−0.016	0.013	0.043	0.059	−0.009
	[−0.002, 0.012]	[0.003, 0.066]	[0.007, 0.063]	[−0.033, 0.001]	[0.005, 0.022]	[0.007, 0.076]	[0.030, 0.086]	[−0.030, 0.008]
τ_{KV}	−0.008	−0.087	−0.180	−0.028	0.021	−0.090	−0.018	−0.043
	[−0.035, 0.026]	[−0.232, 0.124]	[−0.306, 0.008]	[−0.126, 0.075]	[−0.033, 0.070]	[−0.240, 0.153]	[−0.170, 0.180]	[−0.173, 0.072]
τ_{BVN}	−0.003	−0.060	−0.073	−0.020	−0.003	−0.085	−0.022	−0.003
	[−0.021, 0.015]	[−0.156, 0.040]	[−0.141, 0.005]	[−0.068, 0.037]	[−0.028, 0.023]	[−0.171, 0.050]	[−0.097, 0.078]	[−0.068, 0.057]
τ_{BC}	−0.014	−0.154	−0.173	−0.019	−0.018	−0.224	−0.084	0.009
	[−0.047, 0.028]	[−0.341, 0.043]	[−0.322, −0.010]	[−0.125, 0.094]	[−0.072, 0.037]	[−0.402, 0.056]	[−0.243, 0.132]	[−0.116, 0.137]
$\tau_{MB\text{-}BC,\ 0.05}$	−0.019	−0.143	−0.193	−0.008	−0.022	−0.188	−0.059	0.007
	[−0.054, 0.024]	[−0.346, 0.052]	[−0.318, 0.009]	[−0.117, 0.109]	[−0.082, 0.027]	[−0.436, 0.048]	[−0.293, 0.126]	[−0.115, 0.144]
$\tau_{MB\text{-}BC,\ 0.25}$	−0.014	−0.156	−0.176	−0.017	−0.021	−0.234	−0.095	0.016
	[−0.048, 0.025]	[−0.337, 0.032]	[−0.310, −0.005]	[−0.116, 0.088]	[−0.074, 0.032]	[−0.399, 0.040]	[−0.239, 0.122]	[−0.112, 0.137]
P^*	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]	[0.500, 0.500]
P^*_{EE}	0.787	0.828	0.863	0.744	0.950	0.801	0.874	0.781
	[0.737, 0.980]	[0.778, 0.951]	[0.661, 0.980]	[0.538, 0.960]	[0.300, 0.980]	[0.643, 0.979]	[0.754, 0.980]	[0.672, 0.958]

Note: See Table VI.

If we consider the model to be under-specified with heteroskedastic, non-normal errors, then the MB (MB-EE) estimates with a small radius ($\theta = 0.05$) are preferred for the ATE (ATT).[31] For the ATE, the MB estimator fails to find a statistically meaningful relationship between SBP and any of the outcomes. Notably, while this arises in part due to a loss in precision, it also reflects a reduction (in absolute value) in the point estimates. Specifically, the point estimates fall in absolute value for seven of the eight outcomes in Table VI; the lone exception is third-grade BMI growth. For the ATT, however, the MB-EE estimator yields point estimates that are larger than IPW for all eight outcomes, as well as statistically significant for BMI growth and obesity.

To interpret these results, one must consider the value of the BMPS. For the ATE, the BMPS lies in the upper tail of the propensity score distribution, with values ranging from 0.46 to 0.87; the median of the distribution of the estimated propensity score is roughly 0.25. Thus the proper interpretation of the results in Table VI is that there is no evidence of a statistically meaningful causal effect of first-grade SBP participation on subsequent health outcomes for the average student with a relatively high probability of program participation. For the ATT, the BMPS is always above 0.74. As such, the correct interpretation of the results in Table VII is that there is statistically meaningful evidence of a detrimental causal effect of first-grade SBP participation on subsequent health outcomes for the average program participant with a relatively high probability of program participation. Further inspection of the propensity score model reveals that those with a high value of propensity score are much less likely to be white and have a mother with a high school diploma and are much more likely to reside in a food-insecure, low-SES household located in an urban, southern setting.

If we consider the model to be correctly or over-specified with heteroskedastic, non-normal errors, then KV is the preferred estimator, although BVN and BC are also interesting to consider. Unlike the MB or MB-EE estimator, each of these estimators produces an estimate of the unconditional ATE or ATT. The estimates produce two main insights. First, all three estimators yield predominantly *negative* point estimates when analyzing BMI growth, overweight status, and obesity status, regardless of whether one focuses on the ATE or ATT. The point estimates are statistically significant in a few cases, such as the ATE for third-grade obesity status and fifth-grade overweight status using BVN or BC or the ATT for third-grade obesity status using BC. Moreover, the point estimates for all three estimators are extremely large, particularly for overweight and obesity status. This may be attributable to the fact that we examine more long-run outcomes than the majority of the prior literature. That said, given the wide confidence intervals, as well as the fact that the Empirical Monte Carlo results in Table IV indicate that KV is much noisier when examining overweight and obesity status, we are hesitant to draw conclusions regarding the exact magnitudes of the effects. However, the signs of the estimated effects are consistent with the prior literature. Second, all three estimators yield statistically insignificant, but predominantly *negative*, point estimates when assessing underweight status. This is consistent with the SBP achieving its original goal of ensuring a minimal nutrition level to students.

As noted, the combined results yield a picture consistent with the prior literature, particularly MTH. Without controlling for selection on unobserved variables, SBP appears to contribute to childhood obesity. However, after addressing some of the possible selection on unobserved variables by reducing the bias, the results are weakened for the ATE (but not the ATT). Finally, utilizing methods that aim to remove the bias due to unobserved variables – KV, BVN, and BC – there is no longer any evidence that SBP contributes to childhood obesity; there is some statistically meaningful evidence to the

[31] We focus on the case of heteroskedastic, non-normal errors for two reasons. First, normality seems unlikely in general. Second, estimation of the propensity score using a heteroskedastic probit easily rejects the null of homoskedasticity at the $p < 0.01$ confidence level using a likelihood ratio test. In particular, we find that age and urban status are statistically significant, at conventional levels, and increase the error variance. The greater conditional variability with age may reflect greater variation in the independence of children that comes with age and, hence, the ability of children to prepare their own breakfast at home. Urban status may increase the conditional variance due to greater variation in neighborhood attributes (e.g. safety or the presence of fast food restaurants).

DOI: 10.1002/jae

contrary. Thus, as suggested in MTH, it is important to address positive selection into SBP in order to estimate the causal effect of participation.

We also restrict the sample to households with an income less than 200% of the federal poverty guidelines during spring first grade. To conserve space, we highlight three main differences from the full-sample results and relegate the results to Appendix D (supporting information). First, while the pattern of IPW estimates is similar to the full sample, most estimates are smaller (in absolute value), consistent with less nonrandom selection on unobserved variables in this more homogeneous sample. Second, the majority of the estimates are not statistically significant. Third, while the point estimates from BVN and BC are almost exclusively negative, many of the KV point estimates are now positive. However, all are imprecise.

6.3.2. Measurement Error

If program participation is measured with error, the preceding estimates are likely to be attenuated even though measurement error in a binary variable is necessarily nonclassical (Black *et al*., 2000). While we are not aware of any specific evidence regarding misclassification rates for SBP participation, there is ample evidence of measurement error in survey data regarding participation in other programs (Millimet, 2011). To mitigate the impact of possible classification errors, we follow a strategy loosely based on Black *et al*. (2000), who note that one can reduce the bias from measurement error in a binary regressor if one has two mismeasured indicators and the measurement errors are independent. In such a case, an improved estimate is obtained by defining a binary variable equal to one if both mismeasured indicators are equal to one. In this spirit, we re-estimate the preceding models except now we define the treatment as one if the student reported eating breakfast at school in first and third grade (when using third-grade outcomes) and first, third, and fifth grade (when using fifth grade outcomes).[32]

In the interest of brevity, the results are placed in Appendix D (supporting information) and we simply note that the results are qualitatively similar. The two most noteworthy differences are that the MB-EE with a small radius ($\theta = 0.05$) estimates of the ATT in the full sample are no longer statistically significant, and the fact that the KV estimates are now negative and statistically significant when examining obesity status in the full sample (along with BVN and BC continuing to be statistically significant). Arguably, both these results may be due more to the new definition of the treatment rather than a correction for classification errors. As such, there is stronger evidence of a beneficial causal effect of SBP participation for persistent participators.

6.3.3. NSLP Participation

As a final analysis, we replicate the results presented in Tables VI and VII and Appendix D (supporting information), except using NSLP participation as the treatment. As discussed previously, there appears to be credible evidence that the NSLP contributes to childhood obesity. There is also less evidence of nonrandom selection on unobserved variables when it comes to the NSLP. Thus we expect to find evidence of a harmful effect; this constitutes a useful check on the estimators considered here. The results are in Appendix E (supporting information).

As expected, the IPW estimates are consistently positive when assessing BMI growth, overweight status, and obesity status; the estimates are statistically significant in the full sample, less so in the low income sample. Moreover, while the MB and MB-EE with a small radius ($\theta = 0.05$) estimates are positive, but not statistically significant, as was predominantly the case when analyzing SBP

[32] This exercise is merely meant to be suggestive. First, the assumption that measurement error in participation reports across waves are independent is not likely to hold. Second, unlike Black *et al*. (2000), our multiple indicators come from different points in time. As such, regardless of whether there is measurement error in the data, we are implicitly redefining the treatment being evaluated. In other words, the procedure here redefines the treatment of interest to one that is arguably less prone to classification error.

DOI: 10.1002/jae

participation, the KV, BVN, and BC estimates are now positive, large in magnitude, and occasionally statistically significant. Specifically, the estimates are statistically significant for obesity status when analyzing the full sample. The fact that these estimators point to a detrimental effect of NSLP participation, but a beneficial effect of SBP participation, gives us more confidence in our findings.

7. CONCLUSION

The program evaluation literature has expanded rapidly over the past decade. While our understanding of methods designed to provide consistent estimates of some measure of the causal effect of a binary treatment under conditional independence, as well as typical IV methods when conditional independence fails, is relatively well developed, researchers are less informed about how to proceed when conditional independence fails yet the usual type of exclusion restrictions is unavailable. In this study, we propose two new estimators for this situation, and evaluate the performance of our estimators as well as that of the bivariate normal selection model, the selection model extended to the case of non-normality, a parametric version of an IV estimator recently proposed in Klein and Vella (2009) that relies on heteroskedasticity, and an alternative control function approach. In addition, we use these estimators to assess the causal impact of a program of interest to policymakers: the School Breakfast Program.

Our analysis leads to some general guidelines that applied researchers may wish to follow in similar situations moving forward. First, in applications where the researcher believes conditional independence holds, our minimum-biased estimator offers a nice robustness check since it performs nearly as well when the model is correctly specified or over-specified, but vastly better when the model is under-specified. Second, when conditional independence does not hold but the model is correctly specified or over-specified, our bias-corrected estimator does quite well in large samples, even outperforming estimators relying purely on functional form for identification. Third, our parametric Klein and Vella (2009) estimator performs well when the error in the treatment equation is in fact heteroskedastic; the bivariate normal selection model tends to perform best when the errors are homoskedastic. Finally, the penalty to over-specifying the model, if there is one at all, pales in comparison to the penalty from under-specifying the model. However, the KV estimator appears to be most sensitive to over-fitting.

In terms of our analysis of the SBP, the various estimators offer a coherent picture of the causal effect of the program. Specifically, we find a positive and statistically significant association between SBP and child weight when using estimators that require conditional independence. The association remains positive, but becomes statistically insignificant, when we use our minimum-biased estimator to assess the ATE; it remains statistically significant when estimating the ATT. Finally, consistent with the suggestive evidence in MTH, as well as Bhattacharya *et al.* (2006), we find a negative and occasionally statistically significant causal effect of SBP participation on child weight using the bivariate normal selection model and Klein and Vella's (2009) estimator, as well as our bias-corrected estimator. Moreover, consistent with Schanzenbach (2009), these same estimators point a positive causal effect of school lunch on child weight. These findings have important implications for policymakers concerned with school nutrition. That said, as indicated in the Monte Carlo study and suggested by the width of the confidence intervals, the finite sample performance of the estimators relied on here is perhaps not ideal. Thus future attempts to uncover plausible exclusion restrictions, or create exogenous variation via experimental designs, are still needed.

ACKNOWLEDGEMENTS

This study was conducted by Georgia State University and Southern Methodist University under a cooperative agreement with the US Department of Agriculture, Economic Research Service, Food and

DOI: 10.1002/jae

Nutrition Assistance Research Program (agreement no. 58-5000-8-0097). The views expressed here are those of the authors and do not necessarily reflect those of the USDA or ERS. The authors benefited from comments from Ed Vytlacil, four anonymous referees, Jay Bhattacharya, Chris Bollinger, Ozkan Eren, Juan Carlos Escanciano, Jason Fletcher, James Heckman, Keisuke Hirano, Michael Lechner, Arthur Lewbel, Richard Luger, Salvador Navarro-Lozano, Denis Nekipelov, Joon Park, Kosali Simon, Jeff Smith, Justin Tobias, Jessica Todd, Aaron Yelowitz, seminar participants at Clemson University, Colorado College, Cornell University, University Georgia, Indiana University, University of Kentucky, Purdue University, Tulane University, and SMU, and conference participants at the 2009 Econometric Society Winter Meetings, Fifth IZA Conference on Labor Market Policy Evaluation, 2010 American Economic Association Annual Meetings, 2010 Western Economic Association Meetings, 2011 Society for Labor Economists Annual Conference, 2011 Conference on 'Emerging Issues in Agricultural Economics' in Rehovot, Israel, 2011 Econometric Society North American Summer Meetings, and Texas Camp Econometrics XV. Mehtabul Azam provided valuable research assistance. A previous version of this paper was circulated under the title 'Minimizing Bias in Selection on Observables Estimators When Unconfoundness Fails'. Stata code to implement the estimators used is available from the author's web page.

REFERENCES

Abadie A, Imbens G. 2009. Matching on the estimated propensity score. NBER Working Paper No. 15301.

Altonji J, Elder T, Taber C. 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *Journal of Political Economy* **113**: 151–184.

Bhattacharya J, Currie J, Haider S. 2006. Breakfast of champions? The School Breakfast Program and the nutrition of children and families. *Journal of Human Resources* **41**: 445–466.

Black D, Smith J. 2004. How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics* **121**: 99–124.

Black D, Berger M, Scott F. 2000. Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association* **95**: 739–748.

Bushway S, Johnson B, Slocum L. 2007. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology* **23**: 151–178.

Busso M, DiNardo J, McCrary J. 2011. New evidence on the finite sample properties of propensity score reweighting and matching estimators. IZA Discussion Paper No. 3998.

Campbell B, Nayga R, Park J, Silva A. 2011. Does the National School Lunch Program improve children's dietary outcomes? *American Journal of Agricultural Economics* **93**: 1099–1130.

Crump R, Hotz V, Imbens G, Mitnik O. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**: 187–199.

Dehejia R, Wahba S. 1999. Casual effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**: 1053–1062.

Finkelstein E, Trogdon J, Cohen J, Dietz W. 2009. Annual medical spending attributable to obesity: payer- and service-specific estimates. *Health Affairs* **28**: w822–w831.

Fisher R. 1935. The Design of Experiments. Oliver & Boyd: Edinburgh.

Goldberger A. 1983. Abnormal selection bias. In Studies in Econometrics, Time Series and Multivariate Statistics, Karlin S, Amemiya T (eds). Academic Press: New York; 67–84.

Guthrie J, Newman C, Ralston K. 2009. USDA school meal programs face new challenges. *Choices* **24**. Available: http://www.choicesmagazine.org/magazine/article.php?article=83 [4 May 2012].

Headrick T, Sawilowsky S. 1999. Simulating correlated multivariate nonnormal distributions: extending the Fleishman power method. *Psychometrika* **64**: 25–35.

Heckman J, Robb R. 1985. Alternative methods for evaluating the impact of interventions. In Longitudinal Analysis of Labor Market Data, Heckman J, Singer B (eds). Cambridge University Press: Cambridge, UK; 156–245.

Heckman J, Navarro-Lozano S. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics* **86**: 30–57.

Heckman J, Vytlacil E. 2005. Structural equations, treatment effects and econometric policy evaluation. *Econometrica* **73**: 669–738.

Heckman J, Ichimura J, Todd P. 1997. Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**: 261–294.

DOI: 10.1002/jae

Heckman J, Ichimura J, Todd P. 1998. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* **64**: 605–654.
Heckman J, LaLonde R, Smith J. 1999. The economics and econometrics of active labor market programs. In Handbook of Labor Economics, Ashenfelter A, Card D (eds). Elsevier: Amsterdam; **3**: 1865–2097.
Heckman J, Urzua S, Vytlacil E. 2006. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics* **88**: 389–432.
Hirano K, Imbens G 2001. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2**: 259–278.
Hirano K, Imbens G, Ridder G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**: 1161–1189.
Huber M, Lechner M, Wunsch C. 2010. How to control for many covariates? Reliable estimators based on the propensity score. IZA DP No. 5268.
Imbens G. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* **86**: 4–29.
Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* **62**: 467–475.
Imbens G, Wooldridge J. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**: 5–86.
Khan S, Tamer E. 2010. Irregular identification, support conditions, and inverse weight estimation. *Econometrica* **78**: 2021–2042.
Klein R, Vella F. 2009. A semiparametric model for binary response and continuous outcomes under index heteroskedasticity. *Journal of Applied Econometrics* **24**: 735–762.
Lee L-F. 1984. Tests for the bivariate normal distribution in econometric models with selectivity. *Econometrica* **52**: 843–863.
Lee D, Lemieux T. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* **48**: 281–355.
Leung S, Yu S. 2000. Collinearity and two-step estimation of sample selection models: problems, origins, and remedies. *Computational Economics* **15**: 173–199.
Mardia K. 1970. Families of Bivariate Distributions. Hafner: Darien, CT.
Millimet D. 2011. The elephant in the corner: a cautionary tale about measurement error in treatment effects models. *Advances in Econometrics: Missing-Data Methods* **27**: 1–39.
Millimet D, Tchernis R. 2009. On the specification of propensity scores: with applications to the analysis of trade policies. *Journal of Business and Economic Statistics* **27**: 397–415.
Millimet D, Tchernis R, Hussain M. 2010. School nutrition programs and the incidence of childhood obesity. *Journal of Human Resources* **45**: 640–654.
Morgan S, Harding D. 2006. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociological Methods and Research* **35**: 3–60.
Mroz T. 1999. Discrete factor approximations for use in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome. *Journal of Econometrics* **92**: 233–274.
Navarro S. 2008. Control function. In The New Palgrave Dictionary of Economics (2nd edn), Durlauf S, Blume L (eds). Palgrave Macmillan: London.
Neyman J. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (transl.). *Statistical Science* **5**: 465–480.
Puhani P. 2000. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14**: 53–68.
Rosenbaum P, Rubin D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.
Roy A. 1951. Some thoughts on the distribution of income. *Oxford Economic Papers* **3**: 135–146.
Roy M, Millimet D, Tchernis R. 2012. Federal nutrition programs and childhood obesity: inside the black box. *Review of Economics of the Household* (forthcoming).
Rubin D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* **66**: 688–701.
Schanzenbach D. 2009. Does the Federal School Lunch Program contribute to childhood obesity? *Journal of Human Resources* **44**: 684–709.
Serdula M, Ivery D, Coates R, Freedman D, Williamson D, Byers T. 1993. Do obese children become obese adults? A review of the literature. *Preventative Medicine* **22**: 167–177.
Smith J, Todd P. 2005. Does matching overcome LaLonde's critique? *Journal of Econometrics* **125**: 305–353.
Trasande L, Liu Y, Fryer G, Weitzman M. 2009. Effects of childhood obesity on hospital care and costs, 1999–2005. *Health Affairs* **28**: w751–w760.

DOI: 10.1002/jae