MOSTLY HARMLESS SIMULATIONS? ON THE INTERNAL VALIDITY OF EMPIRICAL MONTE CARLO STUDIES[*]

Arun Advani[†] and Tymon Słoczyński[‡]

Abstract

In this paper we evaluate the premise from the recent literature on empirical Monte Carlo studies that an empirically motivated simulation exercise is informative about the actual ranking of various estimators when applied to a particular problem. We use two alternative designs to generate a large number of data sets which mimic the well-known NSW-CPS and NSW-PSID sets. We then compare the "true ranking" of various estimators for the average treatment effect on the treated with rankings implied by the results of our simulations. We conclude that a necessary condition for the simulations to be informative about the true ranking is that the treatment effect in the simulation must be equal to the treatment effect in the original data. This severely limits the usefulness of such procedures in practice, since if the effect were known the procedure would not be necessary.

[*] With thanks to Gil Epstein, Ira Gang, Michael Lechner, Blaise Melly, Cathy Redmond, Jeffrey Wooldridge, and seminar participants at CERGE-EI and the Warsaw School of Economics. This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network. All opinions expressed are those of the authors and have not been endorsed by CERGE-EI or the GDN. Tymon Słoczyński also acknowledges a START scholarship from the Foundation for Polish Science (FNP).

[†] Institute for Fiscal Studies; University College London; and King's College, Cambridge.

[‡] Warsaw School of Economics; and IZA, Bonn.

1 Introduction

Monte Carlo studies constitute a standard approach in econometrics and statistics to examining small-sample properties of various estimators whenever theoretical results are unavailable. Recent papers by Frölich (2004), Lunceford and Davidian (2004), Zhao (2004, 2008), Busso et al. (2009), Millimet and Tchernis (2009), Austin (2010), Abadie and Imbens (2011), Busso et al. (2013), and Huber et al. (2013) have carried out Monte Carlo experiments to assess the relative finite-sample performance of a large number of estimators for various average treatment effects of interest.[1]

Most of these recent papers have used highly stylised data-generating processes (DGPs) which only loosely correspond to any actual evaluation data sets (see, e.g., Frölich 2004 and Busso et al. 2009). This approach has been criticised by Huber et al. (2013) on the grounds that Monte Carlo experiments are design dependent so can only be useful when based on realistic DGPs. They suggest that the conclusions of many Monte Carlo studies may be inapplicable to real-world estimation problems, i.e. the external validity of these studies is low. Instead, they propose an approach to generating artificial data sets which closely mimics the original data of interest, which they term an “empirical Monte Carlo study”. Similar simulation exercises have been carried out by Abadie and Imbens (2011) and Busso et al. (2013), who use a different procedure but it is again adapted to the circumstance of interest.[2]

Busso et al. (2013) encourage empirical researchers to “conduct a small-scale simulation study designed to mimic their empirical context” in order to choose the appropriate estimator(s) for a given research question. This suggestion is based on the premise that a

[1] See, e.g., Blundell and Costa Dias (2009) and Imbens and Wooldridge (2009) for recent reviews of the treatment effects framework.

[2] As noted by Huber et al. (2013), the idea of using data to inform Monte Carlo studies goes back at least as far as Stigler (1977).

carefully designed and empirically motivated Monte Carlo experiment is capable of informing the empirical researcher of the actual ranking of various estimators when applied to a given problem using a given data set. In other words, one must accept a proposition that "the advantage [of an empirical Monte Carlo study] is that it is valid in at least one relevant environment" (Huber et al. 2013), i.e. its internal validity is high by construction. In this paper we evaluate this important premise.

Two different approaches to conducting empirical Monte Carlo simulations have been proposed in the literature. The first, which we term the "structured" design, has been considered by both Abadie and Imbens (2011) and Busso et al. (2013). Loosely speaking, in this setting covariate values are drawn from a distribution similar to that in the data, and then treatment status and outcomes are generated using parameters estimated from the data. The second approach, which we term the "placebo" design, was proposed by Huber et al. (2013). Here both covariates and the outcome are drawn from the control data with replacement, and treatment status is assigned using parameters from the data. However, since all observations come from the control data and the original outcomes are retained, the treatment effect is known to be zero by construction.

We implement both of these approaches using the well-known NSW-CPS and NSW-PSID data sets, previously analysed by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2005), Abadie and Imbens (2011), Diamond and Sekhon (2012), and many others.[3] Since this programme originally had an experimental control group, an unbiased estimate of the effect of the NSW programme can be computed. Following LaLonde (1986) we use this true effect to calculate the bias (in these data) for a large set of estimators. We can then compare these biases, and the ranking of the estimators, to those we find from using the simulation designs considered. If empirical Monte

[3] Also, the NSW data were the subject of several recent empirically motivated Monte Carlo experiments (see, e.g., Lee and Whang 2009, Abadie and Imbens 2011, Diamond and Sekhon 2012, and Busso et al. 2013).

Carlo methods are internally valid, there should be a strong positive correlation between the biases found in the data and those found in the simulations.

We find that the structured approach to empirical Monte Carlo studies is valid only under the restrictive assumption that the treatment effect in the original data is equal to the treatment effect implied by the simulation procedure. This result precludes the use of this method in the practical choice of estimators: if we know that this assumption holds then we already know the true treatment effect, and if not then the method can provide severely misleading answers.

The placebo design is similarly problematic, but for an additional reason. As with the structured design the true effect in simulations is likely to be different than the actual effect of a given programme. However, it is also less representative of the original data, since it uses only the control data. This means that when the treated observations are quite different to the controls it is less good at replicating the original data.

Hence we conclude that there is little support for the chief premise of the recent literature on empirical Monte Carlo studies: that they are at least informative about the appropriate choice of estimator for the data at hand. We caution researchers against seeing these methods as a panacea which provides information about estimator choice, and to instead continue using several different estimators as a form of robustness check.

2 The National Supported Work (NSW) Data

The National Supported Work Demonstration (NSW) was a work experience programme which operated in the mid-1970s at 15 locations in the United States (for a more detailed description of the programme, see Smith and Todd 2005). It served several groups of disadvantaged workers, such as women with dependent children receiving welfare, former drug addicts, ex-criminals, and school dropouts. Unlike many similar programmes, the NSW programme selected its participants randomly, and such a method of selection into the programme allowed for its straightforward evaluation via a comparison of mean outcomes in the treatment and control groups.

In an influential paper, LaLonde (1986) suggested that one could use the design of this programme to assess the performance of various nonexperimental estimators of the average treatment effect. He discarded the original control group from the NSW data and created several alternative comparison groups using data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID), two standard data sets on the U.S. working population. LaLonde (1986) suggested that a reasonable estimator of the average treatment effect should be able to closely replicate the experimental estimate of the effect of the NSW programme on the outcomes of its participants, using data from the treatment group and the nonexperimental comparison groups. He found that very few of the estimates were close to the experimental benchmark. This result motivated a large number of replications and follow-ups, and established a testbed for new estimators for various average treatment effects of interest (see, e.g., Heckman and Hotz 1989, Dehejia and Wahba 1999, Smith and Todd 2005, Abadie and Imbens 2011, Diamond and Sekhon 2012).

The key insight of LaLonde (1986) was that a sensible estimator for the average treatment effect should be able to closely replicate the "true" experimental estimate of this effect using nonexperimental data. In this paper we suggest that a reasonable empirical Monte Carlo study should be able to closely replicate the "true" *ranking* of nonexperimental estimators, based on their ability to uncover this "true" estimate. In our analysis, we use the subset of the treatment group (185 observations) which was created by Dehejia and Wahba (1999) as well as the original CPS and PSID comparison groups (15,992 and 2,490 observations, respectively) which were created by LaLonde (1986), and we aim at creating a large num-

Table 1: Descriptive Statistics for the NSW-CPS and NSW-PSID Data Sets

	NSW		*CPS*		*PSID*	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Number of observations	185		15,992		2,490	
Outcome variable						
Nonemployed ‘78	0.24	0.43	0.14	0.34	0.11	0.32
Control variables						
Age	25.82	7.16	33.23	11.05	34.85	10.44
Black	0.84	0.36	0.07	0.26	0.25	0.43
Education	10.35	2.01	12.03	2.87	12.12	3.08
Married	0.19	0.39	0.71	0.45	0.87	0.34
‘Earnings ‘74’	2,096	4,887	14,017	9,57	19,429	13,407
‘Nonemployed ‘74’	0.71	0.46	0.12	0.32	0.09	0.28
Earnings ‘75	1,532	3,219	13,651	9,27	19,063	13,597
Nonemployed ‘75	0.60	0.49	0.11	0.31	0.10	0.30

NOTE: Earnings variables are all expressed in 1982 dollars.

ber of data sets mimicking these NSW-CPS and NSW-PSID sets. Descriptive statistics for these data are presented in Table 1.

3 Empirical Monte Carlo Designs

3.1 The structured design

What we term a “structured” design is based on the Monte Carlo studies implemented by Abadie and Imbens (2011) and Busso et al. (2013). We test both an “uncorrelated” and a “correlated” version of this design.

First we generate a fixed number of 185 treated and either 2,490 (PSID) or 15,992 (CPS) nontreated observations per replication. We then draw nonemployment status in 1974 and 1975 jointly, with the probability of each joint nonemployment status matching the observed joint probability in the data for individuals with that treatment status. For individuals who are employed in only one period, an income is drawn from a log normal distribution with mean and variance that match those in the data for individuals with the same treatment and employment status. Where individuals are employed in both periods a joint log normal distribution is used. Also, whenever drawn income in a particular year lies outside the support of income in that year observed in the data, the observation is replaced with the limit point of the support, as suggested by Busso et al. (2013) .

In our initial *uncorrelated* design we closely replicate Abadie and Imbens (2011), drawing all other covariates – black, married, education, and age – conditional only on treatment status. Black and married are binary outcomes, so draws are taken from a Bernoulli with appropriate probability of success. Age is drawn from a log normal, with matched conditional mean and conditional variance from the data. As with income, censoring is performed, replacing any generated observations which lie outside the support with the limit point of the support from the original data.

In the original data education is coded as the number of years of education completed, taking integer values four to sixteen. Since the data do not follow any smooth distribution, Abadie and Imbens (2011) use a discrete distribution with support at each possible value. Unlike them, we collapse the discrete distribution into two indicator variables, one indicating whether the individual has at least 12 years of education, and the other whether the individual has at least 16 years. These points are chosen because of the large probability masses observed at these points in the distribution. We can then match the probabilities for each of these to those in the data, conditioning on treatment status. This reduction in support is done for consistency with our correlated design, so that we could focus on the importance of using a rich correlation structure in the data generating process.[4]

In the correlated design we model the joint distribution of the covariates as a tree-structured conditional probability distribution, where the conditional distributions are learned from the data. This contrasts with the uncorrelated design where one imposes that the joint distribution is the product of the marginals conditioned only on the treatment status. We begin by deterministically assigning treatment status, and then generating employment status and

[4] We find qualitatively identical results for the uncorrelated design whether or not we perform this reduction in the support of education. These results are available on request.

income as above. The process for generating other covariates is as follows:

1. The covariates are ordered: treatment status, employment statuses, income in each period, whether black, whether married, whether received at least 12 years of education, whether received at least 16 years of education, and age.

2. Using the original data, each covariate from "black" onwards is regressed on all the covariates listed before it.[5] These regressions are not to be interpreted causally; they simply give the conditional mean of each outcome given all preceding covariates. Where coefficients are insignificantly different from zero, they are set to zero, and the other coefficients are recorded.

3. In the new (Monte Carlo) data set, covariates are drawn sequentially in the same order. For binary covariates a temporary value is drawn from a $Unif(0,1)$ distribution. Then the covariate is equal to one if the temporary value is less than the conditional probability for that observation. The conditional probability is found using the values of the existing generated covariates and the estimated coefficients from (2). Age is drawn from a log normal whose mean depends on the other covariates and whose variance is allowed to depend on treatment status, and again we replace extreme values with the limit of the support, as in the uncorrelated case.

In both designs (correlated and uncorrelated) the binary outcome, y_i, is then generated in two steps. In the first step, a probability of employment is generated conditional on the covariates, using the parameters of a logit model fitted from the original data (see Table A.1). Each covariate is included linearly within the inverse logit function, except for treatment status, which is interacted with all other covariates so that the coefficients may differ de-

[5] One exception is "at least 16 years of education" which is regressed on the prior listed covariates conditional on having at least 12 years of education, since it is clearly not possible to have at least 16 years without having at least 12.

pending on treatment status. Precisely, the estimated coefficients, γ_0 and γ_1, from estimation using the control and treatment subsamples are used to calculate the linear index, $X_i\gamma_d$ (for $d = 0,1$), from which we calculate $p_i = \Pr(y_i = 1|X_i, d_i = d) = e^{X_i\gamma_d}/(1+e^{X_i\gamma_d})$. Finally, employment status is then determined as a draw from a Bernoulli distribution with the estimated conditional probability p_i.

3.2 The placebo design

The "placebo" design follows the approach suggested by Huber et al. (2013), and applied also by Lechner and Wunsch (2013). Covariates are drawn jointly with outcomes from the empirical distribution, rather than a parametrised approximation. In particular, pairs (y_i, X_i) are drawn with replacement from the sample of nontreated observations. The data on the treated sample are used with the control data to parametrically (logit) estimate the probability an individual is treated conditional on their characteristics X_i.

We assign treatment status to observations in the sampled data using the estimated coefficients, β (see Table A.2); iid logistic errors, ε_i; and two parameters λ and α, where λ determines the degree of covariate overlap between the "placebo treated" and "nontreated" observations and α determines the expected proportion of the "placebo treated". Formally $d_i = \mathbf{1}(d_i^* > 0)$ where $d_i^* = \alpha + \lambda X_i\beta + \varepsilon_i$. Since the original outcome, y_i, is drawn directly from the data together with X_i, we do not need to specify any DGP for the outcome. Instead we know that by construction the effect of the assigned treatment status is zero.[6] Hence we can judge estimators based on their ability to replicate this true effect of zero.

This design requires some choice of α and λ. We choose α to ensure that the proportion of the "placebo treated" in each simulated sample is as close as possible to the propor-

[6] A similar approach was previously developed by Bertrand et al. (2004) who studied inference in difference-in-differences (DiD) designs using simulations with randomly generated "placebo laws" in state-level data (i.e. policy changes which never actually happened). See also Hansen (2007), Cameron et al. (2008), and Brewer et al. (2013) for follow-up studies.

tion of treated in the corresponding original data set (1.14% in NSW-CPS and 6.92% in NSW-PSID). Huber et al. (2013) suggest that choosing $\lambda = 1$ should guarantee "selection [into treatment] that corresponds roughly to the one in our 'population'". However, this is not necessarily true: it would be true only if the degree of overlap between the treated and nontreated in the original data was roughly equal to the degree of overlap between the placebo treated and placebo nontreated in the simulated samples. There is no reason to expect such a relationship, so we conduct a small-scale calibration to determine the "optimal" value of λ in these data.

We choose a search grid of possible values for λ, using $\{0.01, 0.03, \dots, 0.99\}$ for NSW-CPS and $\{0.01, 0.02, \dots, 0.99\}$ for the smaller NSW-PSID. For each value we generate data and calculate "overlap" for each sample, which we define to be the proportion of treated individuals for whom the estimated propensity score is larger than the minimum and smaller than the maximum estimated propensity score among the nontreated. We perform 100 replications for each λ in NSW-CPS and 500 in NSW-PSID. We choose this λ which minimises the root-mean-square deviation of our simulated overlap from the one in the original data. This gives $\lambda = 0.51$ in the NSW-CPS and $\lambda = 0.19$ in the NSW-PSID. As a comparison with Huber et al. (2013), however, we also perform simulations with $\lambda = 1$, and we refer to these two versions of the placebo design as *calibrated* and *uncalibrated*, respectively.

4 Estimators

As mentioned above, in this paper we reverse the usual ordering, using a number of estimators to compare different types of empirical Monte Carlo designs, rather than using the generated data to rank estimators. We implement many common estimators to see how good the various designs are at replicating the true biases, absolute biases, and corresponding rankings. Generally, we consider treatment effect estimators which belong to one of five main classes: standard parametric (regression-based), flexible parametric (Oaxaca–Blinder), kernel-based (matching, local linear regression, and local logit), nearest-neighbour matching, and inverse probability weighting estimators.

In each case we estimate the average treatment effect on the treated (ATT) using these estimators, and then calculate the bias for each replication via a comparison to an "oracle" estimator which provides the true value. In the placebo design, the true value in the population is equal to zero by construction. In the structured design, we use our knowledge of both the potential outcome equations to compute the probability of success under both regimes for each individual. The true value is then obtained by averaging the difference between these two probabilities over the subpopulation of treated individuals.

In particular, we use as regression-based estimators the linear probability model (LPM) as well as the logit, probit, and complementary log-log models. The complementary log-log model is a parametric estimator using an asymmetric binary link function, which makes it more appropriate when the probability of success takes values close to zero or one (see, e.g., Cameron and Trivedi 2005 for a textbook treatment), as is the case in our application. We also follow Kline (2011) in using the Oaxaca–Blinder (OB) decomposition to compute the ATT.[7] Since we consider a binary outcome, we apply both linear and non-linear OB estimators. The linear OB decomposition is equivalent to the LPM but with the treatment dummy interacted with appropriately demeaned covariates. Similarly, the non-linear OB decompositions impose either a logit or probit link function around the linear index, separately for both subpopulations of interest (see, e.g., Yun 2004 and Fairlie 2005).

[7] Kline (2011) has shown that the OB decomposition is equivalent to a particular reweighting estimator and that it therefore satisfies the property of double robustness. See also Oaxaca (1973) and Blinder (1973) for seminal formulations of this method as well as Fortin et al. (2011) for a recent review of the decomposition framework.

Turning to more standard treatment effect estimators, we consider several kernel-based methods, in particular kernel matching, local linear regression, and local logit. Kernel matching estimators have played a prominent role in the programme evaluation literature (see, e.g., Heckman et al. 1997 and Frölich 2004), and their asymptotic properties were established by Heckman et al. (1998). Similarly, local linear regression was studied by Fan (1992, 1993), Heckman et al. (1998), and others. Because our outcome is binary, we also consider local logit, as applied in Frölich and Melly (2010). Note that each of these estimators requires estimating the propensity score in the first step (based on a logit model) as well as choosing a bandwidth. For each of the methods, we select the bandwidth on the basis of leave-one-out cross-validation (see, e.g., Busso et al. 2009 and Huber et al. 2013) from a search grid $0.005 \times 1.25^{g-1}$ for $g = 1, 2, \dots, 15$, and repeat this process in each replication.[8]

We also apply the popular nearest-neighbour matching estimators, including both matching on covariates and on the estimated propensity score. Large sample properties for some of these estimators were derived by Abadie and Imbens (2006). Since nearest-neighbour matching estimators were shown not to be $\sqrt{n}$-consistent in general, we also consider the bias-adjusted variant of both versions of matching (Abadie and Imbens 2011). Like kernel-based methods, also nearest-neighbour matching estimators require choosing a tuning parameter, N, i.e. the number of neighbours. We consider the workhorse case of $N = 1$ (pair matching) and also $N = 40$,[9] so we apply eight nearest-neighbour matching estimators in total.

[8] Note that the computation time is already quite large in the case of the NSW-PSID data, but it is completely prohibitive for NSW-CPS. Consequently, in the case of the NSW-CPS data set, we calculate optimal bandwidths only once, for the original data set, and use these values in our simulations. We find qualitatively identical results for the NSW-CPS data set when we exclude all the kernel-based estimators. These results are available on request.

[9] While the latter number of matches is unusually big, results from the early stage of this project suggested a negative monotonic relationship between N and the root-mean-square error (RMSE) of an estimator (in the range 1–64).

The last class of estimators includes three versions of inverse probability weighting (see Busso et al. 2009 for a thorough discussion) as well as the so-called double robust regression (see, e.g., Robins et al. 1994, Robins and Rotnitzky 1995, and Busso et al. 2009). We consider unnormalised reweighting, in which the sum of weights is stochastic; normalised reweighting, in which the weights are rescaled to sum to 1; as well as (asymptotically) efficient reweighting, which is a linear combination of normalised and unnormalised reweighting (Lunceford and Davidian 2004). Also, the double robust regression is in practice a combination of regression and reweighting, and the resulting estimator is consistent if at least one of the two models is well-specified (see Imbens and Wooldridge 2009 for a discussion).

Moreover, for regression-based, Oaxaca–Blinder, and inverse probability weighting estimators we also consider a separate case in which we restrict our estimation procedures to those treated (or placebo treated) whose estimated propensity scores are larger than the minimum and smaller than the maximum estimated propensity score among the nontreated, i.e. to those who are located in the common support region. In consequence, our total number of estimators is equal to 35, including 8 regression-based estimators, 6 Oaxaca–Blinder estimators, 5 kernel-based estimators, 8 nearest-neighbour matching estimators, and 8 inverse probability weighting estimators.[10]

5 Results

Empirical Monte Carlo studies (EMCS) seek to persuade one of the benefits of using a particular estimator, showing that it is preferred to many others in a particular circumstance. Here we are able to test the internal validity of such a procedure, by comparing the biases

[10] We perform our simulations in Stata and use several user-written commands in our estimation procedures: `locreg` (Frölich and Melly 2010), `nnmatch` (Abadie et al. 2004), `oaxaca` (Jann 2008), and `psmatch2` (Leuven and Sianesi 2003).

produced by estimators from the original data with those produced using the Monte Carlo data. Since we know the true effects in the original data – the programme reduced nonemployment among its participants by 11.06 percentage points – and the generated data, we can calculate both of these biases.

Typically one would choose estimators on the basis of minimising either the root-mean-square error (RMSE) or the absolute bias. Minimising the RMSE accounts for both the bias and variance of an estimator, so might be the preferred measure for an analyst in many contexts. Unfortunately from a single sample of data it is only possible to measure the bias of an estimator, not the variance of the estimates produced. However, a minimum condition for an EMCS to be able to reproduce the appropriate RMSE is that it should produce the correct biases, and absolute biases. Hence we look at these metrics, comparing the correlation in bias, absolute bias, and in the ranking of estimators by absolute bias between the various EMCS procedures and the original data.[11]

5.1 The structured design

In this subsection we report simulation results for the uncorrelated and correlated structured designs, and comment on their ability to replicate the "true ranking" of various nonexperimental estimators for the average treatment effect on the treated.[12]

The baseline correlations in the NSW-PSID design are shown in the first and third columns of Table 2. Mean biases are positively and significantly correlated with the true biases, whilst absolute mean biases are significantly negatively correlated with the true absolute

[11] In order to reduce the impact of outliers on our final results, we discard all the estimates whose absolute value is larger than 10. Note that the outcome in our application is binary, so the true effect cannot deviate from the $[-1, 1]$ interval. Our rule should not therefore be viewed as particularly restrictive.

[12] Tables B.1 and B.2 present "true" biases and rankings of these estimators. Table B.3 provides evidence on their relative performance in the uncorrelated structured design, when the DGP attempts to mimic the NSW-CPS data-generating process; similarly Table B.4 provides the results for the NSW-PSID case. Tables B.5 and B.6 present simulation results for the correlated structured design.

Table 2: Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-PSID Data Set

	"True biases"				"Hypothetical biases"			
	Uncorrelated		Correlated		Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Correlations								
Bias–Mean bias	0.371**	0.256	0.643***	0.549***	0.371**	0.256	0.643***	0.549***
	(0.031)	(0.189)	(0.000)	(0.002)	(0.031)	(0.189)	(0.000)	(0.002)
Abs. bias–Abs. mean bias	–0.363**	–0.217	–0.435***	–0.216	0.408**	0.297	0.698***	0.616***
	(0.035)	(0.267)	(0.009)	(0.260)	(0.017)	(0.125)	(0.000)	(0.000)
Rank–Rank	–0.357**	–0.169	–0.380**	–0.142	0.408**	0.222	0.693***	0.599***
	(0.038)	(0.391)	(0.025)	(0.461)	(0.017)	(0.256)	(0.000)	(0.001)
Sample restrictions								
Exclude outliers	Y	Y	Y	Y	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y	N	Y	N	Y
Number of estimators	34	28	35	29	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first and fifth columns).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

biases. The second and fourth columns test for robustness of this result to the exclusion of all the Oaxaca–Blinder estimators, since the logit OB decomposition can be regarded as the "true" model for the structured design, which might improve the performance of various OB decompositions in such designs in an artificial way. Although the correlations generally get weaker, and in some cases become insignificant as the number of estimators falls, the signs are unchanged.[13]

The positive correlation in bias implies that estimators which have relatively high biases in the original data continue to have relatively high biases in the simulations. Since bias is calculated as the difference between the estimate and a constant, this positive correlation in biases simply reflects a positive correlation in the underlying estimates.

However, for a researcher performing an empirical Monte Carlo study the appropriate decision criterion to choose estimators is *absolute* bias, and on this criterion the researcher would choose the wrong estimators. This result here differs from unadjusted bias because when taking absolute values it becomes important what value is used as the constant "true"

[13] We also perform additional robustness checks, such as reweighting the effect of each estimator-observation on our correlations in a way which would guarantee an equal impact of each of the classes of estimators. Since these additional robustness checks never have an effect on our conclusions, we do not report the results here. These results are available on request.

Table 3: Correlations Between the Biases in the Uncorrelated and Correlated Structured Designs and in the Original NSW-CPS Data Set

	“True biases”				“Hypothetical biases”			
	Uncorrelated		Correlated		Uncorrelated		Correlated	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Correlations								
Bias–Mean bias	0.390**	0.259	0.530***	0.379**	0.390**	0.259	0.530***	0.379**
	(0.023)	(0.184)	(0.001)	(0.042)	(0.023)	(0.184)	(0.001)	(0.042)
Abs. bias–Abs. mean bias	0.458***	0.420**	0.396**	0.333*	0.322*	0.326*	0.301*	0.290
	(0.007)	(0.026)	(0.019)	(0.078)	(0.063)	(0.090)	(0.079)	(0.128)
Rank–Rank	0.484***	0.428**	0.426**	0.334*	0.330*	0.323*	0.360**	0.330*
	(0.004)	(0.023)	(0.011)	(0.077)	(0.057)	(0.093)	(0.034)	(0.080)
Sample restrictions								
Exclude outliers	Y	Y	Y	Y	Y	Y	Y	Y
Exclude Oaxaca–Blinder	N	Y	N	Y	N	Y	N	Y
Number of estimators	34	28	35	29	34	28	35	29

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: unnormalised reweighting with the common support restriction (first and fifth columns).

*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

value against which the bias is calculated.

With the NSW-PSID data, the structured design generates true values equal to –0.2554 and –0.2604, on average, in the uncorrelated and correlated versions respectively. These are far from the true value of –0.1106 in the original data, since they are in effect based on the logit Oaxaca–Blinder decomposition, which estimates a true effect of –0.2568.

In the fifth to eighth columns of Table 2 we test the hypothesis that the structured design is informative about the ability of estimators to replicate the estimate *from the model*, rather than the true effect in the data. To do this we replace the “true effect” in the original NSW data with the effect suggested by the model, and use this to compute the corresponding hypothetical biases. Hence this transformation provides some evidence on what the results would be if the model generated the correct treatment effect.

These results are striking. Indeed, all the correlations turn positive, and most of them highly statistically significant. The results are stronger for the correlated structured design, and in that case remain significant even upon the exclusion of a number of estimators.

We further test our hypothesis that a structured empirical Monte Carlo design is informative only when the implied treatment effect is correct by applying the method to the NSW-CPS data. Here the estimated effect is equal to –0.1174, close to the true value of –0.1106.

The results in Table 3 are supportive of our interpretation. We find similar results on absolute bias in each of these, since the true effect is already close to the estimated one, and correlations are generally positive. Again the relationships get weaker, and sometimes insignificant, when we exclude all the OB estimators, but the broad picture does not seem to change.

Hence a structured Monte Carlo design is able to be informative about the absolute bias of an estimator only under the assumption that the true effect is equal to the estimated effect used in the data generating process. However, this assumption is not testable. Further, if one were to take this assumption seriously there would be no reason to use any Monte Carlo procedure, since the true effect would already be known.

5.2 The placebo design

In light of our earlier findings, the placebo design offers some hope that it should provide valuable information to the empirical researcher, since it is clear that the treatment effect here is always known. However, the results in Table 4 show that this procedure is unable to even generally replicate the biases from the true data, with significant negative correlations in many cases, and no correlation in absolute bias.[14] Whilst the calibrated design does always dominate the uncalibrated design, it remains unable to provide useful guidance on the choice of estimators.[15]

Although the placebo design avoids the problem of needing to correctly specify a parametric model for the outcome, the treatment effect is now clearly different from that in the original data. Additionally, only a subset of the original data are used. To the extent that

[14] This design has no "optimal estimator", so we do not include the additional columns we had in the earlier tables.

[15] Simulation results are presented in Tables B.7 and B.8 (uncalibrated placebo design) as well as Tables B.9 and B.10 (calibrated placebo design).

Table 4: **Correlations Between the Biases in the Uncalibrated and Calibrated Placebo Designs and in the Original NSW-CPS and NSW-PSID Data Sets**

	Uncalibrated		Calibrated	
	NSW-PSID	*NSW-CPS*	*NSW-PSID*	*NSW-CPS*
Correlations				
Bias–Mean bias	–0.337**	–0.353**	–0.403**	0.470***
	(0.048)	(0.041)	(0.018)	(0.004)
Abs. bias–Abs. mean bias	–0.022	0.045	0.273	–0.015
	(0.900)	(0.801)	(0.119)	(0.930)
Rank–Rank	0.061	–0.187	0.351**	–0.178
	(0.730)	(0.289)	(0.042)	(0.307)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Number of estimators	35	34	34	35

NOTE: P-values are in parentheses. We define outliers as those estimators whose mean biases are more than three standard deviations away from the average mean bias. The following estimators are treated as outliers: matching on the propensity score, $N = 40$ (second column) and bias-adjusted matching on covariates, $N = 40$ (third column).
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

these control observations differ from the treated ones, this will create a second difference between the original data and our simulations. This effect is important as demonstrated by the results in Table 4. With this design it is generally not possible to even match the mean bias. Although it is partially improved through the use of calibration to better match the overlap between treated and control observations, this remains insufficient to generally solve the problem. Hence the results of this procedure are also not informative about the performance of estimators in finding the treatment effect in the original data.

6 Conclusions

In this paper we investigate the internal validity of empirical Monte Carlo studies, which we define as the ability of such simulation exercises to replicate the “true ranking” of various nonexperimental estimators for the average treatment effect on the treated. This problem is of high practical relevance, since several recent papers have put forward the idea that empirical Monte Carlo studies might provide a solution to the oft-cited design dependence of simulation exercises and their reliance on unrealistic DGPs. For example, Busso et al. (2013) suggest that empirical researchers should “conduct a small-scale simulation study designed to mimic their empirical context" in order to choose the estimator with best properties.

We consider two different empirical Monte Carlo designs. The first, which we term the "structured" design, is based on Abadie and Imbens (2011) and Busso et al. (2013). Here we generate new data which match particular features of the original data set, and then generate outcomes using parameters estimated from the original data.

We show that this method can only be informative about the true ranking of the estimators if the treatment effect in the original data is the same as that implied by the data generating process. This is clearly untestable, and if it were to be true then one would already know the treatment effect of interest, precluding the need for a simulation process. This severely limits the practical usefulness of the structured design.

We also consider the "placebo" design suggested by Huber et al. (2013). Here a sample of observations is drawn from the control data, and a placebo treatment is assigned using a parametric conditional probability of treatment estimated from the full data. Now the true treatment effect is known – it is zero by construction – and one might hope that this would solve the earlier problem.

Our results show that this method is even more problematic than the structured design. The treatment effect in simulations is still likely to be different than the true effect in the original data. Additionally since only the control observations are used, the simulated data may differ significantly from the original data, depending on the overlap in the original data. This can partly be corrected by adjusting the overlap between treated and control observations, but the support of the covariates and outcome may still be very different.

Our results are unfortunately very negative, although in line with a long-standing literature: there is unfortunately no silver bullet for researchers when choosing which estimators to use in a particular circumstance. The finite-sample performance of these estimators continues to be an important issue and finding grounds on which to judge their suitability remains an open research question. For now empirical researchers would be best advised to continue using several different approaches, as Busso et al. (2013) also suggest, and reporting these potentially varying estimates as an important robustness check.

References

A. Abadie and G. W. Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica*, 74:235–267, 2006.

A. Abadie and G. W. Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects". *Journal of Business & Economic Statistics*, 29:1–11, 2011.

A. Abadie, D. Drukker, J. L. Herr, and G. W. Imbens. "Implementing Matching Estimators for Average Treatment Effects in Stata". *Stata Journal*, 4:290–311, 2004.

P. C. Austin. "The Performance of Different Propensity-Score Methods for Estimating Differences in Proportions (Risk Differences or Absolute Risk Reductions) in Observational Studies". *Statistics in Medicine*, 29:2137–2148, 2010.

M. Bertrand, E. Duflo, and S. Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?". *Quarterly Journal of Economics*, 119:249–275, 2004.

A. S. Blinder. "Wage Discrimination: Reduced Form and Structural Estimates". *Journal of Human Resources*, 8:436–455, 1973.

R. Blundell and M. Costa Dias. "Alternative Approaches to Evaluation in Empirical Microeconomics". *Journal of Human Resources*, 44:565–640, 2009.

M. Brewer, T. F. Crossley, and R. Joyce. "Inference with Difference-in-Differences Revisited". Unpublished, 2013.

M. Busso, J. DiNardo, and J. McCrary. "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects". Unpublished, 2009.

M. Busso, J. DiNardo, and J. McCrary. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators". Unpublished, 2013.

A. C. Cameron and P. K. Trivedi. *"Microeconometrics: Methods and Applications"*. Cambridge University Press, 2005.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. "Bootstrap-Based Improvements for Inference with Clustered Errors". *Review of Economics and Statistics*, 90:414–427, 2008.

R. H. Dehejia and S. Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". *Journal of the American Statistical Association*, 94:1053–1062, 1999.

A. Diamond and J. S. Sekhon. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies". *Review of Economics and Statistics*, forthcoming, 2012.

R. W. Fairlie. "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models". *Journal of Economic and Social Measurement*, 30:305–316, 2005.

J. Fan. "Design-adaptive Nonparametric Regression". *Journal of the American Statistical Association*, 87:998–1004, 1992.

J. Fan. "Local Linear Regression Smoothers and Their Minimax Efficiencies". *Annals of Statistics*, 21:196–216, 1993.

N. Fortin, T. Lemieux, and S. Firpo. *"Decomposition Methods in Economics"*, volume 4 of *"Handbook of Labor Economics"*, pages 1–102. Elsevier, 2011.

M. Frölich. "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators". *Review of Economics and Statistics*, 86:77–90, 2004.

M. Frölich and B. Melly. "Estimation of Quantile Treatment Effects with Stata". *Stata Journal*, 10:423–457, 2010.

C. B. Hansen. "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects". *Journal of Econometrics*, 140:670–694, 2007.

J. J. Heckman and V. J. Hotz. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training". *Journal of the American Statistical Association*, 84:862–874, 1989.

J. J. Heckman, H. Ichimura, and P. E. Todd. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies*, 64:605–654, 1997.

J. J. Heckman, H. Ichimura, and P. Todd. "Matching as an Econometric Evaluation Estimator". *Review of Economic Studies*, 65:261–294, 1998.

M. Huber, M. Lechner, and C. Wunsch. "The Performance of Estimators Based on the Propensity Score". *Journal of Econometrics*, 175:1–21, 2013.

G. W. Imbens and J. M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature*, 47:5–86, 2009.

B. Jann. "The Blinder–Oaxaca Decomposition for Linear Regression Models". *Stata Journal*, 8:453–479, 2008.

P. Kline. "Oaxaca-Blinder as a Reweighting Estimator". *American Economic Review: Papers & Proceedings*, 101:532–37, 2011.

R. J. LaLonde. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*, 76:604–620, 1986.

M. Lechner and C. Wunsch. "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables". *Labour Economics*, 21:111–121, 2013.

S. Lee and Y.-J. Whang. "Nonparametric Tests of Conditional Treatment Effects". Cemmap Working Paper no. CWP36/09, 2009.

E. Leuven and B. Sianesi. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing", 2003. URL `http://ideas.repec.org/c/boc/bocode/s432001.html`. This version 4.0.6.

J. K. Lunceford and M. Davidian. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study". *Statistics in Medicine*, 23:2937–2960, 2004.

D. L. Millimet and R. Tchernis. "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies". *Journal of Business & Economic Statistics*, 27: 397–415, 2009.

R. Oaxaca. "Male-Female Wage Differentials in Urban Labor Markets". *International Economic Review*, 14:693–709, 1973.

J. M. Robins and A. Rotnitzky. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data". *Journal of the American Statistical Association*, 90:122–129, 1995.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. "Estimation of Regression Coefficients when Some Regressors Are Not Always Observed". *Journal of the American Statistical Association*, 89:846–866, 1994.

J. A. Smith and P. E. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?". *Journal of Econometrics*, 125:305–353, 2005.

S. M. Stigler. "Do Robust Estimators Work with Real Data?". *Annals of Statistics*, 5: 1055–1098, 1977.

M.-S. Yun. "Decomposing Differences in the First Moment". *Economics Letters*, 82:275–280, 2004.

Z. Zhao. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence". *Review of Economics and Statistics*, 86:91–107, 2004.

Z. Zhao. "Sensitivity of Propensity Score Methods to the Specifications". *Economics Letters*, 98:309–319, 2008.

A Potential Outcome and Treatment Equations

Table A.1 presents potential outcome equations which are used in the uncorrelated and correlated structured designs, separately for the NSW-CPS and NSW-PSID data sets as well as for the treated and nontreated subsamples (γ_1 and γ_0, respectively). These equations are based on the logit coefficients estimated using the original data sets.

Table A.1: Potential Outcome Equations in the Structured Design

	NSW-CPS		*NSW-PSID*	
	γ_1	γ_0	γ_1	γ_0
Age	-0.0068	0.0461	-0.0068	0.0335
Black	1.5818	0.0937	1.5818	-0.2514
Education-12	-0.3608	0.5363	-0.3608	-0.0056
Education-16	(omitted)	-0.0675	(omitted)	-0.1078
Married	-0.6001	0.2558	-0.6001	-0.2182
‘Earnings ‘74’	0.000010	-0.000034	0.000010	0.000010
‘Nonemployed ‘74’	-1.7371	0.5564	-1.7371	1.8915
Earnings ‘75	-0.000145	-0.000060	-0.000145	-0.000068
Nonemployed ‘75	1.3457	1.2479	1.3457	1.3282
Intercept	-1.6669	-3.2891	-1.6669	-2.8314

Similarly, Table A.2 presents treatment equations which are used in the uncalibrated and calibrated placebo designs, separately for the NSW-CPS and NSW-PSID data sets. Again, the coefficients are taken from logit models estimated using the original data sets.

Table A.2: Treatment Equations in the Placebo Design

	NSW-CPS	*NSW-PSID*
Age	–0.0266	–0.1136
Black	3.8887	2.1466
Education	–0.1072	–0.1366
Married	–0.9979	–1.6143
‘Earnings ‘74’	0.000063	0.000024
‘Nonemployed ‘74’	1.6595	3.1840
Earnings ‘75	–0.000180	–0.000276
Nonemployed ‘75	0.1821	–1.2951
Intercept	–3.8391	2.7444

B The Performance of Individual Estimators

B.1 The true ranking

Table B.1 presents nonexperimental estimates of the effect of the NSW programme using the NSW-CPS data set and 35 various nonexperimental estimators. Generally, the estimators perform very well, with the average bias being slightly smaller than 0.01 (less than 9% of the absolute value of the "true effect"). Several regression-based estimators perform best, especially the complementary log-log and logit models. Also, the logit OB decomposition performs very well, as do selected bias-adjusted nearest-neighbour matching estimators. Inverse probability weighting and kernel-based estimators (especially local linear regression and local logit) perform relatively badly, although the corresponding biases can still be regarded as quite low.

Similarly, Table B.2 presents analogous estimates and rankings on the basis of the NSW-PSID data set. The average bias is now much larger than in the previous case (and equal to –0.047), and many estimators, especially all variants of the OB decomposition, suffer from large (absolute) biases in the order of 0.08–0.17. On the other hand, unnormalised reweighting as well as selected nearest-neighbour matching and kernel-based estimators (especially matching with the Gaussian kernel and local logit) perform best. Note that the correlation between the rankings in Tables B.1 and B.2 is insignificant and close to zero.

Table B.1: Nonexperimental Estimates for the NSW-CPS Data

	Comsup?	Estimate	Bias	Rank
Regression-based				
Linear probability		–0.1331	–0.0225	23
Linear probability	X	–0.1293	–0.0187	16
Logit		–0.1076	0.0030	3
Logit	X	–0.1060	0.0047	5
Probit		–0.1002	0.0104	9
Probit	X	–0.0978	0.0128	12
Complementary log-log		–0.1125	–0.0019	2
Complementary log-log	X	–0.1117	–0.0011	1
Oaxaca–Blinder				
Linear probability		–0.1358	–0.0252	26
Linear probability	X	–0.1317	–0.0211	22
Logit		–0.1174	–0.0068	6
Logit	X	–0.1152	–0.0046	4
Probit		–0.1249	–0.0143	13
Probit	X	–0.1222	–0.0116	10
Kernel-based				
Kernel matching, uniform		–0.0962	0.0144	14
Kernel matching, Gaussian		–0.0912	0.0194	19
Kernel matching, Epan.		–0.0876	0.0230	24
Local linear regression		–0.0719	0.0387	34
Local logit		–0.0709	0.0397	35
Matching				
On pscore, $N = 1$		–0.0805	0.0302	28
On pscore, $N = 40$		–0.0859	0.0247	25
On pscore, $N = 1$, bias-adj.		–0.1208	–0.0102	8
On pscore, $N = 40$, bias-adj.		–0.0897	0.0209	21
On covs, $N = 1$		–0.1277	–0.0171	15
On covs, $N = 40$		–0.0749	0.0357	33
On covs, $N = 1$, bias-adj.		–0.1223	–0.0117	11
On covs, $N = 40$, bias-adj.		–0.1019	0.0087	7
Weighting				
Unnormalised		–0.0826	0.0280	27
Unnormalised	X	–0.0905	0.0201	20
Normalised		–0.0793	0.0313	29
Normalised	X	–0.0781	0.0325	31
Efficient		–0.0793	0.0313	30
Efficient	X	–0.078	0.0326	32
Double robust		–0.0913	0.0193	18
Double robust	X	–0.0914	0.0192	17

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on absolute bias.

Table B.2: Nonexperimental Estimates for the NSW-PSID Data

	Comsup?	Estimate	Bias	Rank
Regression-based				
Linear probability		–0.2030	–0.0924	25
Linear probability	X	–0.2017	–0.0911	24
Logit		–0.1941	–0.0835	22
Logit	X	–0.1944	–0.0838	23
Probit		–0.1527	–0.0421	15
Probit	X	–0.1525	–0.0419	14
Complementary log-log		–0.1900	–0.0794	19
Complementary log-log	X	–0.1909	–0.0803	20
Oaxaca–Blinder				
Linear probability		–0.2721	–0.1615	34
Linear probability	X	–0.2701	–0.1595	33
Logit		–0.2568	–0.1462	30
Logit	X	–0.2553	–0.1447	28
Probit		–0.2590	–0.1484	32
Probit	X	–0.2576	–0.1470	31
Kernel-based				
Kernel matching, uniform		–0.1507	–0.0401	12
Kernel matching, Gaussian		–0.0957	0.0149	4
Kernel matching, Epan.		–0.1504	–0.0398	11
Local linear regression		–0.2811	–0.1705	35
Local logit		–0.0842	0.0264	7
Matching				
On pscore, $N = 1$		–0.0703	0.0403	13
On pscore, $N = 40$		–0.0878	0.0228	6
On pscore, $N = 1$, bias-adj.		–0.1381	–0.0275	10
On pscore, $N = 40$, bias-adj.		–0.1914	–0.0808	21
On covs, $N = 1$		–0.1279	–0.0173	5
On covs, $N = 40$		–0.2554	–0.1448	29
On covs, $N = 1$, bias-adj.		–0.1240	–0.0134	3
On covs, $N = 40$, bias-adj.		–0.1789	–0.0683	18
Weighting				
Unnormalised		–0.1110	–0.0004	1
Unnormalised	X	–0.1129	–0.0023	2
Normalised		–0.0142	0.0964	26
Normalised	X	–0.0102	0.1004	27
Efficient		–0.0839	0.0267	9
Efficient	X	–0.0841	0.0266	8
Double robust		–0.0531	0.0575	16
Double robust	X	–0.0518	0.0588	17

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on absolute bias.

B.2 The structured design

Table B.3: Simulation Results for the Uncorrelated Structured Design (NSW-CPS)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		–0.0453	0.0550	0.0362	31
Linear probability	X	–0.0417	0.0547	0.0399	28
Logit		0.0062	0.0282	0.0311	10
Logit	X	0.0025	0.0283	0.0314	4
Probit		0.0127	0.0317	0.0331	18
Probit	X	0.0130	0.0334	0.0344	19
Complementary log-log		0.0113	0.0268	0.0269	16
Complementary log-log	X	0.0058	0.0247	0.0262	9
Oaxaca–Blinder					
Linear probability		–0.0471	0.0567	0.0365	32
Linear probability	X	–0.0424	0.0555	0.0404	29
Logit		–0.0001	0.0353	0.0394	2
Logit	X	–0.0089	0.0397	0.0425	13
Probit		–0.0102	0.0359	0.0387	14
Probit	X	–0.0160	0.0412	0.0420	21
Kernel-based					
Kernel matching, uniform		–0.0041	0.1093	0.1105	7
Kernel matching, Gaussian		0.0286	0.1817	0.1802	26
Kernel matching, Epan.		–0.0043	0.1098	0.1109	8
Local linear regression		–0.0080	0.4852	0.4856	12
Local logit		0.0225	0.1841	0.1834	22
Matching					
On pscore, $N = 1$		0.0247	0.1845	0.1837	25
On pscore, $N = 40$		0.0447	0.0790	0.0660	30
On pscore, $N = 1$, bias-adj.		–0.0041	0.1611	0.1621	6
On pscore, $N = 40$, bias-adj.		0.0000	0.0929	0.0942	1
On covs, $N = 1$		–0.0148	0.1401	0.1400	20
On covs, $N = 40$		–0.0595	0.0734	0.0472	33
On covs, $N = 1$, bias-adj.		–0.0074	0.1681	0.1688	11
On covs, $N = 40$, bias-adj.		–0.0035	0.0927	0.0936	5
Weighting					
Unnormalised		–0.0772	0.6085	0.6046	34
Unnormalised	X	–0.1923	0.6293	0.6003	35
Normalised		0.0236	0.1827	0.1826	24
Normalised	X	–0.0019	0.1805	0.1818	3
Efficient		0.0291	0.1504	0.1486	27
Efficient	X	0.0116	0.1180	0.1182	17
Double robust		0.0226	0.1453	0.1450	23
Double robust	X	–0.0106	0.1359	0.1369	15

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

Table B.4: Simulation Results for the Uncorrelated Structured Design (NSW-PSID)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		0.0501	0.0644	0.0470	14
Linear probability	X	0.0604	0.1270	0.1135	19
Logit		0.0379	0.0631	0.0538	12
Logit	X	0.1037	0.1287	0.0767	26
Probit		0.0692	0.0850	0.0535	21
Probit	X	0.1208	0.1442	0.0798	31
Complementary log-log		0.0459	0.0654	0.0474	13
Complementary log-log	X	0.1135	0.1290	0.0605	30
Oaxaca–Blinder					
Linear probability		0.0003	0.0416	0.0482	1
Linear probability	X	0.0542	0.1212	0.1102	15
Logit		0.0035	0.0629	0.0671	3
Logit	X	0.0562	0.1084	0.0948	17
Probit		0.0007	0.0574	0.0621	2
Probit	X	0.0555	0.1072	0.0938	16
Kernel-based					
Kernel matching, uniform		0.0700	0.3682	0.3622	22
Kernel matching, Gaussian		0.1058	0.3829	0.3689	27
Kernel matching, Epan.		0.0685	0.3720	0.3664	20
Local linear regression		0.1126	0.8083	0.8012	29
Local logit		0.0879	0.4511	0.4435	23
Matching					
On pscore, $N = 1$		0.0943	0.4549	0.4459	24
On pscore, $N = 40$		0.2080	0.2267	0.0914	32
On pscore, $N = 1$, bias-adj.		0.0095	1.5200	1.5208	7
On pscore, $N = 40$, bias-adj.		0.0054	0.3691	0.3706	5
On covs, $N = 1$		–0.0099	0.1574	0.1600	8
On covs, $N = 40$		0.0082	0.0458	0.0515	6
On covs, $N = 1$, bias-adj.		–0.0130	0.4554	0.4566	9
On covs, $N = 40$, bias-adj.		0.0044	0.1595	0.1613	4
Weighting					
Unnormalised		0.2319	0.8523	0.8207	34
Unnormalised	X	–0.5662	1.4889	1.3778	35
Normalised		0.1105	0.3366	0.3189	28
Normalised	X	0.0307	0.3256	0.3250	11
Efficient		0.1012	0.3854	0.3729	25
Efficient	X	0.0562	0.2548	0.2491	18
Double robust		0.2245	0.4953	0.4419	33
Double robust	X	0.0240	0.2788	0.2789	10

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

Table B.5: Simulation Results for the Correlated Structured Design (NSW-CPS)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		–0.0228	0.0398	0.0376	32
Linear probability	X	–0.0222	0.0396	0.0379	30
Logit		0.0075	0.0279	0.0310	17
Logit	X	0.0072	0.0278	0.0310	15
Probit		0.0206	0.0355	0.0335	28
Probit	X	0.0208	0.0357	0.0336	29
Complementary log-log		0.0068	0.0232	0.0248	13
Complementary log-log	X	0.0063	0.0229	0.0247	12
Oaxaca–Blinder					
Linear probability		–0.0253	0.0416	0.0380	34
Linear probability	X	–0.0247	0.0414	0.0383	33
Logit		–0.0016	0.0350	0.0399	4
Logit	X	–0.0022	0.0352	0.0401	7
Probit		–0.0072	0.0353	0.0395	14
Probit	X	–0.0075	0.0355	0.0398	19
Kernel-based					
Kernel matching, uniform		0.0110	0.0528	0.0549	23
Kernel matching, Gaussian		0.0147	0.0632	0.0638	26
Kernel matching, Epan.		0.0099	0.0527	0.0552	21
Local linear regression		0.0407	0.3824	0.3806	35
Local logit		0.0074	0.0676	0.0698	16
Matching					
On pscore, $N = 1$		0.0062	0.0682	0.0702	11
On pscore, $N = 40$		0.0123	0.0477	0.0500	24
On pscore, $N = 1$, bias-adj.		0.0008	0.0647	0.0675	1
On pscore, $N = 40$, bias-adj.		0.0021	0.0478	0.0516	6
On covs, $N = 1$		–0.0044	0.0696	0.0716	9
On covs, $N = 40$		–0.0227	0.0490	0.0474	31
On covs, $N = 1$, bias-adj.		–0.0033	0.0727	0.0748	8
On covs, $N = 40$, bias-adj.		–0.0013	0.0503	0.0538	3
Weighting					
Unnormalised		–0.0134	0.0887	0.0895	25
Unnormalised	X	–0.0174	0.0866	0.0866	27
Normalised		0.0075	0.0635	0.0651	18
Normalised	X	0.0055	0.0632	0.0650	10
Efficient		0.0104	0.0525	0.0544	22
Efficient	X	0.0086	0.0521	0.0545	20
Double robust		0.0010	0.0563	0.0587	2
Double robust	X	–0.0019	0.0560	0.0583	5

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

Table B.6: Simulation Results for the Correlated Structured Design (NSW-PSID)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		0.0673	0.0793	0.0495	13
Linear probability	X	0.0963	0.1295	0.0897	17
Logit		0.0820	0.1012	0.0643	16
Logit	X	0.1167	0.1380	0.0761	19
Probit		0.1212	0.1338	0.0620	21
Probit	X	0.1514	0.1673	0.0739	23
Complementary log-log		0.0780	0.0940	0.0549	15
Complementary log-log	X	0.1113	0.1257	0.0589	18
Oaxaca–Blinder					
Linear probability		–0.0134	0.0441	0.0492	5
Linear probability	X	0.0568	0.1112	0.0985	10
Logit		–0.0007	0.0658	0.0702	1
Logit	X	0.0622	0.1120	0.0962	12
Probit		–0.0048	0.0600	0.0646	2
Probit	X	0.0596	0.1088	0.0941	11
Kernel-based					
Kernel matching, uniform		0.2795	0.4367	0.3359	32
Kernel matching, Gaussian		0.2523	0.3901	0.2983	28
Kernel matching, Epan.		0.2822	0.4360	0.3327	33
Local linear regression		0.1273	0.8406	0.8308	22
Local logit		0.2792	0.4364	0.3360	31
Matching					
On pscore, $N = 1$		0.2838	0.4382	0.3345	34
On pscore, $N = 40$		0.2161	0.2318	0.0843	25
On pscore, $N = 1$, bias-adj.		0.0674	1.2470	1.2449	14
On pscore, $N = 40$, bias-adj.		0.0053	0.2988	0.3001	3
On covs, $N = 1$		–0.0257	0.1478	0.1485	9
On covs, $N = 40$		0.0162	0.0466	0.0504	6
On covs, $N = 1$, bias-adj.		–0.0199	0.4232	0.4241	8
On covs, $N = 40$, bias-adj.		–0.0062	0.1356	0.1384	4
Weighting					
Unnormalised		0.2457	0.9771	0.9469	27
Unnormalised	X	–0.0176	1.1270	1.1275	7
Normalised		0.3065	0.4483	0.3274	35
Normalised	X	0.2741	0.4357	0.3389	30
Efficient		0.2678	0.6234	0.5639	29
Efficient	X	0.2162	0.5945	0.5546	26
Double robust		0.1541	0.2989	0.2576	24
Double robust	X	0.1203	0.2870	0.2622	20

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

B.3 The placebo design

Table B.7: Simulation Results for the Uncalibrated Placebo Design (NSW-CPS)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		–0.0099	0.0348	0.0334	22
Linear probability	X	0.0036	0.0375	0.0374	8
Logit		0.0057	0.0358	0.0353	17
Logit	X	0.0118	0.0398	0.0380	28
Probit		0.0058	0.0359	0.0355	18
Probit	X	0.0132	0.0405	0.0383	30
Complementary log-log		0.0065	0.0353	0.0347	19
Complementary log-log	X	0.0109	0.0386	0.0370	24
Oaxaca–Blinder					
Linear probability		–0.0107	0.0353	0.0337	23
Linear probability	X	0.0037	0.0381	0.0379	9
Logit		0.0049	0.0360	0.0357	14
Logit	X	0.0116	0.0401	0.0384	27
Probit		0.0023	0.0352	0.0352	5
Probit	X	0.0113	0.0398	0.0382	26
Kernel-based					
Kernel matching, uniform		–0.0022	0.0657	0.0657	4
Kernel matching, Gaussian		–0.0014	0.1066	0.1067	3
Kernel matching, Epan.		–0.0031	0.0652	0.0652	7
Local linear regression		–0.0110	0.2786	0.2787	25
Local logit		–0.0048	0.1105	0.1105	13
Matching					
On pscore, $N = 1$		–0.0079	0.1102	0.1101	21
On pscore, $N = 40$		–0.0610	0.0846	0.0586	35
On pscore, $N = 1$, bias-adj.		–0.0055	0.1012	0.1011	16
On pscore, $N = 40$, bias-adj.		–0.0379	0.0764	0.0664	33
On covs, $N = 1$		–0.0029	0.0947	0.0947	6
On covs, $N = 40$		–0.0376	0.0598	0.0466	32
On covs, $N = 1$, bias-adj.		0.0072	0.0991	0.0989	20
On covs, $N = 40$, bias-adj.		–0.0225	0.0661	0.0622	31
Weighting					
Unnormalised		–0.0043	0.1042	0.1042	11
Unnormalised	X	–0.0397	0.1157	0.1087	34
Normalised		–0.0040	0.0953	0.0953	10
Normalised	X	0.0043	0.0951	0.0951	12
Efficient		–0.0054	0.0938	0.0938	15
Efficient	X	–0.0011	0.0824	0.0825	2
Double robust		–0.0007	0.0872	0.0873	1
Double robust	X	–0.0126	0.0853	0.0845	29

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

Table B.8: Simulation Results for the Uncalibrated Placebo Design (NSW-PSID)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		0.0224	0.0386	0.0315	18
Linear probability	X	0.0346	0.0487	0.0342	26
Logit		0.0305	0.0485	0.0376	24
Logit	X	0.0359	0.0524	0.0381	28
Probit		0.0389	0.0532	0.0363	30
Probit	X	0.0439	0.0577	0.0374	32
Complementary log-log		0.0266	0.0445	0.0357	20
Complementary log-log	X	0.0293	0.0449	0.0340	22
Oaxaca–Blinder					
Linear probability		0.0120	0.0354	0.0333	11
Linear probability	X	0.0350	0.0495	0.0350	27
Logit		0.0142	0.0410	0.0385	12
Logit	X	0.0339	0.0500	0.0367	25
Probit		0.0218	0.0427	0.0367	17
Probit	X	0.0401	0.0539	0.0360	31
Kernel-based					
Kernel matching, uniform		0.0018	0.0705	0.0705	6
Kernel matching, Gaussian		0.0042	0.1581	0.1581	7
Kernel matching, Epan.		–0.0005	0.0692	0.0692	2
Local linear regression		0.0173	0.5220	0.5219	14
Local logit		–0.0018	0.1619	0.1619	5
Matching					
On pscore, $N = 1$		–0.0015	0.1619	0.1620	4
On pscore, $N = 40$		–0.0300	0.0668	0.0597	23
On pscore, $N = 1$, bias-adj.		–0.0244	0.1382	0.1361	19
On pscore, $N = 40$, bias-adj.		–0.0378	0.0868	0.0782	29
On covs, $N = 1$		–0.0275	0.0626	0.0562	21
On covs, $N = 40$		0.0179	0.0384	0.0340	15
On covs, $N = 1$, bias-adj.		–0.0510	0.1063	0.0933	33
On covs, $N = 40$, bias-adj.		–0.0896	0.1037	0.0522	35
Weighting					
Unnormalised		0.0207	0.2565	0.2558	16
Unnormalised	X	–0.0767	0.2779	0.2671	34
Normalised		–0.0013	0.1231	0.1231	3
Normalised	X	0.0090	0.1224	0.1221	10
Efficient		0.0002	0.1228	0.1228	1
Efficient	X	0.0060	0.0838	0.0836	9
Double robust		0.0143	0.1065	0.1055	13
Double robust	X	–0.0043	0.1019	0.1018	8

NOTE: "Comsup?" denotes the estimates which are obtained after removing all the treated observations from outside the common support region. "Rank" is based on the absolute value of mean bias.

Table B.9: Simulation Results for the Calibrated Placebo Design (NSW-CPS)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		–0.0241	0.0396	0.0315	31
Linear probability	X	–0.0247	0.0401	0.0316	33
Logit		–0.0084	0.0323	0.0313	20
Logit	X	–0.0093	0.0327	0.0313	22
Probit		–0.0103	0.0330	0.0314	25
Probit	X	–0.0113	0.0334	0.0315	26
Complementary log-log		–0.0054	0.0309	0.0304	15
Complementary log-log	X	–0.0062	0.0312	0.0306	17
Oaxaca–Blinder					
Linear probability		–0.0245	0.0401	0.0317	32
Linear probability	X	–0.0251	0.0405	0.0318	35
Logit		–0.0091	0.0332	0.0320	21
Logit	X	–0.0101	0.0336	0.0321	24
Probit		–0.0122	0.0342	0.0320	27
Probit	X	–0.0132	0.0347	0.0321	28
Kernel-based					
Kernel matching, uniform		0.0075	0.0381	0.0374	19
Kernel matching, Gaussian		0.0161	0.0422	0.0390	30
Kernel matching, Epan.		0.0051	0.0373	0.0369	14
Local linear regression		0.0031	0.4097	0.4101	12
Local logit		0.0096	0.0430	0.0420	23
Matching					
On pscore, $N = 1$		–0.0018	0.0406	0.0406	8
On pscore, $N = 40$		–0.0159	0.0422	0.0391	29
On pscore, $N = 1$, bias-adj.		0.0025	0.0370	0.0370	11
On pscore, $N = 40$, bias-adj.		–0.0016	0.0376	0.0376	7
On covs, $N = 1$		0.0062	0.0361	0.0356	18
On covs, $N = 40$		–0.0248	0.0437	0.0360	34
On covs, $N = 1$, bias-adj.		0.0058	0.0360	0.0355	16
On covs, $N = 40$, bias-adj.		–0.0020	0.0348	0.0348	9
Weighting					
Unnormalised		–0.0003	0.0370	0.0370	2
Unnormalised	X	–0.0039	0.0373	0.0371	13
Normalised		–0.0004	0.0369	0.0370	3
Normalised	X	–0.0014	0.0369	0.0369	6
Efficient		–0.0005	0.0371	0.0371	4
Efficient	X	–0.0012	0.0369	0.0369	5
Double robust		–0.0001	0.0362	0.0362	1
Double robust	X	–0.0022	0.0362	0.0362	10

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.

Table B.10: Simulation Results for the Calibrated Placebo Design (NSW-PSID)

	Comsup?	Mean bias	RMSE	SD	Rank
Regression-based					
Linear probability		0.0060	0.0254	0.0247	26
Linear probability	X	0.0084	0.0262	0.0248	32
Logit		0.0055	0.0270	0.0265	25
Logit	X	0.0076	0.0276	0.0265	29
Probit		0.0080	0.0273	0.0261	31
Probit	X	0.0099	0.0280	0.0262	34
Complementary log-log		0.0036	0.0253	0.0251	16
Complementary log-log	X	0.0052	0.0255	0.0250	23
Oaxaca–Blinder					
Linear probability		0.0040	0.0260	0.0257	18
Linear probability	X	0.0073	0.0267	0.0257	28
Logit		0.0023	0.0263	0.0262	14
Logit	X	0.0054	0.0266	0.0261	24
Probit		0.0051	0.0265	0.0260	22
Probit	X	0.0080	0.0271	0.0259	30
Kernel-based					
Kernel matching, uniform		0.0023	0.0284	0.0283	13
Kernel matching, Gaussian		0.0013	0.0297	0.0297	10
Kernel matching, Epan.		0.0012	0.0283	0.0283	8
Local linear regression		0.0024	0.0741	0.0740	15
Local logit		0.0071	0.0322	0.0314	27
Matching					
On pscore, $N = 1$		–0.0051	0.0334	0.0330	21
On pscore, $N = 40$		–0.0018	0.0282	0.0281	11
On pscore, $N = 1$, bias-adj.		–0.0006	0.0277	0.0277	2
On pscore, $N = 40$, bias-adj.		–0.0092	0.0291	0.0276	33
On covs, $N = 1$		0.0047	0.0262	0.0258	20
On covs, $N = 40$		0.0039	0.0262	0.0259	17
On covs, $N = 1$, bias-adj.		0.0042	0.0264	0.0260	19
On covs, $N = 40$, bias-adj.		–0.0142	0.0303	0.0268	35
Weighting					
Unnormalised		0.0011	0.0275	0.0275	7
Unnormalised	X	–0.0008	0.0276	0.0276	4
Normalised		0.0010	0.0275	0.0275	5
Normalised	X	0.0012	0.0274	0.0274	9
Efficient		0.0008	0.0275	0.0275	3
Efficient	X	0.0020	0.0272	0.0271	12
Double robust		0.0010	0.0274	0.0274	6
Double robust	X	0.0001	0.0273	0.0273	1

NOTE: “Comsup?” denotes the estimates which are obtained after removing all the treated observations from outside the common support region. “Rank” is based on the absolute value of mean bias.