DISCUSSION PAPER SERIES

IZA DP No. 7874

Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies

Arun Advani
Tymon Słoczyński

December 2013

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies

Arun Advani
IFS, University College London and King's College, Cambridge

Tymon Słoczyński
Michigan State University, Warsaw School of Economics and IZA

Discussion Paper No. 7874
December 2013

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7874
December 2013

ABSTRACT

Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies*

In this paper we evaluate the premise from the recent literature on 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 consider two alternative designs and provide an empirical test for both of them. We conclude that a necessary condition for the simulations to be informative about the true ranking is that the treatment effect in simulations must be equal to the (unknown) true effect. This severely limits the usefulness of such procedures, since were the effect known, the procedure would not be necessary.

JEL Classification: C15, C21, C25, C52

Keywords: empirical Monte Carlo studies, programme evaluation, treatment effects

Corresponding author:

Arun Advani
Institute for Fiscal Studies
7 Ridgmount Street
London WC1E 7AE
United Kingdom
E-mail: arun.advani@ifs.org.uk

* With thanks to Cathy Balfe, A. Colin Cameron, Mónica Costa Dias, Gil Epstein, Alfonso Flores-Lagunes, Ira Gang, Martin Huber, Michael Lechner, Justin McCrary, Blaise Melly, Mateusz Myśliwski, Anthony Strittmatter, Timothy Vogelsang, Jeffrey Wooldridge, and seminar participants at CERGE-EI, the Institute for Fiscal Studies, Michigan State University, 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. Arun Advani also acknowledges support from Programme Evaluation for Policy Analysis, a node of the National Centre for Research Methods, supported by the UK Economic and Social Research Council. Tymon Słoczyński also acknowledges a START scholarship from the Foundation for Polish Science (FNP).

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), Khwaja et al. (2011), Busso et al. (2013), Diamond and Sekhon (2013), and Huber et al. (2013) carry out Monte Carlo experiments to assess the relative finite-sample performance of a large number of estimators for various average treatment effects.[1]

Most of these recent papers use highly stylised data-generating processes (DGPs) which only loosely correspond to any actual data sets (see, *e.g.*, Frölich 2004, Busso et al. 2009). This approach is 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" (EMCS). Similar simulation exercises are carried out by Abadie and Imbens (2011) and Busso et al. (2013), who use a different procedure but again adapt it to the circumstance of interest.[2]

What is more, Busso et al. (2013) explicitly 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 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 are proposed in the literature. The first, which we term the "structured" design, is considered by both Abadie and Imbens (2011) and Busso et al. (2013). Loosely speaking, in this setting treat-

[1] Blundell and Costa Dias (2009) and Imbens and Wooldridge (2009) provide 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).

ment status and covariate values are drawn from a distribution similar to that in the data, and then outcomes are generated using parameters estimated from the data. The second approach, which we term the "placebo" design, is proposed by Huber et al. (2013). Here both covariates and outcome are drawn jointly from the control data with replacement, and treatment status is assigned using parameters estimated from the full data. Since all observations come from the control data and the original outcomes are retained, the effect of assigned treatment is known to be zero by construction.

We implement both of these approaches using the 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 (2013), and many others.[3] Since the NSW programme originally had an experimental control group, an unbiased estimate of the effect of this 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 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. Additionally, the placebo design restricts the support of the covariates to be equal to the support of the covariates amongst the control observations. Where the support differs between treated and control groups in the original data, this creates a further reason why the placebo procedure generates samples which differ from the true data-generating process. Hence the conditions under which this procedure is useful are even more stringent, although this latter issue is at least testable.

Hence we conclude that there is little support for the chief premise of the recent literature

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

on empirical Monte Carlo studies: that they are at least informative about the appropriate choice of estimators 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 a 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) suggests that one could use the design of this programme to assess the performance of various nonexperimental estimators of the average treatment effect. He discards the original control group from the NSW data and creates 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. population. LaLonde (1986) suggests 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 finds that very few of the estimates are close to the experimental benchmark. This result has 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 2013).

The key insight of LaLonde (1986) is 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) from Dehejia and Wahba (1999) as

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	9,570	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	9,270	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.

well as the original CPS and PSID comparison groups (15,992 and 2,490 observations, respectively) from LaLonde (1986), and we aim at creating a large number 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 employment status in 1974 and 1975 jointly, with the probability of each joint employment 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. Note that conditioning the distribution of covariates on treatment status means that the probability of treatment conditional on covariates is defined implicitly by this procedure. 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. Since the data do not follow any smooth distribution, Abadie and Imbens (2011) use a discrete distribution with support at each integer from four to sixteen. 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 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. This ordering is chosen purely for convenience, with binary covariates listed before continuous ones.

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;

[4]We also tested a version of the uncorrelated design using the same distribution as Abadie and Imbens (2011), without any consequential effect on our results.

[5]One exception is "at least 16 years of education" which is regressed on the prior listed covari-

they simply give the conditional mean of each variable 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 depending 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, $\boldsymbol{X}_i\gamma_d$ (for $d = 0,1$), from which we calculate $p_i = \Pr(Y_i = 1|\boldsymbol{X}_i, D_i = d) = e^{\boldsymbol{X}_i\gamma_d}/(1+e^{\boldsymbol{X}_i\gamma_d})$. In practice, this model is equivalent to a flexible parametric logit model or – equivalently – to a logit version of the Oaxaca–Blinder decomposition (see, *e.g.*, Fortin et al. 2011). In the second step, employment status is determined as a draw from a Bernoulli distribution with the estimated conditional probability p_i.

We approximate the sample-size selection rule in Huber et al. (2013), which suggests how the number of generated samples should vary with the number of observations, by generating 2,000 samples for NSW-PSID and 500 samples for NSW-CPS.

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

ates 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.

$(Y_i, \boldsymbol{X}_i)$ 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 propensity score, *i.e.* the conditional probability of treatment.

We assign treatment status to observations in the sampled data using the estimated coefficients, $\boldsymbol{\phi}$ (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 $D_i^* = \alpha + \lambda \boldsymbol{X}_i \boldsymbol{\phi} + \varepsilon_i$. Since the original outcome, Y_i, is drawn directly from the data together with $\boldsymbol{X}_i$, 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.

Of course, one should note that the conditional distribution of outcomes for placebo treated individuals might differ significantly from the conditional distribution of outcomes for treated individuals in the original data. This will affect the extent to which knowledge about the relative performance of estimators in the generated samples is informative about the relative performance of estimators in the original data.

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 proportion 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, namely $\{0.01, 0.03, \ldots, 0.99\}$ for NSW-CPS and $\{0.01, 0.02, \ldots, 0.99\}$ for the smaller NSW-PSID.[7] 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 min-

[6] A similar approach is developed by Bertrand et al. (2004) who study 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. For follow-up studies, see also Hansen (2007), Cameron et al. (2008), and Brewer et al. (2013).

[7] On the basis of a presearch, we determined that for both data sets $\lambda \in (0, 1)$.

imum 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 $\lambda = 0.17$ 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.

As before, we generate 2,000 samples for NSW-PSID and 500 samples for the larger NSW-CPS.

4 Method

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. We discuss below the estimators which we use, and the metrics on which we compare the EMCS methods.

4.1 Estimators

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,[8] 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 potential outcome equations to compute the probability of success under both regimes for each individual. The true value is then

[8]Other statistics may also be selected in its place. Since the ATT only needs estimates of the counterfactuals for the treated observations, it is less demanding than the average treatment effect (ATE). Hence, if this method were to be generally useful for the ATE, it would also have to be suitable for the ATT.

obtained by averaging the difference between these two probabilities over the subsample of treated individuals.

In particular, we use as regression-based methods the linear probability model (LPM) as well as the logit, probit, and complementary log-log models. The complementary log-log model uses an asymmetric binary link function, which makes it more appropriate when the probability of success takes values close to zero or one (see 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.[9] 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 (Yun 2004, Fairlie 2005).

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 play a prominent role in the programme evaluation literature (see, *e.g.*, Heckman et al. 1997, Frölich 2004), and their asymptotic properties are established by Heckman et al. (1998). Similarly, local linear regression is 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 (as in Busso et al. 2009 and Huber et al. 2013) from a search grid $0.005 \times 1.25^{g-1}$ for $g = 1, 2, \ldots, 15$, and repeat this process in each replication.[10]

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

[9] Kline (2011) shows 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.

[10] 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.

of these estimators are derived by Abadie and Imbens (2006). Since nearest-neighbour matching estimators are 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, the number of neighbours. We consider the workhorse case of $N = 1$ and also $N = 40$,[11] so we apply eight nearest-neighbour matching estimators in total.

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 (Robins et al. 1994, Robins and Rotnitzky 1995, 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.[12] 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. 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).

[11] 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).

[12] We do not consider such a variant of kernel-based and nearest-neighbour matching estimators for two reasons. First, these estimators explicitly compute a counterfactual for each individual using data from the closest neighbourhood of this individual. Second, these two classes of estimators account for nearly 100% of our computation time, and therefore such an inclusion would be prohibitive timewise. This is not problematic, since our interest is not in how well any particular estimator performs, but rather in comparing the performance of estimators in the original data and in the Monte Carlo samples.

4.2 Metrics

Empirical Monte Carlo studies seek to persuade one of the benefits of using a particular estimator, showing that it is preferred to many others in a particular circumstance. We are able to test the internal validity of such a procedure by comparing the performance of estimators in the original data with their performance using the Monte Carlo data.

Typically one would choose estimators on the basis of minimising either the RMSE or the absolute mean bias between the true value of the statistic of interest and the estimate. 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 biases in both circumstances.

Minimising the RMSE accounts for both the bias and variance of an estimator, so might be preferred as a measure in many contexts. Unfortunately, from a single sample of original data it is not possible to measure the variance of an estimator, only the bias. Hence although one could calculate the RMSE in the Monte Carlo data, this is not possible in the original data. 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 focus on metrics based on bias.

For a researcher comparing the performance of estimators, absolute bias is typically a more relevant metric than bias. We therefore prefer absolute bias to bias as a performance measure which indicates the quality of an EMCS procedure, and in our results we focus on the correlation in absolute (mean) bias between the original data and the Monte Carlo samples ("Abs. bias–Abs. mean bias" in Tables 2–4) as well as on the correlation between (ordinal) rankings of estimators based on absolute (mean) bias ("Rank–Rank"). We also, however, report the correlations for bias.

5 Results

In this section we discuss the performance of various EMCS designs. For each EMCS procedure we implement various nonexperimental estimators for the ATT. We then study the correlation in bias, absolute bias, and ranking, comparing the estimates in the generated data and in the original data.[13]

[13] 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

5.1 The structured design

The baseline correlations in the NSW-PSID design are shown in the first and third columns of Table 2.[14] Mean biases are positively and significantly correlated with the true biases, whilst absolute mean biases are significantly negatively correlated with the true absolute 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 (see Section 3), 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.[15]

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 the true effect in a given replication – which does not vary significantly across replications – this positive correlation in biases simply reflects a positive correlation in the underlying estimates.

However, as noted previously, for a researcher performing an empirical Monte Carlo study the appropriate decision criterion for choosing estimators is typically absolute bias, and on this criterion the researcher would choose the wrong estimators. This result differs from the result on bias, because when taking absolute values it becomes important what value is used as the "true" value against which the bias is calculated.

With the NSW-PSID data, the structured design generates true values equal to –0.2551 and –0.2596, on average, in the uncorrelated and correlated versions, respectively. These

true effect cannot deviate from the $[-1, 1]$ interval. Our rule should not therefore be viewed as particularly restrictive. This leads us to dropping at most 1.8% (0.2%) of the observations for an estimator-design pair in the NSW-PSID (NSW-CPS) data. The only exception is unnormalised reweighting in the correlated structured design for NSW-PSID. In this case we drop up to 8.3% of the observations. We find qualitatively identical results for all the designs when we do not discard any outliers, but instead remove these estimators for which we detect more than 1% of outliers in the first place. These results are available on request.

[14]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.

[15]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. These additional robustness checks never have an effect on our conclusions. These results are available on request.

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)	(3)	(4)	(5)	(6)	(7)	(8)
Correlations								
Bias–Mean bias	0.369**	0.254	0.625***	0.532***	0.369**	0.254	0.625***	0.532***
	(0.032)	(0.192)	(0.000)	(0.003)	(0.032)	(0.192)	(0.000)	(0.003)
Abs. bias–Abs. mean bias	–0.401**	–0.280	–0.447***	–0.236	0.397**	0.286	0.686***	0.603***
	(0.019)	(0.149)	(0.007)	(0.217)	(0.020)	(0.140)	(0.000)	(0.001)
Rank–Rank	–0.360**	–0.194	–0.386**	–0.178	0.395**	0.236	0.656***	0.571***
	(0.036)	(0.320)	(0.022)	(0.356)	(0.021)	(0.226)	(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. Stars indicate significance: *at the 10% level; **at the 5% level; ***at the 1% level.
Columns (1)–(4) correlate biases and rankings in the simulations with biases and rankings in the original data, as measured against the true effect. Columns (5)–(8) correlate biases and rankings in the simulations with hypothetical biases and hypothetical rankings in the original data, as measured against the effect estimated in the original data with the logit OB decomposition. Columns (1), (2), (5), and (6) are based on a DGP which allows covariates to be drawn conditional only on treatment status, whilst in the remaining columns the correlation structure is matched to the data. Odd-numbered columns use all the estimators, whilst even-numbered columns drop OB-based estimators. Outliers are defined as those estimators whose mean biases are more than three standard deviations away from the average mean bias of all the estimators. In columns (1), (2), (5), and (6) unnormalised reweighting with the common support restriction is treated as an outlier.

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, which we term the "hypothetical effect", 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 the first to fourth columns ("true biases") and in the fifth to eighth columns ("hypothetical biases"), since the true effect is already close to the estimated one, and correlations are generally positive. Again the relationships get weaker, and sometimes

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)	(3)	(4)	(5)	(6)	(7)	(8)
Correlations								
Bias–Mean bias	0.389**	0.227	0.547***	0.379**	0.389**	0.227	0.547***	0.379**
	(0.023)	(0.245)	(0.001)	(0.043)	(0.023)	(0.245)	(0.001)	(0.043)
Abs. bias–Abs. mean bias	0.618***	0.599***	0.375**	0.295	0.467***	0.491***	0.293*	0.270
	(0.000)	(0.001)	(0.027)	(0.120)	(0.005)	(0.008)	(0.087)	(0.156)
Rank–Rank	0.554***	0.524***	0.435***	0.345*	0.388**	0.417**	0.406**	0.357*
	(0.001)	(0.004)	(0.009)	(0.067)	(0.023)	(0.027)	(0.016)	(0.058)
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. Stars indicate significance: *at the 10% level; **at the 5% level; ***at the 1% level.
Columns (1)–(4) correlate biases and rankings in the simulations with biases and rankings in the original data, as measured against the true effect. Columns (5)–(8) correlate biases and rankings in the simulations with hypothetical biases and hypothetical rankings in the original data, as measured against the effect estimated in the original data with the logit OB decomposition. Columns (1), (2), (5), and (6) are based on a DGP which allows covariates to be drawn conditional only on treatment status, whilst in the remaining columns the correlation structure is matched to the data. Odd-numbered columns use all the estimators, whilst even-numbered columns drop OB-based estimators. Outliers are defined as those estimators whose mean biases are more than three standard deviations away from the average mean bias of all the estimators. In columns (1), (2), (5), and (6) unnormalised reweighting with the common support restriction is treated as an outlier.

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 which is implicitly used in the data-generating process. However, this assumption is not testable. Further, if one were to take this assumption seriously, then there would be no reason to use any Monte Carlo procedure, since the true effect would already be known.

5.2 The placebo design

The results in Table 4 show that the placebo design 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.[16] Hence, this procedure, as with the structured design, remains unable to provide useful guidance on the choice of estimators.[17]

Although the placebo design avoids the problem of needing to correctly specify a parametric model for the outcome,[18] the treatment effect is now clearly different from that in

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

[17]Simulation results are presented in Tables B.7 and B.8 for the uncalibrated placebo design, and Tables B.9 and B.10 for the calibrated placebo design.

[18]Instead the assumption is made that the distribution of the outcome, conditional on covariates,

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	–0.383**	–0.422**	–0.439***	0.470***
	(0.023)	(0.013)	(0.009)	(0.004)
Abs. bias–Abs. mean bias	–0.101	0.269	0.122	0.094
	(0.564)	(0.124)	(0.492)	(0.593)
Rank–Rank	0.016	0.189	0.213	–0.036
	(0.929)	(0.284)	(0.225)	(0.837)
Sample restrictions				
Exclude outliers	Y	Y	Y	Y
Number of estimators	35	34	34	35

NOTE: P-values are in parentheses. Stars indicate significance: *at the 10% level; **at the 5% level; ***at the 1% level.

Columns with the heading "Uncalibrated" use a DGP which draws observations without adjustment to match the covariate overlap between the samples and the original data, whilst columns with the heading "Calibrated" correct for this. Columns with the heading "NSW-PSID" ("NSW-CPS") use data drawn from the PSID (CPS) sample. Outliers are defined as those estimators whose mean biases are more than three standard deviations away from the average mean bias of all the estimators. The following estimators are treated as outliers: matching on the propensity score ($N = 40$) in the second column and bias-adjusted matching on covariates ($N = 40$) in the third column.

the original data. Additionally, only a subset of the original data is used. To the extent that the distribution of these control observations differs from that of 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 match either the mean bias or the absolute mean bias. Although it is sometimes 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,

is the same for the treated observations as for controls.

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, as suggested by Huber et al. (2013). Here a sample of observations is drawn from the control data, and a placebo treatment is assigned using the propensity score from the full data. The treatment effect in the sample is therefore zero by construction.

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

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

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

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

Austin, P. C. (2010). "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.

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

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

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

Brewer, M., Crossley, T. F., and Joyce, R. (2013). "Inference with Difference-in-Differences Revisited". IZA Discussion Paper no. 7742.

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

Busso, M., DiNardo, J., and McCrary, J. (2013). "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators". *Review of Economics and Statistics*, forthcoming.

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

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

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

Diamond, A. and Sekhon, J. S. (2013). "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies". *Review of Economics and Statistics*, 95:932–945.

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

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

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

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

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

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

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

Heckman, J. J. and Hotz, V. J. (1989). "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.

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

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

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

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

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

Khwaja, A., Picone, G., Salm, M., and Trogdon, J. G. (2011). "A Comparison of Treatment Effects Estimators Using a Structural Model of AMI Treatment Choices and Severity of Illness Information from Hospital Charts". *Journal of Applied Econometrics*, 26:825–853.

Kline, P. (2011). "Oaxaca-Blinder as a Reweighting Estimator". *American Economic Review: Papers & Proceedings*, 101:532–537.

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

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

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

Leuven, E. and Sianesi, B. (2003). "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing". This version 4.0.6.

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

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

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

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

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

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

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

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

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

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

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 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 –0.044), 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	–0.0780	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		–0.0457	0.0551	0.0353	31
Linear probability	X	–0.0422	0.0550	0.0397	28
Logit		0.0066	0.0280	0.0305	10
Logit	X	0.0034	0.0274	0.0306	7
Probit		0.0126	0.0308	0.0320	19
Probit	X	0.0134	0.0325	0.0334	20
Complementary log-log		0.0118	0.0271	0.0265	16
Complementary log-log	X	0.0066	0.0242	0.0255	9
Oaxaca–Blinder					
Linear probability		–0.0474	0.0568	0.0357	32
Linear probability	X	–0.0428	0.0557	0.0402	29
Logit		0.0007	0.0347	0.0383	3
Logit	X	–0.0079	0.0388	0.0420	13
Probit		–0.0098	0.0351	0.0376	14
Probit	X	–0.0155	0.0404	0.0414	21
Kernel-based					
Kernel matching, uniform		0.0010	0.1126	0.1141	5
Kernel matching, Gaussian		0.0364	0.1862	0.1850	26
Kernel matching, Epan.		0.0009	0.1130	0.1145	4
Local linear regression		–0.0636	0.6594	0.6572	34
Local logit		0.0306	0.1883	0.1883	22
Matching					
On pscore, $N = 1$		0.0326	0.1881	0.1877	24
On pscore, $N = 40$		0.0448	0.0772	0.0643	30
On pscore, $N = 1$, bias-adj.		0.0043	0.1710	0.1733	8
On pscore, $N = 40$, bias-adj.		0.0006	0.0964	0.0978	1
On covs, $N = 1$		–0.0076	0.1562	0.1564	12
On covs, $N = 40$		–0.0633	0.0796	0.0508	33
On covs, $N = 1$, bias-adj.		0.0119	0.1775	0.1782	17
On covs, $N = 40$, bias-adj.		–0.0007	0.0947	0.0957	2
Weighting					
Unnormalised		–0.0074	0.4344	0.4361	11
Unnormalised	X	–0.1355	0.4658	0.4476	35
Normalised		0.0365	0.1806	0.1784	27
Normalised	X	0.0116	0.1794	0.1807	15
Efficient		0.0350	0.1578	0.1557	25
Efficient	X	0.0119	0.1202	0.1215	18
Double robust		0.0318	0.1476	0.1459	23
Double robust	X	–0.0017	0.1393	0.1413	6

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		0.0489	0.0636	0.0460	15
Linear probability	X	0.0574	0.1261	0.1128	19
Logit		0.0361	0.0627	0.0540	11
Logit	X	0.1043	0.1294	0.0761	28
Probit		0.0674	0.0841	0.0536	22
Probit	X	0.1205	0.1454	0.0811	31
Complementary log-log		0.0440	0.0646	0.0477	13
Complementary log-log	X	0.1139	0.1296	0.0604	30
Oaxaca–Blinder					
Linear probability		–0.0008	0.0415	0.0468	1
Linear probability	X	0.0514	0.1243	0.1136	16
Logit		0.0018	0.0633	0.0664	3
Logit	X	0.0561	0.1147	0.1004	18
Probit		–0.0009	0.0577	0.0613	2
Probit	X	0.0555	0.1138	0.0996	17
Kernel-based					
Kernel matching, uniform		0.0603	0.3644	0.3598	21
Kernel matching, Gaussian		0.1034	0.3808	0.3672	27
Kernel matching, Epan.		0.0599	0.3662	0.3614	20
Local linear regression		0.0911	0.9063	0.9018	24
Local logit		0.0889	0.4500	0.4413	23
Matching					
On pscore, $N = 1$		0.0971	0.4537	0.4433	26
On pscore, $N = 40$		0.2051	0.2241	0.0903	33
On pscore, $N = 1$, bias-adj.		0.0394	1.5065	1.5065	12
On pscore, $N = 40$, bias-adj.		0.0108	0.3700	0.3699	7
On covs, $N = 1$		–0.0112	0.1538	0.1551	8
On covs, $N = 40$		0.0069	0.0465	0.0504	5
On covs, $N = 1$, bias-adj.		–0.0102	0.4737	0.4739	6
On covs, $N = 40$, bias-adj.		0.0034	0.1556	0.1562	4
Weighting					
Unnormalised		0.2557	0.7785	0.7358	34
Unnormalised	X	–0.5432	1.4081	1.2997	35
Normalised		0.1071	0.3325	0.3151	29
Normalised	X	0.0241	0.3218	0.3209	10
Efficient		0.0961	0.3866	0.3749	25
Efficient	X	0.0487	0.2514	0.2469	14
Double robust		0.2019	0.4794	0.4348	32
Double robust	X	0.0143	0.2748	0.2741	9

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	0.0375	0.0346	31
Linear probability	X	–0.0224	0.0373	0.0348	30
Logit		0.0080	0.0261	0.0290	17
Logit	X	0.0076	0.0260	0.0291	16
Probit		0.0207	0.0336	0.0310	28
Probit	X	0.0209	0.0338	0.0311	29
Complementary log-log		0.0076	0.0223	0.0237	15
Complementary log-log	X	0.0072	0.0221	0.0238	13
Oaxaca–Blinder					
Linear probability		–0.0255	0.0395	0.0350	35
Linear probability	X	–0.0249	0.0392	0.0352	34
Logit		–0.0010	0.0319	0.0365	3
Logit	X	–0.0016	0.0321	0.0368	6
Probit		–0.0068	0.0322	0.0361	11
Probit	X	–0.0071	0.0324	0.0364	12
Kernel-based					
Kernel matching, uniform		0.0149	0.0505	0.0516	26
Kernel matching, Gaussian		0.0185	0.0606	0.0601	27
Kernel matching, Epan.		0.0144	0.0511	0.0523	25
Local linear regression		0.0234	0.6576	0.6571	33
Local logit		0.0132	0.0637	0.0649	23
Matching					
On pscore, $N = 1$		0.0116	0.0671	0.0681	22
On pscore, $N = 40$		0.0109	0.0454	0.0476	21
On pscore, $N = 1$, bias-adj.		0.0037	0.0651	0.0671	9
On pscore, $N = 40$, bias-adj.		0.0005	0.0460	0.0490	2
On covs, $N = 1$		–0.0016	0.0696	0.0717	5
On covs, $N = 40$		–0.0231	0.0474	0.0457	32
On covs, $N = 1$, bias-adj.		–0.0001	0.0727	0.0746	1
On covs, $N = 40$, bias-adj.		–0.0020	0.0483	0.0516	8
Weighting					
Unnormalised		–0.0061	0.0731	0.0748	10
Unnormalised	X	–0.0143	0.0728	0.0736	24
Normalised		0.0095	0.0618	0.0636	19
Normalised	X	0.0075	0.0616	0.0637	14
Efficient		0.0096	0.0521	0.0542	20
Efficient	X	0.0085	0.0509	0.0533	18
Double robust		0.0018	0.0559	0.0589	7
Double robust	X	–0.0012	0.0557	0.0588	4

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		0.0741	0.0847	0.0492	14
Linear probability	X	0.1061	0.1383	0.0920	17
Logit		0.0921	0.1094	0.0641	16
Logit	X	0.1268	0.1480	0.0789	21
Probit		0.1310	0.1425	0.0615	22
Probit	X	0.1601	0.1760	0.0754	24
Complementary log-log		0.0845	0.0994	0.0550	15
Complementary log-log	X	0.1180	0.1322	0.0601	20
Oaxaca–Blinder					
Linear probability		–0.0118	0.0431	0.0489	5
Linear probability	X	0.0666	0.1227	0.1059	11
Logit		0.0039	0.0660	0.0701	3
Logit	X	0.0718	0.1202	0.0993	13
Probit		–0.0008	0.0597	0.0645	1
Probit	X	0.0695	0.1181	0.0985	12
Kernel-based					
Kernel matching, uniform		0.2686	0.4319	0.3388	32
Kernel matching, Gaussian		0.2437	0.3812	0.2938	27
Kernel matching, Epan.		0.2676	0.4324	0.3405	31
Local linear regression		0.1117	1.0116	1.0060	18
Local logit		0.2669	0.4341	0.3429	30
Matching					
On pscore, $N = 1$		0.2715	0.4354	0.3409	33
On pscore, $N = 40$		0.2218	0.2359	0.0810	26
On pscore, $N = 1$, bias-adj.		0.0577	1.1959	1.1950	10
On pscore, $N = 40$, bias-adj.		0.0106	0.3046	0.3058	4
On covs, $N = 1$		–0.0251	0.1446	0.1452	8
On covs, $N = 40$		0.0219	0.0478	0.0498	7
On covs, $N = 1$, bias-adj.		–0.0119	0.4454	0.4461	6
On covs, $N = 40$, bias-adj.		–0.0022	0.1350	0.1379	2
Weighting					
Unnormalised		0.2746	0.8834	0.8395	34
Unnormalised	X	–0.0342	1.1470	1.1472	9
Normalised		0.2949	0.4438	0.3322	35
Normalised	X	0.2618	0.4310	0.3429	29
Efficient		0.2606	0.5470	0.4812	28
Efficient	X	0.2083	0.5358	0.4938	25
Double robust		0.1553	0.3374	0.3014	23
Double robust	X	0.1168	0.2861	0.2625	19

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		–0.0130	0.0370	0.0347	27
Linear probability	X	0.0005	0.0387	0.0388	2
Logit		0.0027	0.0366	0.0365	7
Logit	X	0.0091	0.0398	0.0388	23
Probit		0.0028	0.0364	0.0363	8
Probit	X	0.0104	0.0404	0.0391	26
Complementary log-log		0.0038	0.0363	0.0361	10
Complementary log-log	X	0.0083	0.0384	0.0375	19
Oaxaca–Blinder					
Linear probability		–0.0138	0.0374	0.0348	28
Linear probability	X	0.0006	0.0391	0.0392	3
Logit		0.0017	0.0367	0.0367	6
Logit	X	0.0089	0.0401	0.0391	22
Probit		–0.0008	0.0360	0.0360	5
Probit	X	0.0085	0.0398	0.0389	20
Kernel-based					
Kernel matching, uniform		–0.0049	0.0692	0.0691	12
Kernel matching, Gaussian		–0.0040	0.1094	0.1095	11
Kernel matching, Epan.		–0.0050	0.0692	0.0691	13
Local linear regression		–0.0310	0.4111	0.4104	31
Local logit		–0.0071	0.1122	0.1121	16
Matching					
On pscore, $N = 1$		–0.0092	0.1115	0.1112	25
On pscore, $N = 40$		–0.0633	0.0867	0.0593	35
On pscore, $N = 1$, bias-adj.		–0.0054	0.1033	0.1033	14
On pscore, $N = 40$, bias-adj.		–0.0401	0.0796	0.0688	32
On covs, $N = 1$		–0.0007	0.0954	0.0955	4
On covs, $N = 40$		–0.0413	0.0632	0.0479	33
On covs, $N = 1$, bias-adj.		0.0088	0.0995	0.0992	21
On covs, $N = 40$, bias-adj.		–0.0242	0.0679	0.0636	30
Weighting					
Unnormalised		–0.0082	0.1034	0.1032	18
Unnormalised	X	–0.0413	0.1150	0.1075	34
Normalised		–0.0080	0.0939	0.0937	17
Normalised	X	0.0003	0.0936	0.0936	1
Efficient		–0.0092	0.0958	0.0955	24
Efficient	X	–0.0031	0.0824	0.0824	9
Double robust		–0.0061	0.0898	0.0897	15
Double robust	X	–0.0168	0.0882	0.0866	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		0.0233	0.0391	0.0315	18
Linear probability	X	0.0354	0.0495	0.0346	26
Logit		0.0313	0.0487	0.0373	23
Logit	X	0.0367	0.0532	0.0385	28
Probit		0.0399	0.0537	0.0361	29
Probit	X	0.0449	0.0586	0.0377	32
Complementary log-log		0.0276	0.0452	0.0358	19
Complementary log-log	X	0.0303	0.0461	0.0348	21
Oaxaca–Blinder					
Linear probability		0.0130	0.0356	0.0331	14
Linear probability	X	0.0358	0.0503	0.0353	27
Logit		0.0152	0.0407	0.0378	15
Logit	X	0.0347	0.0506	0.0368	25
Probit		0.0228	0.0426	0.0359	17
Probit	X	0.0410	0.0546	0.0360	30
Kernel-based					
Kernel matching, uniform		–0.0015	0.0725	0.0725	2
Kernel matching, Gaussian		–0.0067	0.1596	0.1595	7
Kernel matching, Epan.		–0.0026	0.0706	0.0705	3
Local linear regression		0.0039	0.4893	0.4894	4
Local logit		–0.0126	0.1636	0.1631	12
Matching					
On pscore, $N = 1$		–0.0113	0.1631	0.1628	11
On pscore, $N = 40$		–0.0339	0.0691	0.0603	24
On pscore, $N = 1$, bias-adj.		–0.0311	0.1358	0.1322	22
On pscore, $N = 40$, bias-adj.		–0.0435	0.0912	0.0802	31
On covs, $N = 1$		–0.0292	0.0636	0.0565	20
On covs, $N = 40$		0.0184	0.0387	0.0341	16
On covs, $N = 1$, bias-adj.		–0.0553	0.1098	0.0949	33
On covs, $N = 40$, bias-adj.		–0.0918	0.1059	0.0528	35
Weighting					
Unnormalised		0.0130	0.2109	0.2105	13
Unnormalised	X	–0.0861	0.2434	0.2278	34
Normalised		–0.0064	0.1287	0.1285	6
Normalised	X	0.0041	0.1275	0.1275	5
Efficient		–0.0090	0.1253	0.1250	9
Efficient	X	0.0007	0.0852	0.0853	1
Double robust		0.0102	0.1114	0.1110	10
Double robust	X	–0.0083	0.1079	0.1076	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		–0.0226	0.0388	0.0315	31
Linear probability	X	–0.0230	0.0390	0.0316	33
Logit		–0.0064	0.0322	0.0316	19
Logit	X	–0.0072	0.0322	0.0315	21
Probit		–0.0083	0.0328	0.0318	23
Probit	X	–0.0091	0.0330	0.0317	25
Complementary log-log		–0.0033	0.0305	0.0304	13
Complementary log-log	X	–0.0039	0.0305	0.0303	14
Oaxaca–Blinder					
Linear probability		–0.0229	0.0392	0.0319	32
Linear probability	X	–0.0233	0.0395	0.0319	34
Logit		–0.0068	0.0330	0.0323	20
Logit	X	–0.0077	0.0332	0.0323	22
Probit		–0.0102	0.0340	0.0325	27
Probit	X	–0.0110	0.0342	0.0324	28
Kernel-based					
Kernel matching, uniform		0.0083	0.0408	0.0400	24
Kernel matching, Gaussian		0.0171	0.0442	0.0408	30
Kernel matching, Epan.		0.0063	0.0401	0.0397	18
Local linear regression		–0.0049	0.3984	0.3988	15
Local logit		0.0101	0.0440	0.0429	26
Matching					
On pscore, $N = 1$		0.0001	0.0403	0.0404	2
On pscore, $N = 40$		–0.0126	0.0409	0.0389	29
On pscore, $N = 1$, bias-adj.		0.0019	0.0380	0.0380	12
On pscore, $N = 40$, bias-adj.		0.0004	0.0376	0.0376	4
On covs, $N = 1$		0.0058	0.0379	0.0375	17
On covs, $N = 40$		–0.0236	0.0432	0.0362	35
On covs, $N = 1$, bias-adj.		0.0054	0.0381	0.0378	16
On covs, $N = 40$, bias-adj.		0.0000	0.0356	0.0356	1
Weighting					
Unnormalised		0.0016	0.0393	0.0393	7
Unnormalised	X	–0.0018	0.0389	0.0389	9
Normalised		0.0019	0.0386	0.0386	11
Normalised	X	0.0011	0.0383	0.0383	5
Efficient		0.0018	0.0386	0.0386	10
Efficient	X	0.0013	0.0381	0.0381	6
Double robust		0.0017	0.0381	0.0381	8
Double robust	X	–0.0002	0.0375	0.0376	3

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		0.0037	0.0251	0.0248	25
Linear probability	X	0.0056	0.0253	0.0246	32
Logit		0.0031	0.0264	0.0262	22
Logit	X	0.0048	0.0265	0.0261	27
Probit		0.0053	0.0263	0.0258	28
Probit	X	0.0068	0.0265	0.0257	33
Complementary log-log		0.0015	0.0246	0.0246	11
Complementary log-log	X	0.0028	0.0245	0.0243	20
Oaxaca–Blinder					
Linear probability		0.0023	0.0258	0.0257	16
Linear probability	X	0.0048	0.0259	0.0255	26
Logit		0.0006	0.0261	0.0261	4
Logit	X	0.0029	0.0259	0.0257	21
Probit		0.0031	0.0261	0.0259	23
Probit	X	0.0054	0.0261	0.0256	29
Kernel-based					
Kernel matching, uniform		0.0008	0.0275	0.0275	5
Kernel matching, Gaussian		–0.0001	0.0289	0.0289	1
Kernel matching, Epan.		0.0001	0.0276	0.0276	2
Local linear regression		0.0023	0.1148	0.1148	17
Local logit		0.0056	0.0305	0.0300	30
Matching					
On pscore, $N = 1$		–0.0056	0.0314	0.0309	31
On pscore, $N = 40$		–0.0033	0.0279	0.0277	24
On pscore, $N = 1$, bias-adj.		–0.0019	0.0259	0.0259	13
On pscore, $N = 40$, bias-adj.		–0.0099	0.0289	0.0271	34
On covs, $N = 1$		0.0027	0.0250	0.0248	19
On covs, $N = 40$		0.0021	0.0257	0.0256	14
On covs, $N = 1$, bias-adj.		0.0022	0.0251	0.0250	15
On covs, $N = 40$, bias-adj.		–0.0143	0.0300	0.0264	35
Weighting					
Unnormalised		–0.0010	0.0268	0.0268	7
Unnormalised	X	–0.0025	0.0272	0.0270	18
Normalised		–0.0011	0.0269	0.0269	9
Normalised	X	–0.0009	0.0268	0.0268	6
Efficient		–0.0012	0.0270	0.0270	10
Efficient	X	–0.0001	0.0267	0.0267	3
Double robust		–0.0011	0.0267	0.0267	8
Double robust	X	–0.0018	0.0268	0.0268	12

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.