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

Arun Advani, University of Warwick, CAGE and Institute for Fiscal Studies

Toru Kitagawa, University College London and cemmap

Tymon Sloczynski, Brandeis University and IZA

Working Paper Series

2018 | 124

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

ARUN ADVANI,† TORU KITAGAWA,‡ AND TYMON SŁOCZYŃSKI§

Abstract

Currently there is little practical advice on which treatment effect estimator to use when trying to adjust for observable differences. A recent suggestion is to compare the performance of estimators in simulations that somehow mimic the empirical context. Two ways to run such 'empirical Monte Carlo studies' (EMCS) have been proposed. We show theoretically that neither is likely to be informative except under restrictive conditions that are unlikely to be satisfied in many contexts. To test empirical relevance, we also apply the approaches to a real-world setting where estimator performance is known. We find that in our setting both EMCS approaches are worse than random at selecting estimators which minimise absolute bias. They are better when selecting estimators that minimise mean squared error. However, using a simple bootstrap is at least as good and often better. For now researchers would be best advised to use a range of estimators and compare estimates for robustness.

JEL Classification: C15, C21, C25, C52

Keywords: empirical Monte Carlo studies, program evaluation, selection on observables, treatment effects

*This version: September 21, 2018. For helpful comments, we thank Thierry Magnac (Co-Editor), four anonymous referees, Alberto Abadie, Cathy Balfe, Richard Blundell, Colin Cameron, Mónica Costa Dias, Gil Epstein, Alfonso Flores-Lagunes, Ira Gang, Martin Huber, Justin McCrary, Blaise Melly, Mateusz Myśliwski, Pedro Sant'Anna, Anthony Strittmatter, Timothy Vogelsang, Jeffrey Wooldridge, Tiemen Woutersen, and seminar and conference participants at Brandeis, Ce2 workshop, CERGE-EI, IAAE, IFS, MEG, MSU, SGH, SOLE, WIEM, and ZEW. We also thank Michael Lechner and Blaise Melly for providing us with their codes, as well as Steven Karel and Francesco Pontiggia for assistance with the Brandeis HPC cluster. This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network (Grant No: RRC12+09). All opinions expressed are those of the authors and have not been endorsed by CERGE-EI or the GDN. Advani also acknowledges support from Programme Evaluation for Policy Analysis, a node of the National Centre for Research Methods, supported by the ESRC (Grant No: RES-576-25-0042). Kitagawa also acknowledges support from the ESRC through the ESRC Centre for Microdata Methods and Practice (cemmap) (Grant No: RES-589-28-0001) and from the ERC through an ERC starting grant (Grant No: EPP-715940). Słoczyński also acknowledges support from the Foundation for Polish Science (FNP) through a START scholarship and from the Theodore and Jane Norman Fund.

†University of Warwick, CAGE, and Institute for Fiscal Studies.

‡University College London and cemmap.

§Brandeis University and IZA. Correspondence: Department of Economics & International Business School, Brandeis University, MS 021, 415 South Street, Waltham, MA 02453. E-mail: tslocz@brandeis.edu.

1 Introduction

A large literature focuses on estimating average treatment effects under unconfoundedness (see, *e.g.*, Blundell and Costa Dias 2009, Imbens and Wooldridge 2009).[1] Many estimators are available to researchers in this context, and many of these estimators have similar asymptotic properties. This can make it difficult to select which estimator to use. Monte Carlo studies are a useful tool for examining the small-sample properties of these estimation methods, which can guide estimator choice.[2] Early contributions, such as Frölich (2004), demonstrate estimator performance in stylised data generating processes (DGPs) which do not resemble any empirical settings. This reliance on unrealistic DGPs is criticised by Huber *et al.* (2013) and Busso *et al.* (2014). Both recommend that Monte Carlo studies should intend to replicate actual datasets of interest, although they suggest different procedures for doing this. Huber *et al.* (2013) describe this approach to examining the small-sample properties of estimators as an 'empirical Monte Carlo study' (EMCS). An important question is whether either type of EMCS can help applied researchers in choosing what estimator(s) to prefer in a given context. Busso *et al.* (2014) indicate this might be possible, noting that their results 'suggest the wisdom of conducting a small-scale simulation study tailored to the features of the data at hand'.[3]

In this paper we evaluate the premise that EMCS is 'internally valid': that it can be informative about the performance of estimators in the particular data which are the basis for the EMCS.[4] We first show theoretically that these approaches are expected to be informative only under very restrictive conditions. These conditions are unlikely to hold in many practical examples faced by a researcher. We then test EMCS performance in a real-world case where we know the actual behaviour of estimators. We find that in terms of selecting estimators on absolute bias they are often worse than choosing randomly. On mean squared error (MSE) they perform better than random, but no better than selecting

[1] The unconfoundedness assumption may also be referred to as exogeneity, ignorability, or selection on observables.

[2] See, for example, 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), Diamond and Sekhon (2013), Huber *et al.* (2013), Busso *et al.* (2014), Frölich *et al.* (2017), and Bodory *et al.* (2018), all studying the finite-sample performance of estimators of average treatment effects under unconfoundedness.

[3] Similarly, Huber *et al.* (2013) suggest that 'the advantage [of an EMCS] is that it is valid in at least one relevant environment', *i.e.* that it is informative at least about the performance of estimators in the dataset on which it was conducted.

[4] This usage of 'internal validity' is somewhat non-standard. However, it is consistent with the definition given by Angrist and Krueger (1999). They define internal validity as the question of 'whether an empirical relationship has a causal interpretation in the setting where it is observed'. In our case the relationship is between the performance of estimators in the original data and their performance in the EMCS implemented on these data.

an estimator based on simple bootstrap estimates of MSEs. Their performance in absolute terms may also still be poor.

The first type of EMCS we consider is the *placebo* EMCS (Huber *et al.*, 2013).[5] This proposes a way to assign 'realistic placebo treatments among the non-treated', using information about the predictors of treatment status in the original data. It then tests how well estimators can recover the zero effect of the placebo treatment. The performance of estimators in this exercise is hypothesised to be informative about their performance in the original data.

The second type we describe as the *structured* EMCS. An exercise of this type is undertaken by Busso *et al.* (2014).[6] Here a parameterised approximation of the original data generating process is created, using functional form assumptions about the distributions of observed covariates. Parameters of their marginal (or conditional) distributions are estimated from the original data.[7] Samples can be drawn from this approximate DGP, to which the estimators can be applied. Since the treatment effect in this DGP can be calculated directly from knowledge of the parameters, performance of the estimators in these samples can be measured. The performance of estimators in this exercise is also hypothesised to be informative about their performance in the original data.

To examine whether or not EMCS can correctly choose a best performing estimator, for various definitions of performance, we first focus on a simple example with two estimators that have Gaussian sampling distributions. We show analytically that both these approaches will only be guaranteed to correctly select the preferred estimator if they can correctly reproduce both the biases and the ordering of the variances of estimators. These are restrictive conditions that we show can easily fail in practical applications, such as when the EMCS procedures fail to recover heteroskedastic errors or misspecify the regression equations or propensity scores. In two sets of simulations based on a stylised DGP, both approaches select the better estimator less than 3% of the time, much worse than 50% achievable by selecting randomly.

To study the extent of the problem in a real-world circumstance, we apply both methods to the National Supported Work (NSW) Demonstration data on men, previously analysed by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), and many others. In these data participation in a job training programme was randomly assigned, so the treatment effect of the programme can

[5] It is also applied by Lechner and Wunsch (2013), Huber *et al.* (2016), Lechner and Strittmatter (2016), Frölich *et al.* (2017), and Bodory *et al.* (2018). A related approach is proposed by Schuler *et al.* (2017).

[6] A similar approach is also used by Abadie and Imbens (2011), Lee (2013), and Díaz *et al.* (2015).

[7] The distribution of particular covariates may be allowed to depend on the realisations of others, in which case parameters of the conditional distributions are needed.

be estimated by comparing sample means. LaLonde (1986) used these data to test the performance of estimators at reproducing this treatment effect when an artificial comparison group (rather than the experimental controls) was used. We instead use the data to test how well the two EMCS procedures can inform us about the performance of the estimators: Can EMCS tell us which estimator to use? On average how much worse than the optimal estimator is the one chosen by EMCS? How well can EMCS reproduce the ranking of performance across all estimators?

Applying the two EMCS procedures we find three main results. First, in terms of absolute bias, the EMCS procedures are no better, and often noticeably worse, than selecting an estimator at random. In two out of three cases we study, the rankings produced are *negatively* correlated with the true ranking. In one case the preferred estimator selected by EMCS is on average 30–37 times worse than the actual best estimator.

Second, EMCS does better at reproducing the performance of estimators in terms of MSE. This is because the MSEs of the estimators are mostly driven by their variances, and EMCS appears more effective at capturing variances. The rankings of estimators are consistently positively correlated with the true rankings, although the estimator preferred by EMCS has an MSE up to twice as high as the best estimator.

Third, given the variance result, we also compare EMCS procedures to choosing estimators based on which has the lowest variance from a simple bootstrap. We find that the bootstrap is as good, and often much better, than either of the EMCS procedures. Hence even when the procedures are somewhat informative, they are not superior to a procedure that relies on fewer design choices.

These results are unfortunate, but nevertheless important. They caution against treating either of these approaches as general solutions to the problem of estimator choice. There remains no silver bullet that can assist empirical researchers with the 'right' or 'best' estimator for a particular context. In the absence of a clear choice driven by research design, the best advice at this stage is likely to be implementing a number of estimators, and then considering the range of estimates provided, as Busso *et al.* (2014) also suggest.

Our results also have implications for researchers studying the small-sample properties of treatment effect estimators (see footnote 2). It has been argued that 'it is preferable to study DGPs that are empirically relevant' (Busso *et al.*, 2014).[8] Our theoretical and empirical results suggest there is little support for this claim. We show theoretically that misspecification in the construction of the DGP can lead the ranking of estimators to be incorrect for the original dataset. In our empirical example, we see that EMCS is not better than using a bootstrap (and sometimes not better than random) to predict performance

[8] A similar argument is made in Huber *et al.* (2013).

in the data on which the EMCS was performed. There seems little reason to then think it is particularly informative about performance in other unrelated real datasets, *i.e.* that testing small-sample properties of estimators in 'real data' is necessarily better than in completely artificial data. A more fruitful path might be to test sensitivity of estimator performance to parameters of the simulation, such as sample size and the degree of heteroskedasticity. This approach is also taken by Huber *et al.* (2013), and might be more helpful in understanding what characteristics of samples most affect the performance of particular estimators.

2 EMCS Designs

We first describe the two main approaches to conducting an EMCS, namely the placebo design of Huber *et al.* (2013) and the structured design of Busso *et al.* (2014). In either EMCS design, one simulates many 'empirical Monte Carlo' replication samples from a known data generating process. By implementing the estimators on the simulated replications, one obtains estimates of the sampling distributions and performance criteria (*e.g.*, MSEs) of the estimators, according to which one ranks the candidate estimators. Note that the researcher needs to make a choice of what criteria to use to rank estimators.

2.1 The Placebo Design

The idea of the placebo design is to assign placebo treatments to some control observations, and attempt to recover the true effect which by construction is zero.[9] In particular, covariates and outcomes $(\mathbf{X}_i, Y_i)$ are first drawn jointly by sampling (with replacement) from the empirical distribution of control observations.[10] Using the original dataset, the propensity score is estimated (*e.g.*, using a logit model). The estimated parameters of this model $\hat{\boldsymbol{\phi}}$ are then used to assign placebo treatments to the generated sample in the following way:

$$D_i = 1[S_i > 0], \tag{1}$$

$$S_i = \pi + \lambda \mathbf{X}_i \hat{\boldsymbol{\phi}} + \epsilon_i, \tag{2}$$

[9] A similar approach is developed by Bertrand *et al.* (2004) who study inference in difference-in-differences methods using simulations with randomly generated 'placebo laws' in state-level data, *i.e.* policy changes which never actually happened. For follow-up studies, see Hansen (2007), Cameron *et al.* (2008), and Brewer *et al.* (2018).

[10] In this paper the sample size is always equal to the size of the original control subsample.

where ϵ_i is an iid error, and both π and λ are additional parameters to be selected. While π shifts the proportion of observations that are treated, λ controls the extent of selection: with $\lambda = 1$ selection on observables takes the same form in the Monte Carlo sample as in the original dataset.

2.2 The Structured Design

The idea of the structured design is instead to create a parameterised approximation to the original (unknown) data generating process, and then draw samples from the approximated process. To begin, a fixed number of treated and control observations are created, to match the number of each in the original dataset. Covariates and outcome variables are then drawn from parameterised distributions where the parameters are estimated from the original dataset. For example, the variable `black` might come from a Bernoulli with mean estimated from the data, and the variable `earnings` from a log-normal distribution with mean and variance estimated from the data. The parameters of these distributions are typically estimated conditional on treatment status. Parameters of some distributions might also be conditional on the value of other variables; *e.g.*, earnings might be conditional on race as well as treatment status. More conditioning will improve the match of the joint distribution of simulated data to the joint distribution of the original data, but will increase the number of parameters that need to be estimated.

3 Theory

To understand the conditions under which an EMCS might be informative about the preferred estimator in some particular dataset, we first construct a simple example. Here we have only two estimators, with a straightforward and restricted joint sampling distribution (bivariate Gaussian). This bivariate Gaussian setting mimics an ideal situation in which the finite sample distribution of the estimators is well approximated by their asymptotic distribution. We show that even in such an ideal, large sample situation, EMCS can fail to select the best estimator if the bias in any one of the estimators or the ranking of variances is not correctly replicated in the simulated samples.[11] We provide simple common cases for treatment effect estimation in which failure to capture the biases and heteroskedasticity contaminates EMCS, and provide results from a simple simulation

[11] In Section 5 we will show that in practice this means that, when estimators are unbiased, rankings based on a simple bootstrap perform at least as well and sometimes better than the more involved EMCS procedures.

illustrating this.[12] We then extend the example to the case of more than two estimators.

3.1 Simple Example: Two Estimator Case

Suppose the researcher wants to rank two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ according to their statistical performances under repeated sampling. These estimators are estimating the same object of interest $\theta \in \mathbb{R}$, but their constructions are different. For simplicity of the illustration, assume that the joint sampling distribution of the two estimators is bivariate Gaussian:

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \Sigma_n \right), \tag{3}$$

where $\Sigma_n = n^{-1}\Sigma$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, and n is the sample size. Here, our implicit assumption is that the estimators $(\hat{\theta}_1, \hat{\theta}_2)$ converge to (θ_1, θ_2) at $\sqrt{n}$-rate. Let θ^0 be the true value of the parameter of interest. We allow $\hat{\theta}_1$ and/or $\hat{\theta}_2$ to be biased so that θ_1 and/or θ_2 can differ from θ^0.

We rank these estimators according to their statistical performances. Given that we often assess the performance of an estimator by its mean squared error (MSE) or mean absolute error (MAE), we may, for instance, rank the estimators according to their MSEs or MAEs.[13] Given the Gaussian assumption, the MSE of each estimator, $j = 1, 2$, is

$$MSE(\hat{\theta}_j) = (\theta_j - \theta^0)^2 + n^{-1}\sigma_j^2.$$

We denote by $j_0 \in \{1, 2\}$ the index of the strictly preferred estimator, assuming it exists. Ranking the estimators is difficult in practice since we do not know the mean and variances of the estimators as well as the true value of θ. Proposals of the EMCS literature aim to infer a best performing estimator j_0 by estimating the sampling distribution of $\hat{\theta}_1$ and $\hat{\theta}_2$ via some Monte Carlo studies. For simplicity, we assume that the estimators

[12] The role of the simulations is to show a quantitative example of how performance might look. Since we are aware this simulation is stylised, in Section 4 we also provide simulation results based on real-world data.

[13] MSE and MAE criteria do not take into account the dependence of the estimators. One way to rank the estimators that takes into account their dependence is based on *the probability of being closer to the truth*, $\Pr(|\hat{e}_1| \leq |\hat{e}_2|)$, where $\hat{e}_1 = \hat{\theta}_1 - \theta^0$ and $\hat{e}_2 = \hat{\theta}_2 - \theta^0$ are the estimation errors of the two estimators. That is, $\hat{\theta}_1$ is preferred to $\hat{\theta}_2$ if $\Pr(|\hat{e}_1| \leq |\hat{e}_2|) > 1/2$ and $\hat{\theta}_2$ is preferred to $\hat{\theta}_1$ if $\Pr(|\hat{e}_1| \leq |\hat{e}_2|) < 1/2$. Considering this criterion instead of MSE does not affect the main results in our simple example.

simulated in EMCS also follow bivariate Gaussian,

$$\begin{pmatrix} \hat{\theta}_1^* \\ \hat{\theta}_2^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix}, \tilde{\Sigma}_n \right), \tag{4}$$

where $\tilde{\Sigma}_n = a_n^{-1}\tilde{\Sigma}$, $\tilde{\Sigma} = \begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_{12} \\ \tilde{\sigma}_{12} & \tilde{\sigma}_2^2 \end{pmatrix}$, and a_n is the size of a simulated sample that may differ from the size of the original sample n. The underlying parameters in EMCS, $(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\Sigma})$, generally depend on the original sample, but we assume for simplicity that the dependence is negligible and they can be treated as constants. EMCS computes $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ repeatedly using simulated samples of size a_n drawn from a data generating process with the parameter value set at known value $\tilde{\theta}^0$. For instance, the placebo EMCS approach of Huber *et al.* (2013) sets $\tilde{\theta}^0 = 0$ and $a_n \leq n_0$, the size of the control group in the original data. The approach of structured EMCS sets $\tilde{\theta}^0$ at an estimate of θ^0 constructed from the original sample. In implementing EMCS, we do not have to know the mean and variance parameters of $(\hat{\theta}_1^*, \hat{\theta}_2^*)$, and they can be estimated with arbitrary accuracy based on the simulated estimators. EMCS accordingly obtains the MSE of each estimator, $j = 1, 2$, by

$$\widehat{MSE}(\hat{\theta}_j) = (\tilde{\theta}_j - \tilde{\theta}^0)^2 + a_n^{-1}\tilde{\sigma}_j^2.$$

We denote by $\hat{j}_0$ the index for a best performing estimator estimated from EMCS, $\hat{j}_0 \equiv \arg\min_j \widehat{MSE}(\hat{\theta}_j)$. To assess the validity of EMCS, we define a criterion of *EMCS internal validity* by the probability that $\hat{j}_0$ coincides with j_0, $\Pr(\hat{j}_0 = j_0)$, where the probability is evaluated under repeated sampling of the original samples. In the examples to follow, we investigate how this criterion of EMCS validity becomes one or zero depending on the parameter values in the bivariate Gaussian distributions of (3) and (4).[14]

We can also consider the *average regret* type criterion such as $\mathbb{E}(MSE(\hat{\theta}_{\hat{j}_0}) - MSE(\hat{\theta}_{j_0})) \geq 0$ to quantify EMCS internal validity. Here, the expectation concerns the sampling distribution of EMCS's selection of an optimal estimator $\hat{j}_0$. This average regret criterion can quantify severity of a wrong choice of the estimators in terms of how much MSE is on average sacrificed relative to the true best-performing estimator.

[14]We assume away the dependence of the parameters in (4) on the original sample for simplicity of illustration. In such a case the MSE estimates in EMCS and resulting selection of a best estimator $\hat{j}_0$ are nonrandom. The criterion of EMCS internal validity in this case is either 1 or 0.

3.1.1 Scenario 1

Denote the biases in $(\hat{\theta}_1, \hat{\theta}_2)'$ by $\mathbf{b} = (b_1, b_2)' = (\theta_1 - \theta^0, \theta_2 - \theta^0)'$ and the biases in $(\hat{\theta}_1^*, \hat{\theta}_2^*)'$ by $\tilde{\mathbf{b}} = (\tilde{b}_1, \tilde{b}_2)' = (\tilde{\theta}_1 - \tilde{\theta}^0, \tilde{\theta}_2 - \tilde{\theta}^0)'$. We start with a scenario in which $(\hat{\theta}_1, \hat{\theta}_2)$ are unbiased and the distribution of $(\hat{\theta}_1^*, \hat{\theta}_2^*)$ well replicates the distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ in the following sense:

$$\mathbf{b} = \tilde{\mathbf{b}} = \mathbf{0}, \qquad \Sigma = \tilde{\Sigma}. \tag{5}$$

Here, the biases and the sample-size-adjusted variances of the estimators simulated in EMCS coincide with those of the estimators in the original data generating process. Note that the true value of parameter assumed in EMCS, $\tilde{\theta}^0$, does not have to agree with the true parameter value in the original sampling process, θ^0.

In the current scenario, the ranking of the true MSEs clearly coincides with the ranking of the MSE estimates in EMCS, implying $\Pr(\hat{j}_0 = j_0) = 1$. This is a benchmark case in which EMCS works. The next two scenarios show that once we depart from the assumptions in (5), EMCS can be no longer valid.

3.1.2 Scenario 2

Assume that the estimators are free from biases both in the original data generating process and EMCS, $\mathbf{b} = \tilde{\mathbf{b}} = \mathbf{0}$, but EMCS fails to replicate the normalised covariance matrix of the estimators, $\Sigma \neq \tilde{\Sigma}$. In this case, the MSE estimates in EMCS correctly rank the true MSEs of the estimators (assuming $\sigma_1^2 \neq \sigma_2^2$) if and only if the ordering of the variances of the two estimators agrees between the original sampling process and the simulated sampling process, *i.e.* $(\sigma_1^2 - \sigma_2^2)(\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2) > 0$. Otherwise, EMCS reverses the ranking of the estimators and incorrectly selects a suboptimal estimator as optimal, $\Pr(\hat{j}_0 = j_0) = 0$.

Hence, even when EMCS well replicates the biases of the estimators, it can fail to select a best performing estimator due to an incorrect variance ordering.

3.1.3 Scenario 3

In the third scenario, we assume that EMCS correctly replicates the variance ordering of the estimators, *i.e.* $(\sigma_1^2 - \sigma_2^2)(\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2) > 0$, but fails to replicate the biases, $(b_1, b_2) \neq (\tilde{b}_1, \tilde{b}_2)$. To be specific, we set $(\tilde{b}_1, \tilde{b}_2) = (0, 0)$, but $(b_1, b_2) = (0, b_2), b_2 \neq 0$. This can correspond to a situation that the estimator 1 is correctly specified and has no bias, whereas estimator 2 is misspecified and is subject to bias in the original data generating process. EMCS, however, fails to capture the misspecification bias in estimator 2.

Suppose $\sigma_1^2 > \sigma_2^2$ holds. The true MSEs are $MSE(\hat{\theta}_1) = n^{-1}\sigma_1^2$ and $MSE(\hat{\theta}_2) = b_2^2 +$

$n^{-1}\sigma_2^2$, while the MSE estimates in EMCS are $\widehat{MSE}(\hat{\theta}_1) = a_n^{-1}\tilde{\sigma}_1^2$ and $\widehat{MSE}(\hat{\theta}_2) = a_n^{-1}\tilde{\sigma}_2^2$. Since we assumed that EMCS correctly replicates the variance of the estimators, EMCS selects $j = 2$ as a best estimator. This selection of the estimator is indeed misleading if b_2 is far from zero, since if $|b_2| > \sqrt{\frac{\sigma_1^2 - \sigma_2^2}{n}}$, $\hat{\theta}_1$ outperforms $\hat{\theta}_2$ in terms of MSE.

This scenario highlights that EMCS-based selection of the estimator can fail if any one of the estimators is misspecified and the simulation design in EMCS does not replicate the misspecification bias.

3.2 Are Scenarios 2 and 3 Relevant in Treatment Effect Estimation?

We next provide simple but empirically relevant examples where we focus on the estimation of treatment effects, and show that both types of EMCS may yield misleading choices of the estimators for the reasons illustrated in Scenarios 2 and 3 above.

Data are given by a random sample of $\{(Y_i, D_i, X_i) : i = 1, \ldots, n\}$, where $Y_i \in \mathbb{R}$ is unit i's observed post-treatment outcome, $D_i \in \{0, 1\}$ is her treatment status, and $X_i \in \mathbb{R}^{d_x}$ is a vector of her pre-treatment characteristics whose support is assumed to be bounded. We denote unit i's potential outcomes by $(Y_i(1), Y_i(0))$. We assume the unconfoundedness assumption, $(Y(1), Y(0)) \perp D|X$, throughout. The propensity score is denoted by $e(x) = \Pr(D = 1|X = x)$.

3.2.1 An Example for Scenario 2

To keep our example as simple as possible, consider the following data generating processes:

$$\begin{aligned} \mathbb{E}(Y(1)|X = x) &= \beta_0 + \beta_1 + x'\beta_2, \\ \mathbb{E}(Y(0)|X = x) &= \beta_0 + x'\beta_2, \\ Var(Y(1)|X = x) = c\sigma_\epsilon^2, \quad & Var(Y(0)|X = x) = \sigma_\epsilon^2, \quad c > 0, \\ e(x) &= \gamma_0 + x'\gamma_1. \end{aligned} \tag{6}$$

The specified mean equations for both potential outcomes imply that the conditional average treatment effects are homogeneous over observable characteristics and equal to β_1. The potential outcomes are heteroskedastic if $c \neq 1$. We assume a linear probability for the propensity score in order to simplify analytical comparisons of the variances of the estimators we introduce below.

Suppose that the parameter of interest is the population average treatment effect for

the treated (ATT), $\theta^0 = \mathbb{E}(Y(1) - Y(0)|D = 1)$. Since specification (6) implies homogeneous treatment effects, $\mathbb{E}(Y(1) - Y(0)|X = x) = \beta_1$, the true value of ATT is $\theta^0 = \beta_1$.

We consider two different estimators to estimate the population ATT. The first estimator $\hat{\theta}_1$ is a semiparametric estimator for ATT, which is consistent without assuming functional forms for the outcome and propensity score equations, and asymptotically attains the semiparametric efficiency bound (SEB) of ATT derived by Hahn (1998). Estimators that attain this property include the inverse probability weighting (IPW) estimator with nonparametrically estimated propensity scores (Hirano *et al.*, 2003), doubly robust estimators of Hahn (1998), covariate or propensity score matching estimators with a single covariate (Abadie and Imbens, 2006, 2016), and covariate balancing estimators of Graham *et al.* (2012, 2016). We can set any one of these estimators as our first estimator without affecting the analysis below.

We specify the second estimator $\hat{\theta}_2$ as the ordinary least squares estimator of β_1 in the following regression equation:

$$Y_i = \beta_0 + \beta_1 D_i + X_i'\beta_2 + \epsilon_i, \quad \mathbb{E}(\epsilon_i|D_i, X_i) = 0. \tag{7}$$

In other words, $\hat{\theta}_2 = \hat{\beta}_{1,OLS}$. The specification of (6) implies that $\hat{\theta}_2$ is unbiased and consistent for the population ATT, θ^0. We consider a situation that the finite sample distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ is well approximated by its large sample normal approximation, *i.e.*

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta^0 \\ \theta^0 \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right),$$

where σ_1^2 is the asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta^0)$ given by SEB for ATT without the knowledge of propensity scores, and σ_2^2 is the asymptotic variance of $\sqrt{n}(\hat{\theta}_2 - \theta^0)$. Under the current specification, they are obtained as

$$\sigma_1^2 = \frac{\sigma^2}{\Pr(D=1)} \left[c + \mathbb{E}\left(\frac{e(X)}{1-e(X)} | D = 1 \right) \right], \tag{8}$$

$$\sigma_2^2 = \frac{\sigma^2}{\Pr(D=1)} \left[c \cdot \frac{\mathbb{E}((1-e(X))^2|D=1)}{[\mathbb{E}(1-e(X)|D=1)]^2} + \frac{\mathbb{E}(e(X)(1-e(X))|D=1)}{[\mathbb{E}(1-e(X)|D=1)]^2} \right]. \tag{9}$$

See Appendix A for their derivations.

When $Y(1)$ and $Y(0)$ share the variance ($c = 1$), it can be shown that the OLS estimator is more efficient than the semiparametric estimator, $\sigma_2^2 < \sigma_1^2$, due to exploitation of the correct functional form of the regression equation. In contrast, if the variance of the treated outcome is higher than the variance of the control outcome ($c > 1$), the simple

OLS estimator that does not take into account the heteroskedastic errors can become less efficient than the semiparametric estimator. Specifically, we show in Appendix A that

$$\begin{aligned} &\sigma_2^2 > \sigma_1^2 \quad \text{iff } c > \frac{\Delta_1}{\Delta_2} + 1, \text{ where} \\ &\Delta_1 = \mathbb{E}\left[\frac{1}{1-e(X)}|D=1\right] - \frac{1}{\mathbb{E}(1-e(X)|D=1)} \geq 0, \\ &\Delta_2 = \frac{\mathbb{E}((1-e(X))^2|D=1)}{[\mathbb{E}(1-e(X)|D=1)]^2} - 1 \geq 0. \end{aligned} \tag{10}$$

Hence, if the degree of heteroskedasticity satisfies the condition in (10), the semiparametric estimator $\hat{\theta}_1$ is strictly preferred to the OLS estimator $\hat{\theta}_2$.

Given that c meets (10), consider applying the placebo EMCS proposed in Huber *et al.* (2013). We assume that the two estimators are centred at zero and their simulated distributions can be well approximated by bivariate Gaussian,

$$\begin{pmatrix} \hat{\theta}_1^* \\ \hat{\theta}_2^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{n_0}\begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_{12} \\ \tilde{\sigma}_{12} & \tilde{\sigma}_2^2 \end{pmatrix}\right),$$

where n_0 is the sample size of control group in the original sample. Suppose also that the propensity scores used to generate the placebo treatment coincide with the true propensity scores in the original data. Since the placebo treated group is generated from the original control group, it fails to replicate the variance of the treatment outcomes in the original data. As a result, the variances of $\sqrt{n_0}\hat{\theta}_1^*$ and $\sqrt{n_0}\hat{\theta}_2^*$ are given by the homoskedastic version ($c = 1$) of (8) and (9),

$$\tilde{\sigma}_1^2 = \frac{\sigma^2}{\widetilde{\Pr}(D=1)}\tilde{\mathbb{E}}\left[\frac{1}{1+e(X)}|D=1\right] \geq \tilde{\sigma}_2^2 = \frac{\sigma^2}{\widetilde{\Pr}(D=1)} \cdot \frac{1}{\tilde{\mathbb{E}}(1-e(X)|D=1)}, \tag{11}$$

where $\widetilde{\Pr}$ and $\tilde{\mathbb{E}}$ are the probability and expectation with respect to the sampling distribution specified in the placebo EMCS. This inequality is strict if $e(X)|D=1$ is nondegenerate. EMCS therefore incorrectly selects the OLS estimator $\hat{\theta}_2$ as a preferred estimator.

The underlying mechanism for why EMCS goes wrong is in line with Scenario 2 in the previous subsection. Even in a rather ideal situation where EMCS well replicates the unbiasedness of the estimators, artificially creating a placebo treated group from the control group in the original sample distorts the variance ordering among the estimators.

Exactly the same reasoning can also invalidate structured EMCS designs if the estimated data generating process from which the data are to be simulated ignores or fails to replicate the underlying heteroskedasticity of the potential outcome distributions.

This problem can be seen in a simple simulation study. We draw 1,000 samples from a data generating process of the form given by equation (6) with 1,000 observations per sample.[15] For each sample we run 1,000 replications of the placebo and structured EMCS procedures, considering IPW and OLS as our two estimators. This gives us 'the true MSE' for each estimator (based on the original samples) as well as 1,000 estimates of the MSE for each combination of an estimator and an EMCS design. Looking at a simple count of how many times each procedure selects the right estimator, we see that the placebo approach selects the superior estimator only 19 times (1.9% of the time) and the structured approach is little better at 30 times (3.0%). This compares with 97.6% and 100% for the placebo and structured procedures, respectively, when there is no heteroskedasticity. Of course this is a single example, and in a very stylised context; in Section 4 we will see that the performance of these methods is also poor in a 'real-world' example.

3.2.2 An Example for Scenario 3

We shift our focus to Scenario 3. We now introduce a bias in one of the estimators in the original data generating process. For this purpose, we maintain the two estimators as in the previous example, but alter the potential outcome equations from (6) with

$$\begin{aligned} \mathbb{E}(Y(1)|X=x) &= \beta_0 + \beta_1 + x'\beta_t, \\ \mathbb{E}(Y(0)|X=x) &= \beta_0 + x'\beta_c, \end{aligned} \tag{12}$$

with distinct slopes, $\beta_t \neq \beta_c$. This causes the regression specification of (7) to be misspecified so that $\hat{\theta}_2$ is no longer consistent for the population ATT, $plim_{n\to\infty}\hat{\theta}_2 = \theta_2 \neq \theta^0 = \beta_1 + \mathbb{E}(X'|D=1)(\beta_t - \beta_c)$. See, *e.g.*, Słoczyński (2017) for analytical characterizations of the bias. On the other hand, the semiparametric estimator $\hat{\theta}_1$ remains consistent and semiparametrically efficient (asymptotically attains SEB).[16] Hence, assuming that the finite sample distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ is well approximated by its asymptotic normal approximation, we have

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta^0 \\ \theta^0 + b_2 \end{pmatrix}, \frac{1}{n}\Sigma \right).$$

As we argued in Scenario 3 above, the bias in $\hat{\theta}_2$ makes $\hat{\theta}_2$ inferior to unbiased estimator $\hat{\theta}_1$ even when $\sigma_2^2 < \sigma_1^2$ if b_2 or the sample size is sufficiently large.

[15] For full details of our procedure and parameter values see Appendix B. For detailed simulation results see Appendix C.

[16] Due to the misspecification of the regression equation, the asymptotic variance of $\sqrt{n}(\hat{\theta}_2 - \theta_2)$ differs from and is generally greater than the variance of (9).

In the placebo EMCS procedure of Huber *et al.* (2013), the fact that the placebo treated group is generated from the original control group removes the misspecification issue of the OLS estimator caused by the non-parallel treatment outcome equation. Hence, $\hat{\theta}_2^*$ behaves as a correctly specified OLS estimator with homoskedastic errors, and the simulated distribution of $\hat{\theta}_2^*$ fails to replicate the bias in $\hat{\theta}_2$. Since the variance ordering in EMCS obtained in (11) is preserved in the current example, EMCS erroneously concludes that the OLS estimator $\hat{\theta}_2$ dominates the semiparametric estimator $\hat{\theta}_1$.

In case of structured EMCS procedures, if the data generating process from which Monte Carlo samples are drawn is estimated under misspecification, the structured EMCS misleads the estimator selection for exactly the same reason. For example, if one were to construct the Monte Carlo data generating process using linear regressions additive in D_i, structured EMCS will then wrongly conclude that the OLS estimator $\hat{\theta}_2$ outperforms the semiparametric estimator $\hat{\theta}_1$.

Again we perform a simple simulation, analogous to the previous subsection but modifying the potential outcome equation as given by equation (12).[17] We perform 1,000 replications of each EMCS procedure using the same estimators, and then compare in how many cases the EMCS correctly selects the estimator with the lower MSE. Again the performance of EMCS is rather poor: placebo EMCS correctly selects IPW 2.3% of the time, and structured EMCS is correct only .2% of the time.

3.3 More Than Two Estimators

Applications of EMCS often consider comparing more than two estimators. Fragility of EMCS-based estimator selection highlighted in the two estimator examples above naturally carries over to settings with more than two estimators, since ranking over multiple estimators consists of transitive pairwise rankings of any two candidate estimators.

The Monte Carlo exercises and the empirical application below consider a setting with seven estimators in the context of program evaluation with observational data. Let $(\hat{\theta}_1, \ldots, \hat{\theta}_J)$ be the pool of J candidate estimators, and let the purpose of EMCS be to obtain a *complete* ordering among these J estimators according to the MSE criteria.[18]

The internal validity criteria for EMCS introduced above, $\Pr(\hat{j}_0 = j_0)$ and $\mathbb{E}(MSE(\hat{\theta}_{\hat{j}_0}) -$

[17] For full details of our procedure and parameter values again see Appendix B. Similarly, for detailed simulation results see Appendix C.

[18] The pairwise ordering criterion defined in footnote 13 is not suitable to generate a complete ordering among several estimators since the ordering criterion described there is not transitive. For instance, consider three random variables (X, Y, Z) such that for $\epsilon \in (0, 1/4)$,

$$\Pr(X > Y > Z) = \frac{1}{2} - \epsilon, \quad \Pr(Z > X > Y) = \frac{1}{4} + \frac{\epsilon}{2}, \quad \Pr(Y > Z > X) = \frac{1}{4} + \frac{\epsilon}{2}$$

$MSE(\hat{\theta}_{j_0})$) can be straightforwardly extended to the case with several estimators. In addition, to measure similarity or dissimilarity between the true ranking and estimated rankings in EMCS, it can be of interest to look at the distribution of the Kendall's tau,

$$\hat{\tau} = \frac{2}{J(J-1)} \sum_{i<j} 1\{(\rho(i) - \rho(j))(\hat{\rho}(i) - \hat{\rho}(j)) > 0\}$$

where $\rho(j)$ and $\hat{\rho}(j)$, $j \in \{1, \ldots, J\}$, are the ranks of estimator j with respect to the true MSE and estimated MSE in EMCS, respectively. Noting $\hat{\tau} \in [-1, 1]$ has a distribution under repeated sampling, its mean or other location parameters can summarise how well EMCS can assess the relative performances among the candidate estimators.

4 Application

To demonstrate the empirical relevance of the theoretical results discussed above, and consider the extent to which they might be a problem in practice, we provide an application of EMCS procedures to a real dataset. In these data we have an experimental estimate of the treatment effect. By (initially) treating the experimental estimate as the true treatment effect, the aim is to show whether (or not) EMCS procedures can accurately recover the ranking of estimators that we see from the experiment. We first discuss the data used, then our approach, next the estimators, and finally the details of how the EMCS procedures were conducted.

4.1 Data and Context

We focus on the data on men from LaLonde (1986), used also by Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), and Smith and Todd (2001, 2005).[19] A subset of these data comes from the National Supported Work (NSW) Demonstration, which was a work experience programme that operated in the mid-1970s at 15 locations in the United States (for a detailed description of the programme see Smith and Todd, 2005). This programme served several groups of disadvantaged workers, such as women with dependent children receiving welfare, former drug addicts, ex-convicts, and school drop-outs. Unlike many similar programmes, the NSW implemented random assignment among el-

hold. Then, $\Pr(X > Y) = \Pr(Y > Z) > 1/2$ are true, but $\Pr(X > Z) < 1/2$.

[19] Recent work by Calónico and Smith (2017) highlights the effects of the NSW programme for women. Prior to this women were largely ignored in the NSW literature subsequent to LaLonde (1986) because the analysis datafile for women was not preserved.

igible participants. This random selection allowed for straightforward evaluation of the programme via a comparison of mean outcomes in the treatment and control groups.

In an influential paper, LaLonde (1986) uses the design of this programme to assess the performance of a large number of nonexperimental estimators of average treatment effects, many of which are based on the assumption of unconfoundedness. 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 datasets on the U.S. population. His key insight is that a 'good' estimator should be able to closely replicate the experimental estimate of the effect of NSW using nonexperimental data. He finds that very few of the estimates are close to this benchmark. This result motivated a large number of replications and follow-ups, and established a testbed for estimators of average treatment effects under unconfoundedness (see, *e.g.*, Heckman and Hotz 1989; Dehejia and Wahba 1999, 2002; Smith and Todd 2001, 2005; Abadie and Imbens 2011; Diamond and Sekhon 2013). Like many other papers, we use the largest of the six nonexperimental comparison groups constructed by LaLonde (1986), which he refers to as CPS-1.

4.2 Approach

In this paper we take the key insight of LaLonde (1986) one step further. We treat the NSW–CPS data from LaLonde (1986) as a finite population, with 185 treated observations and 7,660 comparison observations in our main example.[20] From this we draw 1,000 samples, each composed of 100 treated observations and 1,900 comparison observations. We then implement the estimators described below. For each sample and each estimator we compute the difference between the estimate and the 'true effect' ($1,794), which comes from the experimental estimate of the impact of NSW on earnings. With 1,000 such differences for each estimator, we can compute the MSE and other performance measures for that estimator in these data. Then, on each of the 1,000 samples, we implement the two EMCS procedures described in Section 2, and compare their performances in terms of the criteria introduced in Section 3.

One limitation of this approach is that the 'true effect' we calculate is subject to sampling error. We therefore consider a second case, where we apply the insight of Smith

[20]This comes from taking the treated sample used by Dehejia and Wahba (1999) and a trimmed version of the CPS-1 dataset. We use a logit model to predict propensity to be in the experimental data (either as treatment or control) versus being in the CPS-1 data. We then drop all CPS-1 observations with propensity scores below the minimum or above the maximum in the experimental data. This is the trimmed CPS-1 dataset, which we then combine with the NSW treated observations from Dehejia and Wahba (1999).

and Todd (2005) that the *control* sample from the NSW can be compared to the same non-experimental comparison group. The NSW control sample includes people who were selected in the same way as those actually treated, but who were randomised out of treatment. Now we know that the 'true effect' is a precise zero, since the control sample did not actually receive treatment. Thus, we have an original dataset of 142 'treated' observations (who in reality received no treatment) and 7,467 comparison units.[21] Again we draw samples by selecting 100 treated observations and 1,900 comparison observations from this population, with the true effect being precisely zero in each sample, and then perform EMCS on these samples.

Another possible worry might be that our example applies estimators that are suitable under unconfoundedness, *i.e.* when potential outcomes are independent of treatment assignment, conditional on observables. One of the main conclusions in Smith and Todd (2005) is that such conditional independence is not plausible in the context of the NSW–CPS data. To address this concern, we take a third approach. The basic idea is to construct a population similar to the NSW–CPS data where unconfoundedness holds by construction, and then draw samples from this. We begin with a trimmed version of the Dehejia and Wahba (1999) dataset used in the first case. Next, we perform 4-nearest neighbour matching (with replacement) to impute the 'missing' potential outcome for each observation. This is our new population. We then draw random subsamples of 2,000 observations (covariates, potential outcomes, and propensity scores) from the data, add a logistic error to the estimated propensity score, and assign to treatment the individuals in the top quarter of the adjusted propensity score distribution (giving 500 treated and 1,500 nontreated in each sample). By construction treatment is now independent of potential outcomes. We also use our knowledge of potential outcomes of all individuals to calculate the true value (or 'pseudo-true' value) of ATT.[22] We implement EMCS on the samples drawn in this way.

[21] This comes from taking the control sample used by Smith and Todd (2005) and a trimmed version of the CPS-1 dataset. The size of the comparison subsample (7,467 observations) is different than in the first case (7,660 observations) because our logit model – which we use to predict propensity to be in the experimental data and to trim – is now fitted on a different dataset; while the comparison units remain the same, the treated and control subsamples are different.

[22] More precisely, we use the representation of ATT as $\frac{1}{\Pr(D=1)} \cdot \mathbb{E}\left[e(x) \cdot (Y(1) - Y(0))\right]$ to calculate its value in our application. By design, we know both potential outcomes for all units and we set $\Pr(D = 1)$ at 25%. We approximate $e(x)$ for all units with the empirical probabilities of treatment in 10,000 samples from the original data generating process. This 'pseudo-true' value of ATT is equal to –$405.

4.3 Estimators

In all our simulations we study the impact of the NSW programme on earnings in 1978. We consider seven nonexperimental estimators: linear regression, Oaxaca–Blinder, inverse probability weighting (IPW), doubly-robust regression, uniform kernel matching, nearest neighbour matching, and bias-adjusted nearest neighbour matching. For details see Appendix D. In each case we focus on the average treatment effect on the treated (ATT), unless a given method does not allow for heterogeneity in effects (in which case we estimate the overall effect of treatment). As noted above, all of these estimators are based on the assumption of unconfoundedness.

We use a single set of control variables in all our simulations. Following Dehejia and Wahba (1999) and Smith and Todd (2005), we control for age, age squared, age cubed, education, education squared, whether a high school dropout, whether married, whether black, whether Hispanic, earnings in months 13–24 prior to randomization, earnings in 1975, nonemployment in months 13–24 prior to randomization, nonemployment in 1975, and the interaction of education and earnings in months 13–24 prior to randomization.

We conduct all our simulations in Stata and use several user-written commands in our estimation procedures: `nnmatch` (Abadie *et al.* 2004), `oaxaca` (Jann 2008), and `psmatch2` (Leuven and Sianesi 2003).

4.4 Procedures

In Section 2 we noted that for the placebo design we require some choice of π and λ, where λ determines the degree of covariate overlap between the 'placebo treated' and 'placebo control' observations and π determines the proportion of the 'placebo treated'. We choose π to ensure that the proportion of the 'placebo treated' observations in each placebo EMCS replication is equal to the proportion of treated units in the sample.[23] We also follow Huber *et al.* (2013) in choosing $\lambda = 1$ as well as in using a logit model to estimate the propensity score.

The structured design requires more choices, in particular how we specify the joint probability distribution as the product of the marginal distribution for treatment status and some conditional distributions. As discussed in Section 2, we begin each structured EMCS replication by generating a fixed number of treated and nontreated observations to match the numbers in the sample. We then order the covariates, regress each covariate on

[23]It should be noted, however, that the way these datasets were constructed by LaLonde (1986) results in samples that are best described as choice-based. More precisely, the treatment and control groups are heavily overrepresented relative to their population proportions. See Smith and Todd (2005) for a further discussion of this issue.

the preceding covariates (using logistic regression for binary covariates), and use this to define the conditional distribution for that covariate. In EMCS replications the covariates are then drawn in the same order, from the appropriate conditional distribution. Full details of the procedure are provided in Appendix E.

5 Results

We now describe the results of our tests of the two EMCS procedures – placebo and structured – in the context of our real-world data. As described in Subsection 4.2, we perform three sets of tests. First, we apply the two procedures to the NSW treatment sample, combined with the CPS-1 comparison dataset. We find performance of the procedures to be poor when it comes to finding the estimator with the lowest bias. When we study MSE (*i.e.*, account also for variance), performance is better. This is because the rankings of estimators are mainly being driven by the variance, and both EMCS methods do well at replicating the variance components. However, given this, we also test a simple bootstrap procedure and find that it is more effective at picking the best estimator. Then, we follow Smith and Todd (2005) in using the NSW controls as our 'treated' sample instead: now the effect we intend to estimate must be zero for sure, removing worries that poor performance might be an artefact caused by sampling uncertainty around the true effect. We find that the previous results are maintained. Finally, we use an adjusted version of the original data, constructed so that conditional independence necessarily holds, to allay concerns that poor performance is driven by a context in which unconfoundedness may not hold. Again we find that the EMCS procedures do not perform well on bias, and are better on MSE, although here the bootstrap does not clearly dominate.

5.1 Testing EMCS in the NSW Data

Our first results using 'real-world' data focus on the variant of the original NSW treatment sample constructed by Dehejia and Wahba (1999), combined with a trimmed version of the CPS-1 comparison dataset. We create 1,000 samples from the original dataset by sampling 100 treated and 1,900 nontreated observations from the 185 possible treated and 7,660 comparison units in the original data. We implement the two EMCS procedures 1,000 times on each of the 1,000 samples, giving a total of 1,000,000 replications for each EMCS procedure. In each replication we implement the seven estimators described earlier, and measure how well the two EMCS procedures help us assess the relative performance of the estimators. We might measure performance of an estimator in terms of

absolute bias or MSE (which also takes into account its variance). Performance ('internal validity') of EMCS is then measured by how well the EMCS procedure replicates these features of an estimator in the original samples. In Section 3, we described two measures of EMCS performance suitable for when we have many estimators:

1. the average regret, *i.e.* average difference in absolute bias/MSE between the estimator selected by EMCS and the estimator with the actual minimum absolute bias/MSE; and

2. the average Kendall's tau (Kendall's rank correlation coefficient), which measures the similarity between the ranking of estimators suggested by EMCS and the 'true' ranking from the original samples.[24]

For ease of interpretation, it is also useful to normalise the values of average regret. Our discussion below focuses on the average regret as a percentage of the minimum value of absolute bias/MSE. However, we also consider an alternative normalisation, where we divide the average regret for a given EMCS procedure by the average regret for random selection of estimators (which we discuss further below). Finally, we also consider an additional measure, which is straightforward to interpret, namely

3 the average correlation in absolute bias/MSE (rather than in the rankings, as given by Kendall's tau).

In each case the comparison is between what the EMCS procedure suggests and the results from taking the 'true effect' in the original data, and then calculating the absolute bias/MSE of each estimator across the 1,000 samples.

To provide a benchmark for the performance of EMCS, we also include results from two other procedures. In the first we simply draw nonparametric bootstrap samples.[25] We can then compare estimators on variance, and also see how the ranking compares to the ranking on MSE from the original samples. In the second we do not create any samples, but simply rank estimators randomly. This provides a 'worst-case' benchmark: suppose a researcher knows nothing at all about performance and just picks an estimator blindly, how would they do? Here we cannot compute a result for the correlation, but can for average regret and Kendall's tau. Table 1 shows the results from these simulations. Appendix F provides further details.

The first result is that performance of both EMCS procedures in terms of bias is very poor. The average regret in terms of absolute bias, as a percentage of the absolute bias for

[24] See Subsection 3.3 for the precise calculation.

[25] Precisely, we sample with replacement, and draw replication samples of the same size as the original sample.

Table 1: **Internal validity of EMCS using different performance metrics**

EMCS approach	Placebo	Structured	Bootstrap	Random
Absolute bias (minimum = 16)				
Average regret (% of minimum)	3,067	3,766	—	1,184
Average regret (% of random)	259.0	318.1	—	100.0
Average Kendall's tau	–.214	–.372	—	0
Average correlation	–.437	–.505	—	—
Mean squared error (minimum = 512,322)				
Average regret (% of minimum)	18.2	16.3	7.9	141.9
Average regret (% of random)	12.8	11.5	5.6	100.0
Average Kendall's tau	.599	.635	.828	0
Average correlation	.647	.791	.809	—
Variance (minimum = 454,278)				
Average regret (% of minimum)	2.7	11.2	1.9	148.0
Average regret (% of random)	1.8	7.6	1.3	100.0
Average Kendall's tau	.767	.812	.883	0
Average correlation	.895	.920	.862	—

Notes: 'EMCS approach' denotes the way in which the empirical Monte Carlo samples were generated. 'Placebo' and 'Structured' generate samples using the placebo and structured approaches described in Section 2. 'Bootstrap' generates nonparametric bootstrap samples by sampling with replacement the same number of observations as the original data. 'Random' does not generate samples but instead randomly assigns *rankings* to the estimators (hence statistics are only available for the performance metrics based on rankings). The absolute bias, mean squared error, and variance are features of estimators. The 'minimum' value for each feature is its lowest value among our estimators in the original data generating process (*i.e.* we have one value of each feature for each estimator in the 'original samples' and we report the lowest of these values). See Appendix F for more details. Four performance measures are used for each of these statistics. 'Average regret' measures the average increase in the statistic from choosing the estimator actually selected by the EMCS approach rather than the estimator with the minimum value of this statistic, as a percentage of *(i)* that minimum value or *(ii)* the average regret for random selection of estimators. 'Average Kendall's tau' measures the average correlation in the ranking of estimators provided by the EMCS approach relative to the ranking in the original samples. 'Average correlation' measures the average correlation in the actual values of the statistic (rather than the ranking) provided by the EMCS approach relative to the values in the original samples. All averages are taken with respect to 1,000 original samples; for each sample, a separate simulation study was conducted. We do not report any performance measures for absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. The results for random selection of estimators are analytical; instead of actually generating random rankings, we report the known values of expected Kendall's tau (zero) and expected regret with random rankings. The latter value is equal to the average regret across estimators, with an equal probability of each estimator to be selected as 'best'.

the best estimator, is 3,067% (3,766%) for placebo (structured), *i.e.* an order of magnitude larger than the minimum value. It is worse than choosing completely randomly, which would be 1,184% worse than the best estimator. Looking at the ranking across estimators, the average Kendall's tau is –.21 (–.37) for placebo (structured). So the rankings produced by EMCS are, on average, *negatively* correlated with the ranking in the original samples. This is worse than random, which gives .00. The same pattern is seen in the average correlation coefficients for absolute bias, which are –.44 (–.51).

A researcher might be interested in knowing about performance of estimators in terms of MSE rather than only considering bias. Here EMCS performs significantly better. The average regret for placebo (structured) is now only 18% (16%), much better than random (142%). Similarly, average Kendall's tau is now .60 and .64 for placebo and structured, respectively, much better than .00 for random. The lowest panel of Table 1 shows that this is driven by the much better performance in replicating the variances. Since the rankings here are mostly determined by the variance, being able to reproduce variances significantly improves the measures of performance relative to the metrics based on absolute bias.

However, looking at our other benchmark case – the bootstrap – we see that it outperforms both EMCS methods in terms of MSE. Average regret is lower at 7.9%, and the average Kendall's tau is much higher at .83. Given that MSE performance for EMCS is driven by the variance components, this does not seem surprising. The bootstrap is a simpler procedure than the two EMCS methods, and its ability to help us understand the variability of estimators is well known. It therefore seems like a potentially valuable path which has fewer design choices than EMCS.

5.2 Removing Sampling Error from the 'True Effect'

The previous subsection calculated the MSE for each estimator by comparing the value of the estimate in each sample to a 'true effect' measured using the experiment. One concern might be that the estimate from the experiment is subject to sampling error, and this might somehow negatively affect our performance measures for EMCS. To test this, we now use as our 'treated' observations the NSW control sample from Smith and Todd (2005). Since these individuals were selected for the programme in the same way as those actually treated, but were then randomised out, the actual treatment effect for them is precisely zero. We therefore repeat the exercise on these data, again implementing the two EMCS procedures 1,000 times on each of the 1,000 original samples. Table 2 documents the results. Appendix F provides further details.

Table 2: **Internal validity of EMCS, ensuring no sampling error in the treatment effect**

EMCS approach	Placebo	Structured	Bootstrap	Random
Absolute bias (minimum = 954)				
Average regret (% of minimum)	30.2	41.6	—	19.0
Average regret (% of random)	158.9	219.1	—	100.0
Average Kendall's tau	–.274	–.466	—	0
Average correlation	–.418	–.822	—	—
Mean squared error (minimum = 1,222,627)				
Average regret (% of minimum)	22.5	32.1	17.4	94.4
Average regret (% of random)	23.9	34.0	18.4	100.0
Average Kendall's tau	.645	.549	.809	0
Average correlation	.843	.814	.746	—
Variance (minimum = 296,671)				
Average regret (% of minimum)	1.2	5.0	9.0	262.6
Average regret (% of random)	.5	1.9	3.4	100.0
Average Kendall's tau	.950	.665	.762	0
Average correlation	.959	.934	.833	—

Notes: 'EMCS approach' denotes the way in which the empirical Monte Carlo samples were generated. 'Placebo' and 'Structured' generate samples using the placebo and structured approaches described in Section 2. 'Bootstrap' generates nonparametric bootstrap samples by sampling with replacement the same number of observations as the original data. 'Random' does not generate samples but instead randomly assigns *rankings* to the estimators (hence statistics are only available for the performance metrics based on rankings). The absolute bias, mean squared error, and variance are features of estimators. The 'minimum' value for each feature is its lowest value among our estimators in the original data generating process (*i.e.* we have one value of each feature for each estimator in the 'original samples' and we report the lowest of these values). See Appendix F for more details. Four performance measures are used for each of these statistics. 'Average regret' measures the average increase in the statistic from choosing the estimator actually selected by the EMCS approach rather than the estimator with the minimum value of this statistic, as a percentage of *(i)* that minimum value or *(ii)* the average regret for random selection of estimators. 'Average Kendall's tau' measures the average correlation in the ranking of estimators provided by the EMCS approach relative to the ranking in the original samples. 'Average correlation' measures the average correlation in the actual values of the statistic (rather than the ranking) provided by the EMCS approach relative to the values in the original samples. All averages are taken with respect to 1,000 original samples; for each sample, a separate simulation study was conducted. We do not report any performance measures for absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. The results for random selection of estimators are analytical; instead of actually generating random rankings, we report the known values of expected Kendall's tau (zero) and expected regret with random rankings. The latter value is equal to the average regret across estimators, with an equal probability of each estimator to be selected as 'best'.

Table 3: **Internal validity of EMCS, ensuring unconfoundedness holds**

EMCS approach	Placebo	Structured	Bootstrap	Random
Absolute bias (minimum = 68)				
Average regret (% of minimum)	593.5	670.1	—	522.0
Average regret (% of random)	113.7	128.4	—	100.0
Average Kendall's tau	–.003	.057	—	0
Average correlation	–.048	.217	—	—
Mean squared error (minimum = 340,300)				
Average regret (% of minimum)	260.1	89.7	276.1	682.9
Average regret (% of random)	38.1	13.1	40.4	100.0
Average Kendall's tau	.737	.729	.632	0
Average correlation	.943	.790	.778	—
Variance (minimum = 137,574)				
Average regret (% of minimum)	0	.3	0	1,631
Average regret (% of random)	0	0	0	100.0
Average Kendall's tau	.768	.727	.752	0
Average correlation	.951	.820	.820	—

Notes: 'EMCS approach' denotes the way in which the empirical Monte Carlo samples were generated. 'Placebo' and 'Structured' generate samples using the placebo and structured approaches described in Section 2. 'Bootstrap' generates nonparametric bootstrap samples by sampling with replacement the same number of observations as the original data. 'Random' does not generate samples but instead randomly assigns *rankings* to the estimators (hence statistics are only available for the performance metrics based on rankings). The absolute bias, mean squared error, and variance are features of estimators. The 'minimum' value for each feature is its lowest value among our estimators in the original data generating process (*i.e.* we have one value of each feature for each estimator in the 'original samples' and we report the lowest of these values). See Appendix F for more details. Four performance measures are used for each of these statistics. 'Average regret' measures the average increase in the statistic from choosing the estimator actually selected by the EMCS approach rather than the estimator with the minimum value of this statistic, as a percentage of *(i)* that minimum value or *(ii)* the average regret for random selection of estimators. 'Average Kendall's tau' measures the average correlation in the ranking of estimators provided by the EMCS approach relative to the ranking in the original samples. 'Average correlation' measures the average correlation in the actual values of the statistic (rather than the ranking) provided by the EMCS approach relative to the values in the original samples. All averages are taken with respect to 1,000 original samples; for each sample, a separate simulation study was conducted. We do not report any performance measures for absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. The results for random selection of estimators are analytical; instead of actually generating random rankings, we report the known values of expected Kendall's tau (zero) and expected regret with random rankings. The latter value is equal to the average regret across estimators, with an equal probability of each estimator to be selected as 'best'.

Our conclusions are similar to those in the previous subsection. In terms of absolute bias, the average regret is much lower than previously, at 30% (42%) for placebo (structured), although this is mostly driven by a large increase in the minimum value of absolute bias.[26] Also, this is still worse than choosing at random (19%). As before, the average Kendall's tau is negative for placebo (structured) at –.27 (–.47), which is worse than random (.00) as well. On MSE performance is better, with average regret of 23% (32%) and average Kendall's tau of .65 (.55). These are much better than random (94% and .00), but worse than bootstrap (17% and .81).

5.3 Ensuring Unconfoundedness Holds

Another potential concern is whether the conditional independence assumption holds. Here we take the approach described in Subsection 4.2 to generate 1,000 samples in which conditional independence holds by construction. Then, we implement the two EMCS procedures 500 times on each of these samples.[27] Table 3 displays the results. Appendix F provides further details.

The previous results are broadly maintained even after ensuring conditional independence. In terms of absolute bias the performance of both EMCS approaches is similar to random. In terms of MSE both procedures perform better than random selection of estimators and also marginally better than bootstrap. Average regret in terms of MSE is worse than in the first case, though average Kendall's tau is a little higher, so it is also not obvious that contexts where conditional independence holds should necessarily see better performance of EMCS procedures.

6 Discussion

Advances in econometrics have left the empirical researcher blessed with a wealth of possible treatment effect estimators from which to choose. They have not yet provided clear

[26] In Subsection 5.1, the minimum absolute bias in the original data was equal to 16 (nearest neighbour matching); now, it is equal to 954 (Oaxaca–Blinder). See also Appendix F. As noted by Smith and Todd (2005), it is much more difficult to recover the true effect of NSW in these data. The fact that the difference in the values of average regret between Tables 1 and 2 is driven by the minimum absolute bias can also be seen by normalising these values by the average regret for random selection of estimators. In the first simulation study, the average regret for placebo (structured) is approximately 2.6 (3.2) times larger than for random; in the second simulation study, this metric is approximately 1.6 (2.2) times larger for placebo (structured) than for random. While these values continue to be smaller in the second simulation study, their overall magnitudes are similar in both cases.

[27] The smaller number of replications per sample follows from the significant computational burden of this simulation study, which originates from the larger size of the treated subsamples (500 instead of 100 observations per sample).

guidance on which of these estimators should be preferred in which context. In this paper we studied two proposals which suggest an approach to choosing an appropriate estimator for a given context. The first approach (placebo) suggests a way to introduce placebo treatments to some control observations in a dataset, and studies how well estimators can pick up the true zero effect. The second approach (structured) creates data from a known DGP whose parameters are estimated from features of the original data, and studies how well estimators can pick up the implied true effect in the DGP.

We showed theoretically that both approaches can only be guaranteed to work under rather restrictive conditions: when they can correctly reproduce the biases and the ordering of the variances of estimators. We show simple practical cases where one or other of these might fail, and give an example of the consequences based on simulations from an artificial DGP. To provide a real-world example, we also implement the EMCS procedures in the NSW–CPS data, where we know the 'true effect' of the programme. This allows us to compute actual performance of the estimators in samples from the original data, and compare this to what EMCS would suggest if applied to these samples. We show that in this example EMCS performs badly on ordering estimators in terms of absolute bias, and the estimator it suggests is often many times worse than the best (or even than selecting randomly). In this example both EMCS procedures perform much better in terms of MSE because reproducing the variance term turns out to drive the MSE in these data. But, this leads the methods to be no better (and sometimes significantly worse) than a simple bootstrap procedure.

These results are unfortunate, but nevertheless important. There remains no silver bullet that can assist empirical researchers with the 'right' or 'best' estimator for a particular context. In the absence of a clear choice driven by research design, the best advice at this stage is likely to be implementing a number of estimators, and then considering the range of estimates provided, as Busso *et al.* (2014) also suggest.

One possible future alternative, recently proposed, is *synth-validation* (Schuler *et al.*, 2017). This approach is related to cross-validation and is based on estimating 'the estimation error of causal inference methods applied to a given dataset'. The authors provide simulations which suggest that this 'lowers the expected estimation error relative to consistently using any single method'. Further work is needed to test how general this approach is, and whether it can reliably guide researchers in selecting estimators.

References

ABADIE, A., D. DRUKKER, J. L. HERR, AND G. W. IMBENS (2004): "Implementing Matching Estimators for Average Treatment Effects in Stata," *Stata Journal*, 4, 290–311.

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

——— (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11.

——— (2016): "Matching on the Estimated Propensity Score," *Econometrica*, 84, 781–807.

ANGRIST, J. D. AND A. B. KRUEGER (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter and D. Card, Elsevier, vol. 3, 1277–1366.

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., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–275.

BLUNDELL, R. AND M. COSTA DIAS (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640.

BODORY, H., L. CAMPONOVO, M. HUBER, AND M. LECHNER (2018): "The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators," *Journal of Business & Economic Statistics*, forthcoming.

BREWER, M., T. F. CROSSLEY, AND R. JOYCE (2018): "Inference with Difference-in-Differences Revisited," *Journal of Econometric Methods*, 7.

BUSSO, M., J. DINARDO, AND J. MCCRARY (2009): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," Unpublished.

——— (2014): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 96, 885–897.

CALÓNICO, S. AND J. SMITH (2017): "The Women of the National Supported Work Demonstration," *Journal of Labor Economics*, 35, S65–S97.

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

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

——— (2002): "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161.

DIAMOND, A. AND J. S. SEKHON (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.

DÍAZ, J., T. RAU, AND J. RIVERA (2015): "A Matching Estimator Based on a Bilevel Optimization Problem," *Review of Economics and Statistics*, 97, 803–812.

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

FRÖLICH, M., M. HUBER, AND M. WIESENFARTH (2017): "The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation," *Computational Statistics & Data Analysis*, 115, 91–102.

GRAHAM, B. S., C. CAMPOS DE XAVIER PINTO, AND D. EGEL (2012): "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies*, 79, 1053–1079.

——— (2016): "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business & Economic Statistics*, 31, 288–301.

HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

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 V. J. HOTZ (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.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

HUBER, M., M. LECHNER, AND G. MELLACE (2016): "The Finite Sample Performance of Estimators for Mediation Analysis Under Sequential Conditional Independence," *Journal of Business & Economic Statistics*, 34, 139–160.

HUBER, M., M. LECHNER, AND C. WUNSCH (2013): "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics*, 175, 1–21.

IMBENS, G. W. AND J. M. WOOLDRIDGE (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., G. PICONE, M. SALM, AND J. G. TROGDON (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 A. STRITTMATTER (2016): "Practical Procedures to Deal with Common Support Problems in Matching Estimation," *Econometric Reviews*, forthcoming.

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

LEE, W.-S. (2013): "Propensity Score Matching and Variations on the Balancing Test," *Empirical Economics*, 44, 47–80.

LEUVEN, E. AND B. SIANESI (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 M. DAVIDIAN (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 R. TCHERNIS (2009): "On the Specification of Propensity Scores, with Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics*, 27, 397–415.

SCHULER, A., K. JUNG, R. TIBSHIRANI, T. HASTIE, AND N. SHAH (2017): "Synth-Validation: Selecting the Best Causal Inference Method for a Given Dataset," Unpublished.

SŁOCZYŃSKI, T. (2017): "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," Unpublished.

SŁOCZYŃSKI, T. AND J. M. WOOLDRIDGE (2018): "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 34, 112–133.

SMITH, J. A. AND P. E. TODD (2001): "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review: Papers & Proceedings*, 91, 112–118.

——— (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353.

WOOLDRIDGE, J. M. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301.

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.

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

Appendix

A Theory

Derivations of (8) and (9): A general expression of SEB for ATT in the absence of knowledge of the propensity score is given by

$$SEB_{ATT} = \frac{1}{\Pr(D=1)} E\left[Var(Y(1)|X) + \frac{e(X)}{1-e(X)} Var(Y(0)|X) + (\tau(X) - \theta^0)^2 | D = 1\right].$$

Plugging the current specifications for $Var(Y(1)|X)$ and $Var(Y(0)|X)$ and noting $\tau(X) = \theta^0$ for all X, the expression of (8) follows.

By the partialling out argument of the least squares regression and the linear probability specification of the propensity score, the asymptotic variance of $\sqrt{n}(\hat{\theta}_2 - \theta^0)$ can be written as

$$\begin{aligned} Avar(\sqrt{n}(\hat{\theta}_2 - \theta^0)) &= \frac{E(\epsilon^2 (D - e(X))^2)}{[E((D-e(X))^2)]^2} \\ &= \frac{E\left[Var(Y(1)|X)(1-e(X))^2 e(X) + Var(Y(0)|X) e(X)^2 (1-e(X))\right]}{[E(e(X)(1-e(X))]^2} \\ &= \frac{\sigma^2}{\Pr(D=1)} \cdot \frac{E\left[c(1-e(X))^2 + e(X)(1-e(X)) | D=1\right]}{[E(1-e(X)|D=1)]^2}, \end{aligned}$$

where the third line follows from Bayes rule applied to each denominator and numerator. □

Proof of (10): Rewrite (8) and (9) as

$$\sigma_1^2 = \frac{\sigma^2}{\Pr(D=1)} \left[c - 1 + E\left(\frac{1}{1-e(X)} | D = 1\right)\right], \tag{13}$$

$$\sigma_2^2 = \frac{\sigma^2}{\Pr(D=1)} \left[(c-1) \cdot \frac{E((1-e(X))^2 | D=1)}{[E(1-e(X)|D=1)]^2} + \frac{1}{E(1-e(X)|D=1)}\right]. \tag{14}$$

Hence, we obtain

$$\sigma_2^2 - \sigma_1^2 = \frac{\sigma^2}{\Pr(D=1)} \left[(c-1)\Delta_2 - \Delta_1\right], \tag{15}$$

$\Delta_1 \geq 0$ by Jensen's inequality, and $\Delta_2 \geq 0$. Hence, the condition for $\sigma_2^2 > \sigma_1^2$ follows as in (10). □

B Stylised Simulations: Design

Here we provide further details on the parameters and procedures for the stylised simulations described in Subsection 3.2.

B.1 Details of Simulation for Scenario 2

For each sample we generate 1,000 observations, and for each observation draw a covariate x from a truncated standard normal distribution with the left truncation point at –4 and the right truncation point at 6.

Propensity score $e(x)$ is then constructed as

$$e(x) = .4 + .1x. \tag{16}$$

For each observation we draw a random number from a standard uniform distribution, and assign treated status, $D = 1$, if $e(x)$ exceeds that random number.

We next generate an unobservable ϵ drawn from a normal distribution with mean zero. Since Scenario 2 is the heteroskedastic case, the standard deviation for those not treated is $\sigma_0 = .5$, while for those who are treated it is $\sigma_1 = 1.5$.[28]

Finally the outcome Y is generated as

$$Y = 3 + .5D + .5X + \epsilon, \tag{17}$$

and hence ATT is equal to .5.

This completes the generation of a Scenario 2 sample, which can then be used to implement the two EMCS procedures described in Section 2. For each EMCS design, we consider 1,000 samples and 1,000 replications per sample.

In the placebo design, we additionally require some choice of π and λ, where λ determines the degree of covariate overlap between the 'placebo treated' and 'placebo control' observations and π determines the proportion of the 'placebo treated'. We choose π to ensure that the proportion of the 'placebo treated' observations in each placebo EMCS replication is equal to the proportion of treated units in the sample. We follow Huber *et al.* (2013) in choosing $\lambda = 1$. We also use a linear model to estimate the propensity score, as this corresponds to the true model in equation (16).

In the structured design, we first estimate the mean and variance of X in a given sample, conditional on treatment status. We also regress Y on D and X, excluding the

[28] In the benchmark case (Scenario 1), mentioned at the end of Subsubsection 3.2.1, $\sigma_0 = \sigma_1 = .5$.

interaction of D and X. Next, in the simulated dataset, X is drawn from a normal distribution with mean and variance conditional on treatment status and equal to the estimates above. Whenever the draw of X lies outside the support observed in the data, conditional on treatment status, the observation is replaced with the limit point of the support. Finally, the simulated outcome, Y, is generated in two steps. In the first step, we calculate its conditional mean based on the estimated coefficients from the regression above. In the second step, the simulated outcome is determined as a draw from a normal distribution with the conditional mean determined above and the variance that is equal to the variance of the residuals in the regression model estimated on the original data.[29] Again, we replace extreme values with the limit of the support, conditional on treatment status.

We use two estimators in our stylised simulations: linear regression (OLS) and inverse probability weighting (IPW). In the latter case, we first estimate the propensity score using a linear model, as this corresponds to the true model in equation (16), and then use inverse weighting with normalised weights to estimate the ATT.

B.2 Details of Simulation for Scenario 3

A similar procedure to that detailed in the previous subsection is followed. Two changes are made. First, we now have homoskedasticity so $\sigma_0 = \sigma_1 = .5$. Second, in each sample, we now generate the outcome Y as

$$Y = 3 + .5D + .5X + .5XD + \epsilon, \tag{18}$$

and hence ATT is equal to $.5 + .5 \cdot \mathbb{E}(X|D = 1)$.[30]

The source of misspecification of the structured design in Scenario 3 is in its failure to account for the interaction of D and X when generating the simulated outcomes.

[29] Thus, by using a single value of variance for both treated and control units, we fail to account for heteroskedasticity of the potential outcome equations. This is the source of misspecification of the structured design in Scenario 2.

[30] In practice, we estimate $\mathbb{E}(X|D = 1)$ using the mean of X for the treated observations in 1,000 samples from the true data generating process. As a result, ATT is equal to (approximately) .625.

C Stylised Simulations: Detailed Results

Table C1: **Simulation results for Scenario 1 in Section 3**

	Absolute bias	RMSE	SD
Original samples			
IPW	.000	.034	.034
OLS	.000	.032	.032
Placebo			
IPW	.002	.044	.044
	(.001)	(.002)	(.002)
OLS	.001	.042	.042
	(.001)	(.002)	(.002)
Structured			
IPW	.007	.035	.034
	(.005)	(.002)	(.001)
OLS	.001	.033	.033
	(.001)	(.001)	(.001)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in Appendix B. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. In each case, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as elsewhere in the paper).

Table C2: **Simulation results for Scenario 2 in Section 3**

	Absolute bias	RMSE	SD
Original samples			
IPW	.003	.079	.079
OLS	.002	.080	.080
Placebo			
IPW	.002	.044	.044
	(.001)	(.002)	(.002)
OLS	.001	.042	.042
	(.001)	(.002)	(.002)
Structured			
IPW	.016	.070	.067
	(.012)	(.005)	(.003)
OLS	.010	.067	.066
	(.008)	(.003)	(.003)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in Appendix B. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. In each case, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as elsewhere in the paper).

Table C3: **Simulation results for Scenario 3 in Section 3**

	Absolute bias	RMSE	SD
Original samples			
IPW	.001	.044	.044
OLS	.081	.089	.037
Placebo			
IPW	.002	.043	.043
	(.001)	(.002)	(.002)
OLS	.001	.042	.042
	(.001)	(.002)	(.002)
Structured			
IPW	.011	.040	.038
	(.009)	(.004)	(.001)
OLS	.003	.037	.036
	(.003)	(.001)	(.001)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in Appendix B. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. In each case, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as elsewhere in the paper).

D Empirical Application: Estimators

We use seven estimators in our empirical application.

1. Linear regression (OLS).

2. Oaxaca–Blinder – we follow Kline (2011) in using the Oaxaca–Blinder decomposition to estimate the ATT.

3. Inverse probability weighting (IPW) – we first estimate the propensity score using a logit model, and then use inverse weighting with normalised weights to estimate the ATT.

4. Doubly-robust regression – as in Wooldridge (2007) and Słoczyński and Wooldridge (2018), we use the inverse-probability-weighted regression-adjustment (IPWRA) estimator. This is effectively a combination of the two estimators above, IPW and Oaxaca–Blinder. It satisfies the double robustness property.

5. Uniform kernel matching – we first estimate the propensity score using a logit model, and then match on propensity scores using a uniform kernel. We select the bandwidth on the basis of leave-one-out cross-validation (as in Busso *et al.* 2009 and Huber *et al.* 2013), using a search grid $.005 \times 1.25^{g-1}$ for $g = 1, 2, \ldots, 15$. The computational time of doing this for each replication is prohibitive. Consequently we calculate this once for each original sample, and use the recovered optimal bandwidth in all EMCS replications for that sample.

6. Nearest neighbour matching – nearest neighbour matching on propensity scores, which are first estimated from a logit regression, with matching on the single nearest neighbour. We match with replacement; if there are ties, all of the tied observations are used.

7. Bias-adjusted nearest neighbour matching – as above, but correcting bias as in Abadie and Imbens (2011), since nearest neighbour matching is not $\sqrt{n}$-consistent.

E Empirical Application: Structured EMCS Procedure

Here we detail precisely the procedure followed to implement the structured EMCS in our empirical application. As noted previously, we begin each structured EMCS replication by generating a fixed number of treated and nontreated observations to match the number in the sample. We then draw an employment status pair of `u74` and `u75` (nonemployment in months 13–24 prior to randomization and nonemployment in 1975), conditional on treatment status, to match the observed conditional joint probability. For individuals who are employed in only one period, an income is drawn from a log normal distribution conditional on treatment and employment statuses, with mean and variance calibrated to the respective conditional moments in the data. Where individuals are employed in both periods a joint log normal distribution is used, again conditioning on treatment status. In all cases, whenever the income draw in a particular year lies outside the relevant support observed in the data, conditional on treatment status, the observation is replaced with the limit point of the empirical support, as also suggested by Busso *et al.* (2014).

We model the joint distribution of the remaining control variables as a particular tree-structured conditional probability distribution, so that we can better fit the correlation structure in the data. The process for generating these covariates is as follows:

1. The covariates are ordered: treatment status, employment statuses, income in each period, whether a high school dropout (`nodegree`), education (`educ`), age, whether married, whether black, and whether Hispanic. This ordering is arbitrary, and a similar correlation structure would be generated if the ordering were changed.

2. Using the sample on which the EMCS is being performed, each covariate from `nodegree` onward is regressed on all the covariates listed before it (we use the logit model for binary variables).[31] These regressions are not to be interpreted causally; they simply give the conditional mean of each variable given all preceding covariates.

3. In the simulated dataset, covariates are drawn sequentially in the same order. For binary covariates a temporary value is drawn from a $\mathcal{U}[0,1]$ 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 step 2. Age and

[31] One exception is `educ` which is regressed on the prior listed covariates conditional on `nodegree`. Clearly, it is not possible for a high school dropout to have twelve years of schooling or more; it is also not possible for a non-dropout to have less than twelve years of schooling.

education are drawn from a normal distribution whose mean depends on the other covariates and whose variance is equal to that of the residuals from the relevant model. Again, we replace extreme values with the limit of the support, conditional on treatment status (for education, also conditional on dropout status).

The outcome studied is earnings in 1978, `re78`. The simulated outcome, Y_i for individual i, is then generated in two steps. In the first step, we generate a conditional mean using the parameters of a flexible linear model fitted to the sampled data. Precisely, we estimate (δ_0, δ_1) from the following linear model:

$$\mathrm{E}(Y|D, \mathbf{X}) = (1 - D)\mathbf{X}\delta_0 + D\mathbf{X}\delta_1. \tag{19}$$

The predicted conditional mean in the replication is then calculated using the estimated coefficients $(\hat{\delta}_0, \hat{\delta}_1)$ from above, and the simulated treatment status and covariates, D_i and $\mathbf{X}_i$. In the second step, the simulated outcome, Y_i, is determined as a draw from a normal distribution with the estimated conditional mean $(1 - D_i)\mathbf{X}_i\hat{\delta}_0 + D_i\mathbf{X}_i\hat{\delta}_1$ and the variance that is fitted to that of the residuals from the model in equation (19), conditional on treatment status. Once again, we replace extreme values of `re78` with the limit point of the support, also conditional on treatment status. 'True effects' in each replication, $\tilde{\theta}^0$, are calculated using the conditional means for both treatment statuses, and the difference in conditional means, *i.e.* the individual-level treatment effect, is averaged over the subsample of treated units.[32]

[32]Thus, we implicitly focus on the sample average treatment effect on the treated (SATT), not on the population average treatment effect on the treated (ATT). Both of these measures can be used as the benchmark effect in simulations and we have no particular preference for either.

F Empirical Application: Detailed Results

Table F1: **Simulation results for Subsection 5.1**

	Absolute bias	RMSE	SD
Original samples			
Doubly-robust regression	142	1,019	1,010
IPW	51	1,102	1,101
Kernel matching	818	1,382	1,115
OLS	306	740	674
Oaxaca–Blinder	35	716	715
NN matching	16	1,209	1,209
Bias-adjusted NN matching	102	1,411	1,408
Placebo			
Doubly-robust regression	280	1,817	1,785
	(206)	(236)	(234)
IPW	323	1,960	1,928
	(161)	(225)	(221)
Kernel matching	68	1,244	1,242
	(50)	(141)	(139)
OLS	385	901	776
	(287)	(152)	(37)
Oaxaca–Blinder	420	927	783
	(311)	(171)	(40)
NN matching	241	2,423	2,407
	(172)	(326)	(318)
Bias-adjusted NN matching	319	3,531	3,509
	(392)	(11,085)	(11,086)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 1,000 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 1). The 'minimum' value for each feature, as reported in Table 1, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 16; for MSE, it is 512,322 (or $\simeq 716^2$); for variance, it is 454,278 (or $\simeq 674^2$).

Table F1: **Simulation results for Subsection 5.1 (cont.)**

	Absolute bias	RMSE	SD
Structured			
Doubly-robust regression	620	1,402	1,261
	(492)	(340)	(113)
IPW	591	1,436	1,310
	(488)	(331)	(124)
Kernel matching	408	1,458	1,426
	(371)	(254)	(145)
OLS	558	1,125	1,006
	(476)	(359)	(105)
Oaxaca–Blinder	690	1,192	997
	(495)	(389)	(103)
NN matching	626	1,660	1,533
	(492)	(311)	(122)
Bias-adjusted NN matching	620	1,634	1,509
	(491)	(312)	(119)
Bootstrap			
Doubly-robust regression	—	1,197	1,186
	—	(203)	(193)
IPW	—	1,305	1,301
	—	(235)	(231)
Kernel matching	—	1,789	1,610
	—	(307)	(189)
OLS	—	906	906
	—	(68)	(68)
Oaxaca–Blinder	—	961	961
	—	(93)	(93)
NN matching	—	1,653	1,518
	—	(325)	(250)
Bias-adjusted NN matching	—	3,126	2,980
	—	(3,560)	(3,562)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 1,000 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 1). The 'minimum' value for each feature, as reported in Table 1, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 16; for MSE, it is 512,322 (or $\simeq 716^2$); for variance, it is 454,278 (or $\simeq 674^2$).

Table F2: **Simulation results for Subsection 5.2**

	Absolute bias	RMSE	SD
Original samples			
Doubly-robust regression	1,222	1,566	980
IPW	1,081	1,514	1,060
Kernel matching	1,356	1,652	944
OLS	1,111	1,237	545
Oaxaca–Blinder	954	1,106	559
NN matching	1,122	1,732	1,320
Bias-adjusted NN matching	1,101	1,847	1,484
Placebo			
Doubly-robust regression	263	1,823	1,794
	(196)	(198)	(197)
IPW	203	2,026	2,013
	(132)	(197)	(198)
Kernel matching	78	1,487	1,483
	(88)	(267)	(263)
OLS	379	909	791
	(284)	(153)	(36)
Oaxaca–Blinder	408	930	797
	(304)	(171)	(37)
NN matching	219	2,480	2,467
	(156)	(253)	(250)
Bias-adjusted NN matching	290	3,233	3,214
	(226)	(1,852)	(1,853)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 1,000 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 2). The 'minimum' value for each feature, as reported in Table 2, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 954; for MSE, it is 1,222,627 (or $\simeq 1{,}106^2$); for variance, it is 296,671 (or $\simeq 545^2$).

Table F2: **Simulation results for Subsection 5.2 (cont.)**

	Absolute bias	RMSE	SD
Structured			
Doubly-robust regression	1,009	1,440	1,065
	(398)	(327)	(70)
IPW	1,027	1,500	1,120
	(401)	(326)	(84)
Kernel matching	827	1,411	1,156
	(405)	(301)	(86)
OLS	1,023	1,295	858
	(395)	(351)	(60)
Oaxaca–Blinder	1,072	1,327	851
	(389)	(351)	(57)
NN matching	1,041	1,704	1,364
	(403)	(303)	(84)
Bias-adjusted NN matching	1,010	1,642	1,318
	(398)	(301)	(76)
Bootstrap			
Doubly-robust regression	—	1,152	1,136
	—	(189)	(180)
IPW	—	1,262	1,258
	—	(233)	(231)
Kernel matching	—	1,452	1,352
	—	(284)	(208)
OLS	—	849	849
	—	(45)	(44)
Oaxaca–Blinder	—	853	853
	—	(43)	(43)
NN matching	—	1,789	1,615
	—	(450)	(321)
Bias-adjusted NN matching	—	3,356	3,180
	—	(1,025)	(932)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 1,000 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 1,000 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 2). The 'minimum' value for each feature, as reported in Table 2, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 954; for MSE, it is 1,222,627 (or $\simeq$1,106^2); for variance, it is 296,671 (or $\simeq$545^2).

Table F3: **Simulation results for Subsection 5.3**

	Absolute bias	RMSE	SD
Original samples			
Doubly-robust regression	68	1,573	1,572
IPW	565	1,682	1,585
Kernel matching	540	1,649	1,559
OLS	1,069	1,131	371
Oaxaca–Blinder	171	583	558
NN matching	442	2,374	2,333
Bias-adjusted NN matching	102	1,837	1,835
Placebo			
Doubly-robust regression	174	1,721	1,709
	(134)	(156)	(153)
IPW	187	2,162	2,153
	(117)	(172)	(171)
Kernel matching	144	1,925	1,917
	(140)	(281)	(277)
OLS	208	651	600
	(160)	(72)	(26)
Oaxaca–Blinder	298	753	664
	(224)	(116)	(30)
NN matching	175	2,705	2,699
	(136)	(242)	(238)
Bias-adjusted NN matching	175	1,942	1,931
	(137)	(174)	(171)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 500 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 500 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 3). The 'minimum' value for each feature, as reported in Table 3, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 68; for MSE, it is 340,300 (or $\simeq 583^2$); for variance, it is 137,574 (or $\simeq 371^2$).

Table F3: **Simulation results for Subsection 5.3 (cont.)**

	Absolute bias	RMSE	SD
Structured			
Doubly-robust regression	198	819	816
	(116)	(78)	(65)
IPW	194	1,228	1,224
	(121)	(119)	(116)
Kernel matching	149	910	913
	(116)	(107)	(106)
OLS	405	631	470
	(253)	(172)	(20)
Oaxaca–Blinder	202	509	503
	(114)	(54)	(21)
NN matching	140	1,205	1,210
	(107)	(113)	(111)
Bias-adjusted NN matching	200	942	939
	(117)	(81)	(71)
Bootstrap			
Doubly-robust regression	—	1,347	1,304
	—	(396)	(363)
IPW	—	1,177	1,157
	—	(625)	(570)
Kernel matching	—	1,276	1,187
	—	(512)	(456)
OLS	—	415	415
	—	(23)	(23)
Oaxaca–Blinder	—	620	620
	—	(33)	(33)
NN matching	—	2,245	1,974
	—	(967)	(726)
Bias-adjusted NN matching	—	1,946	1,763
	—	(555)	(435)

Notes: Results for 'Original samples' correspond to the true values of all features of interest (absolute bias, RMSE, and SD) in the original data generating process. Measures of absolute bias and RMSE are centred around the true value of ATT, reported in the main text. All calculations are based on 1,000 samples. For each of these 1,000 samples, 'Placebo' and 'Structured' generate 500 new replications using the placebo and structured approaches described in Section 2. Similarly, 'Bootstrap' generates 500 nonparametric bootstrap replications by sampling with replacement the same number of observations as the original data. In each of the three cases, we report both the mean and the standard deviation (in brackets) of EMCS estimates of all features of interest across all replications. Estimates of absolute bias and RMSE are centred around 0 for placebo and around the model-implied value for structured. Estimates of RMSE are centred around the point estimate in the original sample for bootstrap. We do not report any estimates of absolute bias in bootstrap samples, as there is no clear way to estimate bias in bootstrap. For ease of interpretation, RMSE and SD are reported instead of MSE and variance (as in Table 3). The 'minimum' value for each feature, as reported in Table 3, is its lowest value among our estimators in the original data generating process (*i.e.* the lowest value in the 'Original samples' panel). The minimum value of absolute bias is 68; for MSE, it is 340,300 (or $\simeq 583^2$); for variance, it is 137,574 (or $\simeq 371^2$).