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

ARUN ADVANI† AND TYMON SŁOCZYŃSKI‡

**Abstract**

In this paper we evaluate the premise from the recent literature on Monte Carlo studies that an empirically motivated simulation exercise is informative about the actual performance of various estimators in a particular application. We develop a theoretical framework within which this claim can be assessed. We also provide an empirical test for two leading designs of an empirical Monte Carlo study. We conclude that the internal validity of such simulation exercises is dependent on the value of the parameter of interest. This severely limits the usefulness of such procedures, since were this object known, the procedure would be unnecessary.

*JEL Classification:* C15, C21, C25, C52

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

*This version: May 5, 2017. For helpful comments, we thank Thierry Magnac (Co-Editor), four anonymous referees, Alberto Abadie, Cathy Balfe, Richard Blundell, A. 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, and seminar and conference participants at Brandeis University, Ce2 workshop, CERGE-EI, IAAE (Thessaloniki), Institute for Fiscal Studies, Michigan State University, SOLE (Arlington), Warsaw International Economic Meeting, Warsaw School of Economics, and ZEW Summer Workshop for Young Economists. We also thank Michael Lechner and Blaise Melly for providing us with copies of their codes as well as Francesco Pontiggia for helping us with the HPC cluster at Brandeis. 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. Arun Advani also acknowledges support from Programme Evaluation for Policy Analysis, a node of the National Centre for Research Methods, supported by the UK Economic and Social Research Council (Grant No: RES-576-25-0042). Tymon Słoczyński also acknowledges a START scholarship from the Foundation for Polish Science (FNP).

†University College London and Institute for Fiscal Studies.

‡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 the assumption variously referred to as exogeneity, ignorability, selection on observables, or unconfoundedness (see, *e.g.*, Blundell and Costa Dias 2009, Imbens and Wooldridge 2009). Because the number of estimators that are available to researchers in this context is very large and many of these estimators have similar asymptotic properties, Monte Carlo studies are often seen as a useful tool for examining the small-sample properties of these estimation methods.[1] Early contributions, such as Frölich (2004), focus on very stylized data-generating processes (DGPs) which do not necessarily resemble any empirical settings. This reliance on unrealistic DGPs is criticized by Huber *et al.* (2013) and Busso *et al.* (2014) who also recommend that Monte Carlo studies should intend to replicate actual datasets of interest. Such an approach to examining the small-sample properties of estimators is termed an "empirical Monte Carlo study" (EMCS) by Huber *et al.* (2013). In this sense moving away from unrealistic DGPs can be viewed as a useful improvement.

A further reading of Huber *et al.* (2013) and Busso *et al.* (2014) reveals a second recommendation for practitioners. Because the performance of estimators of average treatment effects under unconfoundedness is highly dependent on the features of the DGP, there is little clear guidance as to which estimator is most appropriate in a particular application. Consequently, Busso *et al.* (2014) recommend that "researchers estimate average treatment effects using a variety of approaches," but also that they might want to "conduct a small-scale simulation study designed to mimic their empirical context." In fact, one of the concluding statements of Busso *et al.* (2014) is that their results "suggest the wisdom of conducting a small-scale simulation study tailored to the features of the data at hand."

This recommendation mirrors one of the statements in the concluding paragraph of Huber *et al.* (2013). In particular, when discussing several possible avenues for future research, Huber *et al.* (2013) suggest that "future work may help to better understand the general external validity of the results presented in [their] paper, as even an [e]mpirical Monte Carlo study has the important limitation that it may not necessarily be valid in a different environment." At the same time, however, Huber *et al.* (2013) maintain that "the advantage [of an empirical Monte Carlo study] is that it is valid in at least one relevant environment," or, in other words, that it necessarily has a high degree of internal validity. Clearly, this usage of the terms "internal validity" and "external validity" is somewhat

[1] 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), and Frölich *et al.* (2015), all studying the finite-sample performance of estimators of average treatment effects under unconfoundedness.

nonstandard, as these terms now refer to the ability of EMCS procedures to provide evidence on the finite-sample performance of various estimators in the initial dataset of interest (internal validity) and in other empirical contexts (external validity).[2] Throughout this paper we follow Huber *et al.* (2013) in using this adapted terminology.

It must be noted, however, that Huber *et al.* (2013) do not offer any formal justification as to why it might be reasonable to expect that EMCS procedures are internally valid, essentially taking it for granted that a high degree of internal validity follows automatically from the design of their empirical Monte Carlo study. In this paper we aim at filling this gap in the recent literature. Our starting point is to develop a simple framework within which the claim of internal validity of EMCS procedures can be assessed. We show that, in general, there is little reason to expect that the performance of estimators in a simulation study is informative about the performance of estimators in the sample of interest. In principle, taking the simulation results as given, the internal validity of an EMCS is functionally related to the value of the population parameter of interest. This severely limits the usefulness of these simulation procedures, since were this parameter known, the procedure would not be necessary.

Do we therefore conclude that EMCS procedures can never be useful? Apparently, the answer to this question is quite complex. First, we develop a rule of thumb for maximizing our preferred measure of the internal validity of a simulation exercise, namely the correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations. This rule of thumb represents the optimal value of the mean benchmark effect in simulations as a function of the parameter of interest and two other parameters. Although these additional objects can be estimated, assuming a particular value for the parameter of interest would render the simulation exercise useless. Hence, our rule of thumb is more interesting from a theoretical than from a practical perspective, and it also reiterates the dependence of the internal validity of an EMCS on the population object of interest. Second, we note that our preferred measure of internal validity can easily be calculated—taking the simulation results as given—for every possible value of the parameter of interest. Consequently, if we were willing to place some bounds on the re-

[2]Following the work of Campbell and Stanley (1963) in psychology, these terms are now widely used in social sciences, including economics. Angrist and Krueger (1999) define internal validity as the question of "whether an empirical relationship has a causal interpretation in the setting where it is observed," while external validity is the question of "whether a set of internally valid estimates has predictive value for groups or values of the response variable other than those observed in a given study." At the same time, other social sciences formulate broader definitions of these terms. For example, Punch (2014) defines internal validity as referring to "the internal logic and consistency of the research." Namely, "[i]f research is seen as an argument ... then internal validity is about the logic and internal consistency of this argument." Then, external validity is simply the question of generalizability of the findings of a given study. Our usage of these terms, and that of Huber *et al.* (2013), can be seen as an adaptation of these broader definitions.

gion in which this parameter must lie—say, assume that the effect of a given treatment is nonnegative—we would sometimes be able to conclude that a given simulation exercise is at least internally valid under this condition. Yet another possibility is that bounding the effect of interest will not allow us to sign the correlation coefficient, in which case we will treat a given study as uninformative.

Another contribution of this paper is to test our theoretical predictions about the internal validity of EMCS procedures, using the data from LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), and Smith and Todd (2001, 2005). These data come from the National Supported Work (NSW) Demonstration—a U.S. job training program that operated in the 1970s and randomized treatment assignment among eligible participants—as well as from two representative samples of the U.S. population, the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID).[3] The key insight of our test is that we can compare the performance of estimators in simulations with their performance in the original data. Whilst performance in the data of interest is usually unknown, in half of our analyses we follow LaLonde (1986) in using both the experimental treatment and experimental control group to recover an unbiased estimate of the average treatment effect on employment and earnings—which, as in LaLonde (1986), becomes our "true effect" as well as our benchmark for nonexperimental estimators. In the second half of our analyses, we follow Smith and Todd (2005) in focusing on estimates using the experimental control group and one of the comparison groups; this approach has the advantage that the "true effect" is zero by construction and is not subject to sampling error. We then use these "true effects" to calculate biases (in the original data) for a number of estimators, and test how well the performance of estimators in simulations predicts their performance in the original data.

We apply this test to two alternative approaches to conducting an EMCS that are proposed in the recent literature. The first, which we term the "structured" design, is considered by Busso *et al.* (2014).[4] Loosely speaking, in this setting treatment status and covariate values are drawn from a distribution similar to that in the data, and then outcomes are generated using parameters estimated from the data. The effect of treatment can be calculated directly from the specified DGP. The second approach, which we term the "placebo" design, is proposed by Huber *et al.* (2013).[5] Here both covariates and outcome

[3]We make use of two alternative versions of the data from the NSW experiment, namely the sample used by Dehejia and Wahba (1999) and the "early random assignment" ("early RA") sample used by Smith and Todd (2005). Thus, we implicitly restrict our attention to two subsamples of men from LaLonde (1986). We also use two nonexperimental comparison groups constructed by LaLonde (1986), CPS-1 and PSID-1.

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

[5]It is also applied by Lechner and Wunsch (2013), Frölich *et al.* (2015), Huber *et al.* (2016), and Lechner and Strittmatter (2016).

are drawn jointly from the comparison data with replacement, and treatment status is assigned using parameters estimated from the full data. Since all observations come from the comparison data and the original outcomes are retained, the effect of this "placebo treatment" is always zero by construction.

We run a total of sixty-four simulation studies, half of which are placebo and half of which are structured, based on the combined NSW-CPS and NSW-PSID datasets. The results of these simulations corroborate our theoretical predictions about the internal validity of EMCS procedures. First, the average correlation between the absolute bias in the original data and the absolute mean bias in simulations is close to zero. Second, our rule of thumb has predictive power for the magnitude of this correlation; indeed, the larger is the difference between the "optimal" value and the actual value of the mean benchmark effect in simulations, the smaller is this correlation. Third, we document the existence of simulation studies in which placing sensible bounds on the parameter of interest allows us to conclude that a given EMCS might potentially be helpful in estimator choice; not surprisingly, we document that the opposite cases also exist. Our general advice to practitioners follows from these considerations. We suggest that empirical Monte Carlo studies are approached with caution; however, for a simulation study that has already been run, it might be possible to assess how likely we are to learn something useful from it. Finally, it is important that researchers continue using several different estimators as a form of a robustness check, as Busso *et al.* (2014) also suggest.

# 2 Theory

This section introduces our theoretical framework and demonstrates that the ability of empirical Monte Carlo studies to provide evidence on the finite-sample performance of various estimators hinges on the relationship between the benchmark effect in simulations and the (unknown) true effect. The notation in this section is adapted from the discussion of bootstrap methods in Horowitz (2001).

## General Framework

We begin by introducing $\{\mathbf{X}_i : i = 1, \ldots, N\}$ as generic notation for observed data, where $i$ indexes observations. $\mathbf{X}_i = (Y_i, D_i, \mathbf{Z}_i)$ is a vector, where $Y_i$ denotes the outcome variable, $D_i$ denotes the treatment variable, and $\mathbf{Z}_i$ denotes the vector of control variables. We assume $\{\mathbf{X}_i\}$ to be a set of iid random draws from an underlying distribution with cdf $F_0(\mathbf{x}) = \mathrm{P}(\mathbf{X} \leq \mathbf{x})$.

Let $\theta$ denote the population parameter of interest, which we take to be a scalar for simplicity; in this paper, we take $\theta$ to be the average treatment effect on the treated (ATT). We also consider a number of estimators (indexed by $j$) of $\theta$, namely $\{\hat{\theta}_j : j = 1, \ldots, K\}$. Each $\hat{\theta}_j$ is a function of observed data, $\hat{\theta}_j = \hat{\theta}_j(\mathbf{X}_1, \ldots, \mathbf{X}_N)$, and has an exact finite-sample distribution with cdf $G_j(t, N, F_0) = \mathrm{P}(\hat{\theta}_j \leq t)$. We can also use $G_j(\cdot, \cdot, F)$ to denote the exact cdf of $\hat{\theta}_j$ when the data are sampled from a distribution whose cdf is $F$.

This distinction between $G_j(t, N, F_0)$—which is unknown, because $F_0$ is unknown—and $G_j(\cdot, \cdot, F)$—which can be estimated for a known distribution $F$—allows us to formalize the sense in which a Monte Carlo study can be "empirical." Such a study is necessarily based on a mapping from $\{\mathbf{X}_i\}$ to $F$, say, $\psi : \{\mathbf{X}_i\} \to F_\psi(\mathbf{x})$. However, note that there exist sensible functions $F_\psi$ which do not translate into an empirical Monte Carlo study. First, the current framework contains the bootstrap as a special case; in particular, in the case of the nonparametric bootstrap, we take $F_\psi$ to be the empirical distribution function of observed data (see, *e.g.*, Horowitz 2001). However, as will be noted below, in an EMCS we also require the existence of a benchmark effect, against which all the estimates can be compared (and this is absent in the bootstrap). Second, a Monte Carlo study is termed "empirical" if $F_\psi$ is a nondegenerate function of $\{\mathbf{X}_i\}$. In other words, a Monte Carlo study will not be considered "empirical" if $F_\psi$ does not indeed depend on observed data.

Once $F_\psi$ has been determined, we can draw random samples (indexed by $s$) from this distribution and analyze the empirical distribution of various statistics of interest. In particular, each sample $\{\mathbf{X}_{i,s} : i = 1, \ldots, N_\psi\}$ yields an estimate $\hat{\theta}_{j,s} = \hat{\theta}_j(\mathbf{X}_{1,s}, \ldots, \mathbf{X}_{N_\psi,s})$. Drawing $R$ such samples, and applying the estimators to each, yields $\{\hat{\theta}_{j,s} : s = 1, \ldots, R\}$. Using $\{\hat{\theta}_{j,s}\}$, we can also estimate $G_j(t, N_\psi, F_\psi)$ by $\hat{G}_j(t, N_\psi, F_\psi) = R^{-1} \sum_{s=1}^{R} 1[\hat{\theta}_{j,s} \leq t]$.

Finally, we need to select some value, $\tilde{\theta}_s$, to be the benchmark effect in simulations, against which all the estimates will be compared. Importantly, $\tilde{\theta}_s$ is determined differently in the two approaches to conducting an EMCS that we consider in this paper. In the placebo design, $\tilde{\theta}_s = 0$ for all $s$. In the structured design, $\tilde{\theta}_s$ is determined by the parametric model for $\mathrm{E}(Y|D, \mathbf{X})$ which is also used to generate the observations on $Y$. In this latter case, $\tilde{\theta}_s$ may vary across replications, together with the observations on $D$ and $\mathbf{X}$.

## Finite-Sample Performance of Estimators

It is now possible to define several measures that can be used to evaluate the finite-sample performance of $\{\hat{\theta}_j : j = 1, \ldots, K\}$. In particular, let

$$b_{j,s} = \hat{\theta}_{j,s} - \tilde{\theta}_s \tag{1}$$

denote the bias of estimator $\hat{\theta}_j$ in sample $s$. We are more likely, however, to be interested in the mean bias of estimator $\hat{\theta}_j$, namely

$$\bar{b}_j = R^{-1} \sum_{s=1}^{R} b_{j,s} = \bar{\hat{\theta}}_j - \bar{\bar{\theta}}. \tag{2}$$

The mean estimate, $\bar{\hat{\theta}}_j$, will also turn out to be useful in what follows. Moreover, it is important to note that minimization of $\bar{b}_j$ is not helpful in evaluating the finite-sample performance of $\{\hat{\theta}_j\}$, since such an approach could simply lead to choosing estimators with "very negative" biases. Instead, in order to finally be able to discriminate between the elements of $\{\hat{\theta}_j : j = 1, \ldots, K\}$, we might focus on minimization of the absolute mean bias of estimator $\hat{\theta}_j$, namely

$$\left|\bar{b}_j\right| = \left|\bar{\hat{\theta}}_j - \bar{\bar{\theta}}\right|. \tag{3}$$

When we focus on $\left|\bar{b}_j\right|$, it is clear that we should prefer estimators with small values of this measure. Of course, other measures of finite-sample performance of estimators also exist, such as (absolute) median bias and mean squared error. Although we focus on absolute mean bias for conciseness, our theoretical discussion could be extended to absolute median bias. On the other hand, the case of mean squared error is more difficult. It is well known, however, that the mean squared error is the sum of the variance and the squared mean bias (or, equivalently, squared absolute mean bias). Since our theoretical predictions on the internal validity of EMCS procedures with respect to absolute mean bias will be negative, the predictions with respect to mean squared error could only be positive if EMCS procedures were able to predict estimator variance very well, and this effect was able to completely offset the negative result on absolute mean bias.[6] Thus, we believe that our results on absolute mean bias have very general implications in the context of internal validity of empirical Monte Carlo studies.

## Internal Validity of Empirical Monte Carlo Studies

Applied researchers may be tempted to choose estimators that have small values of absolute mean bias in a particular simulation study. In what follows, we will discuss a possible approach to evaluating the internal validity of empirical Monte Carlo studies,

[6]Regardless of whether this is plausible, it is unclear why one would use an empirical Monte Carlo study—instead of the bootstrap—as a data-driven method to study estimator variance in a particular setting. In principle, we might prefer an EMCS over the bootstrap in order to control several features of the DGP, thereby increasing the number of relevant settings. But in this paper we focus solely on internal validity of EMCS procedures, not on their external validity.

thereby assessing also the appropriateness of such practice.

Suppose we have an initial set of observed data, $\{\mathbf{X}_i^* : i = 1, \ldots, N^*\}$. Using this dataset, we can obtain $K$ estimates of $\theta$, namely $\left\{\hat{\theta}_j^* : j = 1, \ldots, K\right\}$. Which of these values should we trust? Let

$$b_j^* = \hat{\theta}_j^* - \theta \tag{4}$$

denote the (true) bias of estimator $\hat{\theta}_j$ in the initial dataset. As before, we are more likely to be interested in the absolute bias of $\hat{\theta}_j$, defined as

$$\left|b_j^*\right| = \left|\hat{\theta}_j^* - \theta\right|. \tag{5}$$

Clearly, we would like to choose $\hat{\theta}_j$ such that $\left|b_j^*\right|$ is as small as possible. In practice, of course, we cannot calculate the absolute bias of $\hat{\theta}_j$ because we do not know $\theta$. If we knew $\theta$, we would not need *any* estimators to estimate it. In this situation, we might think that it is at least possible to predict the relative magnitudes of $\left\{\left|b_j^*\right| : j = 1, \ldots, K\right\}$ using an empirical Monte Carlo study. This would amount to applying $\psi$ to $\{\mathbf{X}_i^*\}$, obtaining $F_\psi^*$, and drawing a large number of samples from this distribution.

Recall that our general definition of an internally valid EMCS is that the performance of estimators in this study is informative about the performance of estimators in the initial dataset of interest. To operationalize this definition, given our focus on absolute mean bias, we will say that an empirical Monte Carlo study is *internally valid* if a given $\psi$ ensures that $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|) > 0$. Also, the higher this correlation the better is an EMCS, as larger values of the correlation coefficient translate into better predictive power of $\left|\bar{b}_j\right|$ for $\left|b_j^*\right|$. Intuitively, $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|) > 0$ corresponds to being more likely to get what we need—an estimator with a small value of $\left|b_j^*\right|$, that is, absolute (true) bias—whenever we choose an estimator with small absolute mean bias (in simulations), $\left|\bar{b}_j\right|$. In what follows, we will demonstrate that this is not a realistic expectation, regardless of $\psi$, that is, regardless of whether we use the placebo design, the structured design, or any other approach to conducting an EMCS. (In fact, our results apply also to stylized DGPs, as in Frölich 2004.)

## Relationship between $\left|\bar{b}_j\right|$ and $\left|b_j^*\right|$

We begin by defining several additional objects of interest. First, write the linear projection of $\bar{\hat{\theta}}_j$ onto $\hat{\theta}_j^*$ as

$$\bar{\hat{\theta}}_j = \alpha_\theta + \beta_\theta \hat{\theta}_j^* + v_{j\theta}, \tag{6}$$

where $\beta_\theta = \text{Cov}(\bar{\hat{\theta}}_j, \hat{\theta}_j^*)/\text{V}(\hat{\theta}_j^*)$. Next, the linear projection of $\bar{b}_j$ onto $b_j^*$ can be written as

$$\bar{b}_j = \alpha_b + \beta_b b_j^* + v_{jb}, \tag{7}$$

where $\beta_b = \text{Cov}(\bar{b}_j, b_j^*)/\text{V}(b_j^*) = \text{Cov}(\bar{\hat{\theta}}_j - \bar{\bar{\theta}}, \hat{\theta}_j^* - \theta)/\text{V}(\hat{\theta}_j^* - \theta) = \text{Cov}(\bar{\hat{\theta}}_j, \hat{\theta}_j^*)/\text{V}(\hat{\theta}_j^*) = \beta_\theta$; of course, even though $\beta_b = \beta_\theta$, it is not necessarily the case that $\alpha_b$ is equal to $\alpha_\theta$. Finally, write the linear projection of $\left|\bar{b}_j\right|$ onto $\left|b_j^*\right|$ as

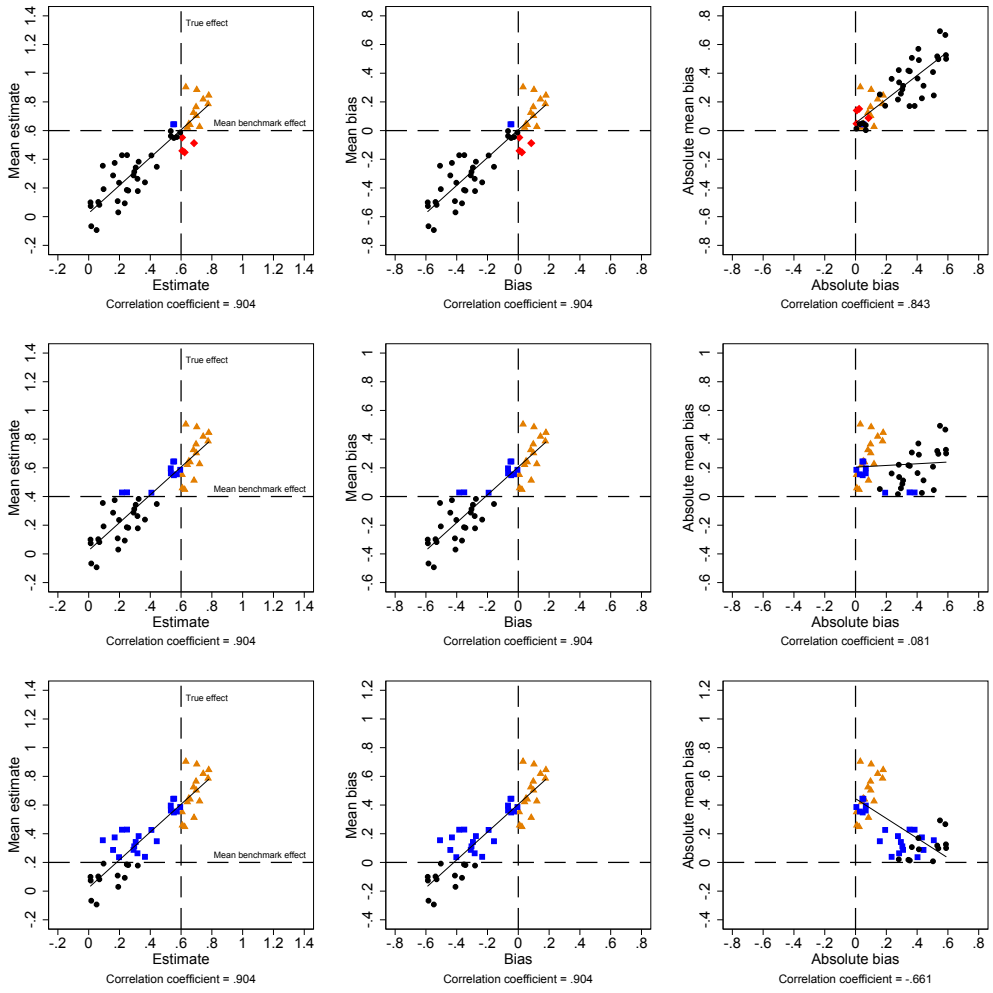
$$\left|\bar{b}_j\right| = \alpha_a + \beta_a \left|b_j^*\right| + v_{ja}, \tag{8}$$

where $\beta_a = \text{Cov}(\left|\bar{b}_j\right|, \left|b_j^*\right|)/\text{V}(\left|b_j^*\right|) = \text{Cov}(\left|\bar{\hat{\theta}}_j - \bar{\bar{\theta}}\right|, \left|\hat{\theta}_j^* - \theta\right|)/\text{V}(\left|\hat{\theta}_j^* - \theta\right|)$. The relationship between $\beta_a$ and $\beta_b$, or $\text{Cov}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ and $\text{Cov}(\bar{b}_j, b_j^*)$, is therefore unclear; in particular, $\beta_a$ depends (unlike $\beta_b = \beta_\theta$) on the exact values of $\bar{\bar{\theta}}$ and $\theta$. This dependence is problematic, as it implies that the ability of an empirical Monte Carlo study to help us find a "good" estimator of $\theta$ depends on the (unknown) value of $\theta$.

Figure 1, which contains three rows and three columns of scatter plots, provides an illustration of this problem using a stylized dataset. This is a purely hypothetical example whose role is to effectively illustrate our argument. Point estimates ($x$ axis), which represent the initial dataset of interest, were drawn from a mixture of two normal distributions, $\mathcal{N}[.2, .1]$ and $\mathcal{N}[.6, .1]$. Mean estimates in simulations ($y$ axis) were generated as the sum of each point estimate and iid random noise with distribution $\mathcal{N}[0, .1]$. Clearly, the assumption that the estimates are iid is unrealistic, but this is just an illustration.

In Figure 1 the first column always depicts raw estimates, that is, the relationship between $\bar{\hat{\theta}}_j$ and $\hat{\theta}_j^*$. Next, the second column always depicts the relationship between mean biases and (true) biases, or $\bar{b}_j$ and $b_j^*$. Finally, the third column always depicts the relationship between the absolute values of these measures, or $\left|\bar{b}_j\right|$ and $\left|b_j^*\right|$. What is particularly important, each row presents the *same* dataset, with the *same* value of $\theta$ (equal to 0.6). The *only* difference between each pair of rows is that $\bar{\bar{\theta}}$, the mean benchmark effect in simulations, changes from 0.6 in the first row to 0.4 in the second row to 0.2 in the third row. Thus, the first row represents a situation in which the mean benchmark effect, $\bar{\bar{\theta}}$, is equal to the true effect, $\theta$. This leads to a strong (positive) correlation between the absolute bias and the absolute mean bias, $\left|b_j^*\right|$ and $\left|\bar{b}_j\right|$. If we were to choose an estimator on the basis of this Monte Carlo study, we would be likely to make a good decision. However, when instead we observe a mild difference between $\bar{\bar{\theta}}$ and $\theta$ in the second row ($\bar{\bar{\theta}} = 0.4 \neq 0.6 = \theta$), this positive correlation disappears. When the difference between $\theta$

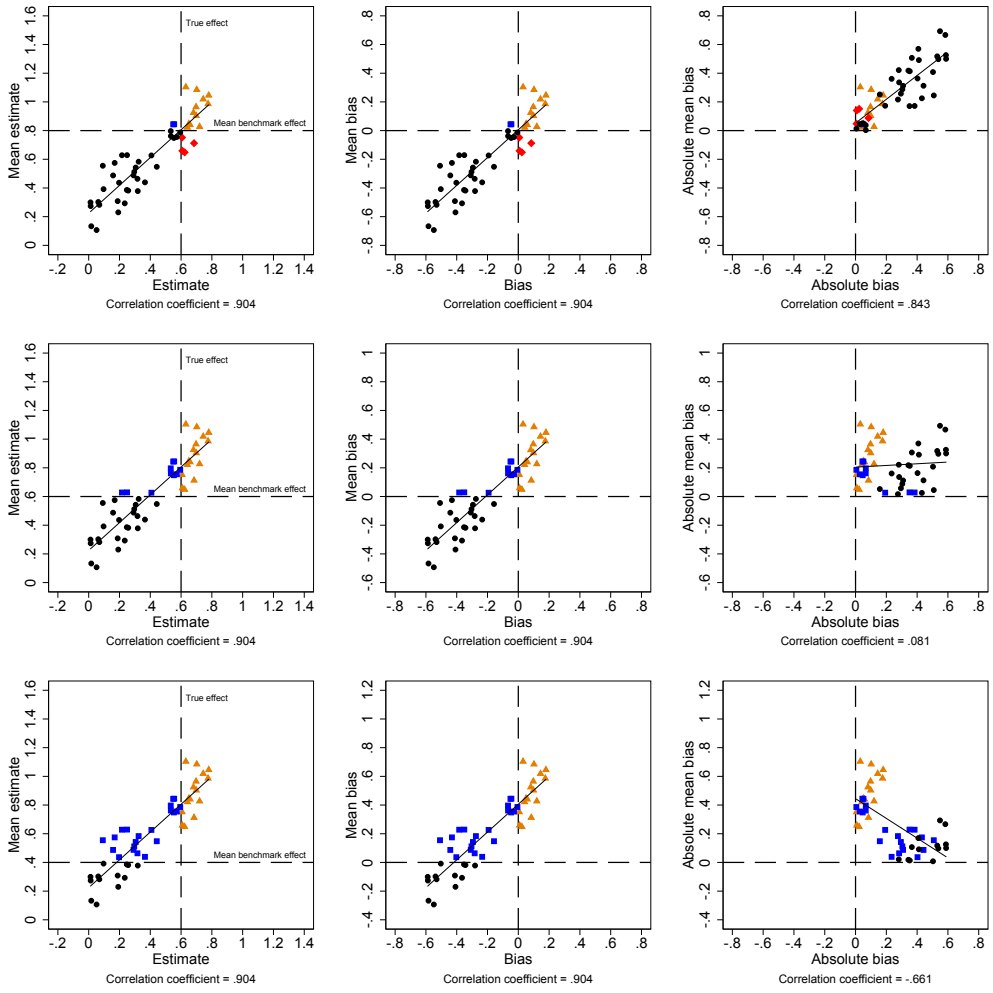
Figure 1: Internal Validity of an EMCS in a Stylized Dataset ($\alpha_\theta \simeq 0$ and $\beta_\theta \simeq 1$)




*Note:* The first column depicts the relationship between $\bar{\hat{\theta}}_j$ and $\hat{\theta}_j^*$. The second column depicts the relationship between $\bar{b}_j$ and $b_j^*$. The third column depicts the relationship between the absolute values of $\bar{b}_j$ and $b_j^*$. Each row presents the same dataset, with the same value of $\theta$ (equal to 0.6). The only difference between each pair of rows is that $\bar{\bar{\theta}}$ changes from 0.6 in the first row to 0.4 in the second row to 0.2 in the third row. Orange triangles (blue squares/black circles/red diamonds) are used for data points that are located in the first (second/third/fourth) quadrant of the second-column plots. Each data point is depicted using the same symbol in all plots.

and $\bar{\bar{\theta}}$ increases from 0.2 to 0.4 in the third row, the correlation again becomes strong, but *negative*. Unlike in the first row, we would now be likely to make a good decision if we were to choose an estimator that performs *badly* in this Monte Carlo study. In practice, however, we would never know which situation applies, unless we actually knew the value of $\theta$. If this parameter were to be known, however, no Monte Carlo study would be necessary—and, in fact, no estimator of $\theta$ would be necessary either.

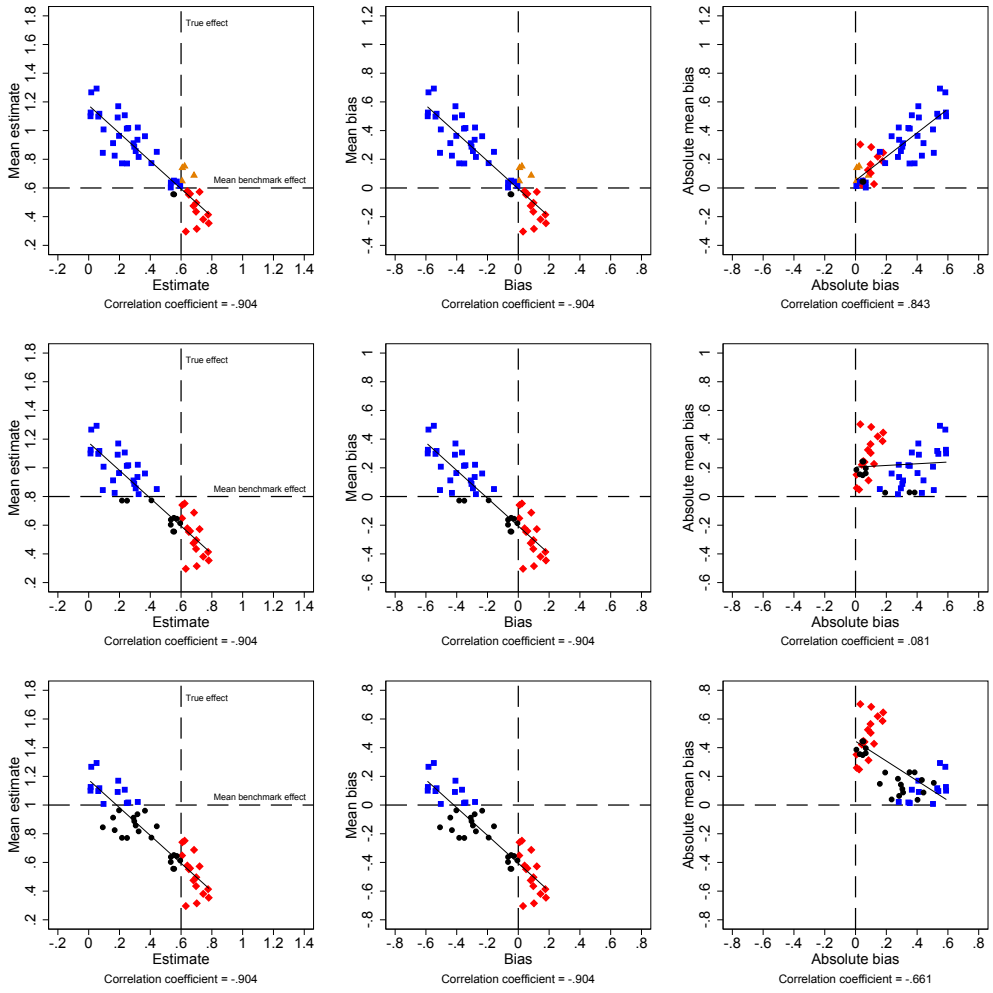
Figure 2: Internal Validity of an EMCS in a Stylized Dataset ($\alpha_\theta \simeq 0.2$ and $\beta_\theta \simeq 1$)




*Note:* The first column depicts the relationship between $\bar{\hat{\theta}}_j$ and $\hat{\theta}_j^*$. The second column depicts the relationship between $\bar{b}_j$ and $b_j^*$. The third column depicts the relationship between the absolute values of $\bar{b}_j$ and $b_j^*$. Each row presents the same dataset, with the same value of $\theta$ (equal to 0.6). The only difference between each pair of rows is that $\bar{\tilde{\theta}}$ changes from 0.8 in the first row to 0.6 in the second row to 0.4 in the third row. Orange triangles (blue squares/black circles/red diamonds) are used for data points that are located in the first (second/third/fourth) quadrant of the second-column plots. Each data point is depicted using the same symbol in all plots.

A potentially problematic aspect of Figure 1 is that it is very stylized. In particular, $\alpha_\theta \simeq 0$ and $\beta_\theta = \beta_b \simeq 1$. This implies that $\mathrm{E}(\bar{\hat{\theta}}_j|\hat{\theta}_j^*) = \hat{\theta}_j^*$; in other words, for each estimator $\hat{\theta}_j$, we expect the mean estimate in simulations to be equal to the estimate in the initial dataset. While this might be a good approximation for some empirical Monte Carlo studies, it is not likely to be true in general. A slightly different example is presented in Figure 2. In fact, it depicts the same dataset as in Figure 1, with each mean estimate—as

Figure 3: Internal Validity of an EMCS in a Stylized Dataset ($\alpha_\theta \simeq 1.2$ and $\beta_\theta \simeq -1$)




*Note:* The first column depicts the relationship between $\bar{\hat{\theta}}_j$ and $\hat{\theta}_j^*$. The second column depicts the relationship between $\bar{b}_j$ and $b_j^*$. The third column depicts the relationship between the absolute values of $\bar{b}_j$ and $b_j^*$. Each row presents the same dataset, with the same value of $\theta$ (equal to 0.6). The only difference between each pair of rows is that $\bar{\bar{\theta}}$ changes from 0.6 in the first row to 0.8 in the second row to 1.0 in the third row. Orange triangles (blue squares/black circles/red diamonds) are used for data points that are located in the first (second/third/fourth) quadrant of the second-column plots. Each data point is depicted using the same symbol in all plots.

well as the mean benchmark effect—now increased by 0.2. Consequently, $\alpha_\theta \simeq 0.2$ and $\beta_\theta = \beta_b \simeq 1$. The second row of Figure 2 demonstrates that the equality between the mean benchmark effect in simulations, $\bar{\bar{\theta}}$, and the true effect, $\theta$, does not guarantee that the correlation between the absolute bias and the absolute mean bias, $\left|b_j^*\right|$ and $\left|\bar{b}_j\right|$, will be strong and positive. Clearly, in this example, $\bar{\bar{\theta}} = \theta = 0.6$, but the correlation between the absolute bias and the absolute mean bias is nevertheless close to zero. At the same

time, we observe a strong and positive correlation between $\left|b_j^*\right|$ and $\left|\bar{b}_j\right|$ in the first row of Figure 2. In this example, there is a mild difference between the mean benchmark effect, $\bar{\bar{\theta}} = 0.8$, and the true effect, $\theta = 0.6$, but this difference is offset by $\alpha_\theta \simeq 0.2$.

We might also wonder what the consequences of a negative relationship between the mean estimate in simulations and the estimate in the initial dataset, $\text{Cov}(\bar{\hat{\theta}}_j, \hat{\theta}_j^*) < 0$, might be. Another stylized example, based on a transformation of the previous data, is therefore presented in Figure 3. Once again we observe that, dependent on the relationship between the mean benchmark effect in simulations and the true effect, it is entirely possible that the correlation between the absolute bias and the absolute mean bias can be strong and positive, close to zero, or strong and negative—even though the underlying estimates once again do *not* change between rows. An example with a strong and positive correlation between $\left|b_j^*\right|$ and $\left|\bar{b}_j\right|$ is presented, again, in the first row of Figure 3.

What do the first rows of Figures 1–3 have in common? In other words, what condition is required to enable the correlation between the absolute bias and the absolute mean bias to be as strong and positive as possible? We might notice that the condition that is shared by the first rows of Figures 1–3 is actually very simple. The linear projection of $\bar{\hat{\theta}}_j$ onto $\hat{\theta}_j^*$ always passes through the point $(\theta, \bar{\bar{\theta}})$; equivalently, the linear projection of $\bar{b}_j$ onto $b_j^*$ always passes through the origin.

We can now examine the consequences of this observation. In fact, we can rewrite equation (7) as

$$\begin{aligned} \bar{b}_j &= \alpha_b + \beta_b b_j^* + v_{jb} && (9) \\ \bar{\hat{\theta}}_j - \bar{\bar{\theta}} &= \alpha_b + \beta_\theta \left(\hat{\theta}_j^* - \theta\right) + v_{jb} && (10) \\ \bar{\hat{\theta}}_j &= \left(\alpha_b + \bar{\bar{\theta}} - \beta_\theta \theta\right) + \beta_\theta \hat{\theta}_j^* + v_{jb}. && (11) \end{aligned}$$

But equation (11) represents the linear projection of $\bar{\hat{\theta}}_j$ onto $\hat{\theta}_j^*$, which was previously defined in equation (6). What follows, $\alpha_\theta = \alpha_b + \bar{\bar{\theta}} - \beta_\theta \theta$. Moreover, if the linear projection of $\bar{b}_j$ onto $b_j^*$ were to pass through the origin, we would also require that $\alpha_b = 0$, which in turn implies that $\alpha_\theta = \bar{\bar{\theta}} - \beta_\theta \theta$. This restriction leads to a rule of thumb for maximizing the correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations, $\left|b_j^*\right|$ and $\left|\bar{b}_j\right|$. If we use $\bar{\bar{\theta}}_{opt}$ to denote the correlation-maximizing ("optimal") value of the mean benchmark effect, this rule of thumb can be written as

$$\bar{\bar{\theta}}_{opt} = \alpha_\theta + \beta_\theta \theta. \quad (12)$$

Equation (12) demonstrates that any claim of internal validity of empirical Monte Carlo studies is necessarily based on circular reasoning. First, one might suggest that we can learn about the value of $\theta$ because an empirical Monte Carlo study has helped us choose "good" estimators of this parameter. Then, however, our ability to discriminate between "good" and "bad" estimators depends on $\theta$, and is maximized if $\bar{\tilde{\theta}} = \bar{\tilde{\theta}}_{opt}$.

## Implications

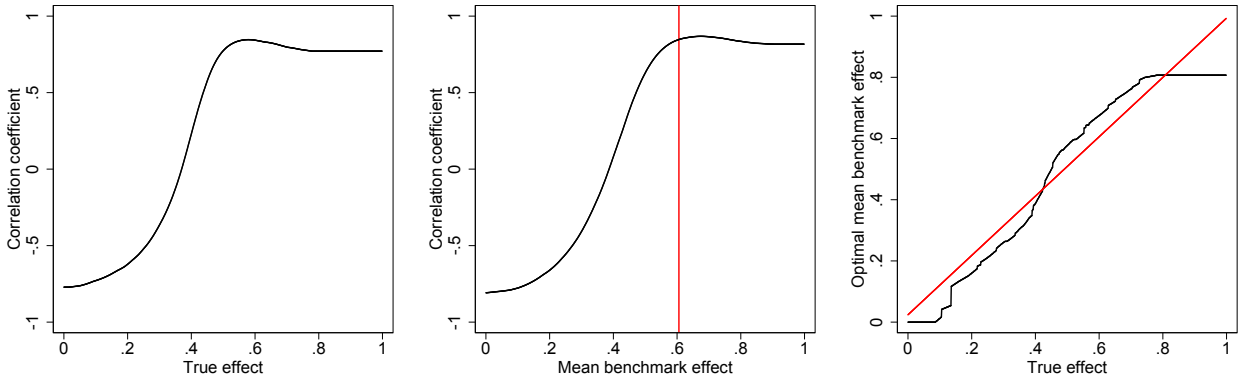
The discussion thus far has important implications for the internal validity of empirical Monte Carlo studies, some of which have already been mentioned. Here we provide a more detailed discussion as well as an illustration which, again, uses the previously introduced stylized dataset (more precisely, its version from Figure 1).

First, it is clear that the internal validity of an empirical Monte Carlo study is functionally related to the value of the parameter of interest, $\theta$. This severely limits the usefulness of EMCS procedures, since the whole point of conducting such a simulation study is to find "good" estimators of $\theta$. This conclusion is not related to the fact that Figures 1–3 present an oversimplified example of a possible simulation study. It is clear that our target measure, $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|) = \text{Cor}(\left|\bar{\tilde{\theta}}_j - \bar{\tilde{\theta}}\right|, \left|\hat{\theta}_j^* - \theta\right|)$, does in general depend on $\theta$.[7] In the case of our stylized dataset, the dependence of $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ on $\theta$, holding $\bar{\tilde{\theta}}$ fixed at 0.6, is presented in the left-hand segment of Figure 4. The correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations fluctuates between –0.771 and 0.846, and the *only* source of variation is the value of $\theta$. In practice, if we decided to conduct an empirical Monte Carlo study, we would never know which situation applies because we would not know the value of $\theta$. Thus, we should not, in general, expect that EMCS procedures can provide reliable information about the performance of estimators in the sample of interest.

Second, equation (12) suggests that $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ will be maximized if $\bar{\tilde{\theta}} = \alpha_\theta + \beta_\theta \theta$. It is important to see that the accuracy of this rule of thumb does, in general, depend on whether the linear projection of $\bar{\tilde{\theta}}_j$ onto $\hat{\theta}_j^*$ provides a good fit to the data. If the data are highly dispersed or the relationship between $\bar{\tilde{\theta}}_j$ and $\hat{\theta}_j^*$ is nonlinear, this rule of thumb will be imperfect. In the case of our stylized dataset, the accuracy of equation (12) in predicting the maximum of $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ can be seen visually in the central segment of Figure 4, which displays the dependence of $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ on $\bar{\tilde{\theta}}$, holding $\theta$ fixed at 0.6.

[7] It should be straightforward to realize that, for any other set of simulation results and for fixed $\bar{\tilde{\theta}}$, manipulating $\theta$ (the vertical dashed line in the first column of Figures 1–3) would induce variation in the correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations.

Figure 4: The Impact of $\theta$ and $\bar{\tilde{\theta}}$ on the Internal Validity of an EMCS




*Note:* Data come from the stylized dataset introduced in Figure 1. The left-hand segment presents the dependence of $\text{Cor}(|\bar{b}_j|, |b_j^*|)$ on $\theta$, holding $\bar{\tilde{\theta}}$ fixed at 0.6. The central segment presents the dependence of $\text{Cor}(|\bar{b}_j|, |b_j^*|)$ on $\bar{\tilde{\theta}}$, holding $\theta$ fixed at 0.6. The vertical (red) line presents the value of $\bar{\tilde{\theta}}_{opt}$ implied by equation (12). The right-hand segment presents the dependence of $\bar{\tilde{\theta}}_{opt}$ on $\theta$, using two formulas for $\bar{\tilde{\theta}}_{opt}$; the straight (red) line uses equation (12) and the curved (black) line uses equation (13).

Clearly, the value of $\bar{\tilde{\theta}}$ that is suggested by the rule of thumb (represented by the vertical red line) does a reasonably good job of predicting the maximum of $\text{Cor}(|\bar{b}_j|, |b_j^*|)$.

It is important to note that equation (12) can be thought of as a linear approximation to the true value of $\bar{\tilde{\theta}}_{opt}$, where

$$\bar{\tilde{\theta}}_{opt} = \arg\max_{\bar{\tilde{\theta}}} \text{Cor}(|\bar{b}_j|, |b_j^*|) = \arg\max_{\bar{\tilde{\theta}}} \text{Cor}(|\bar{\hat{\theta}}_j - \bar{\tilde{\theta}}|, |\hat{\theta}_j^* - \theta|). \tag{13}$$

In the case of our stylized dataset, the fact that equation (12) is indeed a linear approximation to equation (13) can be seen in the right-hand segment of Figure 4. For each value of $\theta$, we compare two "optimal" values of $\bar{\tilde{\theta}}$, namely the value that is suggested by the rule of thumb (straight red line) and the actual correlation-maximizing value (curved black line), which we can find numerically. Clearly, the interpretation of our rule of thumb as a linear approximation to equation (13) is quite accurate in this case.

Third, even though our results on the internal validity of EMCS procedures are generally quite negative, it is not impossible to formulate a potentially more optimistic implication of our findings. Namely, it can sometimes be helpful to place some bounds on the region in which our population parameter of interest is assumed to lie. Since our target measure, $\text{Cor}(|\bar{b}_j|, |b_j^*|)$, can easily be calculated for every possible value of $\theta$, we can also determine the bounds on $\theta$ which guarantee that this correlation coefficient will be positive (or negative). If these bounds are sensible, an empirical Monte Carlo study

might be genuinely helpful. In the case of our stylized dataset, this information can be gathered from the left-hand segment of Figure 4. For example, if we were convinced that $\theta \geq 0.370$, we would know that our correlation of interest must be nonnegative (which might not be informative enough); if we were willing to increase this lower bound, say, to 0.438, we would know that $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|) \geq 0.5$, and then this simulation exercise could certainly be useful. On the other hand, these particular bounds might have questionable practical value, given that many point estimates in our hypothetical example are smaller than the proposed lower bounds on $\theta$.

# 3 Data

This section discusses the data that we use as the basis for our empirical Monte Carlo studies; the role of these simulations is to provide a test for our theoretical predictions. As noted in Section 1, we focus on the data on men from LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), and Smith and Todd (2001, 2005). A subset of these data comes from the National Supported Work (NSW) Demonstration, which was a work experience program that operated in the mid-1970s at 15 locations in the United States (for a detailed description of the program see Smith and Todd, 2005). This program 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 programs, the NSW implemented random assignment among eligible participants. This random selection allowed for straightforward evaluation of the program via a comparison of mean outcomes in the treatment and control groups.

In an influential paper, LaLonde (1986) uses the design of this program 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 and PSID-1.

In this paper we take the key insight of LaLonde (1986) one step further. We note that if we treat the experimental estimate of the impact of NSW as the "true effect," we can calculate "true biases" of various estimators in the original nonexperimental datasets. We can also conduct an empirical Monte Carlo study in an attempt to replicate these data, and compare the performance of estimators in simulations with their performance in the original data. It must be noted, however, that in this scenario the "true effect" is estimated and, therefore, is subject to sampling error. Thus, we also use an insight of Smith and Todd (2005) that if we discard the original treatment group from the NSW data and study the control group and the nonexperimental comparison groups, the "true effect" in these data will be zero by construction and will not be subject to sampling error. This approach has a clear advantage if we are concerned about the uncertainty in the "true effect."

Moreover, we use two alternative versions of the data from the NSW experiment, namely the sample used by Dehejia and Wahba (1999), henceforth DW, and the "early random assigment" ("early RA") sample used by Smith and Todd (2005), henceforth ST. What follows, we construct a total of eight datasets, namely "DW control/CPS," "DW treated/CPS," "ST control/CPS," "ST treated/CPS" (jointly referred to as "NSW-CPS"), "DW control/PSID," "DW treated/PSID," "ST control/PSID," and "ST treated/PSID" (jointly referred to as "NSW-PSID"). For each of these datasets, we conduct eight empirical Monte Carlo studies, where we vary three aspects of a study: the outcome variable (earnings or nonemployment), the set of control variables ("simple" or "balanced"), and the design of the study itself ("placebo" or "structured").[8] Descriptive statistics as well as "true effects" for our "original datasets" are presented in the Appendix.

# 4 Designs

This section provides a more detailed discussion of our application of both approaches to conducting an EMCS, namely the structured design of Busso *et al.* (2014) and the placebo design of Huber *et al.* (2013).

[8]Both outcome variables are measured in 1978. The "simple" set of control variables is taken from Abadie and Imbens (2011) and includes age (`age`), education (`educ`), earnings in months 13–24 prior to randomization (`re74`), earnings in 1975 (`re75`), as well as indicators for whether black (`black`), whether married (`married`), whether had zero earnings in months 13–24 prior to randomization (`u74`), and whether had zero earnings in 1975 (`u75`). The "balanced" set of control variables is borrowed from Dehejia and Wahba (2002). For CPS, it includes all the variables in the "simple" specification as well as age squared (`age2`), age cubed (`age3`), education squared (`educ2`), an indicator for whether a high school dropout (`nodegree`), an indicator for whether Hispanic (`hispanic`), and an interaction between `re74` and `educ` (`re74ed`). For PSID, it includes all the variables in the "simple" specification as well as `age2`, `educ2`, `nodegree`, `hispanic`, `re74` squared (`re742`), `re75` squared (`re752`), and an interaction between `u74` and `hispanic` (`u74h`).

## The Structured Design

In the structured design we begin by generating a fixed number of treated and nontreated observations in each replication, so that the original sample sizes and proportions of both groups are retained. We then draw an employment status pair of `u74` and `u75`, 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, 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 original data, each covariate from `nodegree` onward is regressed on all the covariates listed before it (we use the logit model for binary variables).[9] 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 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).

[9] 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.

The simulated outcome, $Y_{i,s}$, is then generated in two steps. In the first step, we generate a conditional mean using the parameters of a flexible logit model (for `u78`) or a flexible linear model (for `re78`) fitted from the original data. Precisely, we estimate either $(\gamma_0, \gamma_1)$ from the following logit model:

$$\mathrm{P}(Y_i^* = 1 | D_i^*, \mathbf{Z}_i^*) = \Lambda((1 - D_i^*)\mathbf{Z}_i^*\gamma_0 + D_i^*\mathbf{Z}_i^*\gamma_1), \tag{14}$$

or $(\delta_0, \delta_1)$ from the following linear model:

$$\mathrm{E}(Y_i^* | D_i^*, \mathbf{Z}_i^*) = (1 - D_i^*)\mathbf{Z}_i^*\delta_0 + D_i^*\mathbf{Z}_i^*\delta_1. \tag{15}$$

Importantly, $\mathbf{Z}_i^*$ contains all the control variables that are included in a given specification, "simple" or "balanced." The predicted conditional mean in the simulated data is then calculated using the estimated coefficients $(\hat{\gamma}_0, \hat{\gamma}_1)$ or $(\hat{\delta}_0, \hat{\delta}_1)$, and the simulated treatment status and covariates, $D_{i,s}$ and $\mathbf{Z}_{i,s}$. In the second step, the simulated outcome, $Y_{i,s}$, is determined either—in the case of nonemployment—as a draw from a Bernoulli distribution with the estimated conditional probability $\Lambda((1 - D_{i,s})\mathbf{Z}_{i,s}\hat{\gamma}_0 + D_{i,s}\mathbf{Z}_{i,s}\hat{\gamma}_1)$ or—in the case of earnings—as a draw from a normal distribution with the estimated conditional mean $(1 - D_{i,s})\mathbf{Z}_{i,s}\hat{\delta}_0 + D_{i,s}\mathbf{Z}_{i,s}\hat{\delta}_1$ and the variance that is fitted to that of the residuals from the model in equation (15), 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, $\tilde{\theta}_s$, 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.[10]

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

## The Placebo Design

In the placebo design covariates are drawn jointly with outcomes from the empirical distribution, rather than a parameterized approximation. In particular, pairs $(Y_{i,s}, \mathbf{Z}_{i,s})$ are drawn with replacement from the sample of comparison observations from CPS or PSID.

[10]Thus, we implicitly focus on the sample average treatment effect on the treated (SATT), not on the population average treatment effect on the treated (PATT). Both of these measures can be used as the benchmark effect in simulations and we have no particular preference for either. Our theoretical predictions from Section 2 also support using either of these parameters.

The data from the NSW experiment—both the treatment and control groups, as in Smith and Todd (2005)—are used with the comparison data to estimate the propensity scores using the logit model. The vector of coefficients from this model is referred to as $\boldsymbol{\phi}$, and in each case we estimate the propensity scores using all the control variables that are included in a given specification, "simple" or "balanced." The inclusion of the "balanced" specification might be particularly important, given that Huber *et al.* (2013) stress the importance of the correct specification of the model for the propensity scores.

We then assign treatment status to observations in the simulated data using the estimated vector, $\hat{\boldsymbol{\phi}}$; iid logistic errors, $\epsilon_{i,s}$; and two scalar parameters, $\lambda$ and $\pi$, where $\lambda$ determines the degree of covariate overlap between the "placebo treated" and "nontreated" observations and $\pi$ determines the proportion of the "placebo treated." More precisely,

$$D_{i,s} = 1[S_{i,s} > 0], \tag{16}$$

$$S_{i,s} = \pi + \lambda \mathbf{Z}_{i,s}\hat{\boldsymbol{\phi}} + \epsilon_{i,s}. \tag{17}$$

Since the outcome variable, $Y_{i,s}$, is drawn directly from the data together with $\mathbf{Z}_{i,s}$, we do not need to specify any DGP for the outcome. Instead we know that the effect of the "placebo treatment" is zero, and that it is also constant across samples, $\tilde{\theta}_s = 0$ for all $s$.[11]

This design requires some choice of $\pi$ and $\lambda$. We choose $\pi$ to ensure that the proportion of the "placebo treated" observations in each simulated sample is as close as possible to the proportion of treated units in the corresponding original dataset.[12] We also follow Huber *et al.* (2013) in choosing $\lambda = 1$. As before, we generate 2,000 samples in each EMCS based on NSW-PSID and 500 samples in each EMCS based on NSW-CPS.

# 5 Estimators

This section provides an overview of estimators which we use in our EMCS procedures. All of these estimators are based on the assumption of unconfoundedness which, loosely speaking, requires that there is no omitted variables bias. Unfortunately, this assumption is not uncontroversial in the context of the NSW-CPS and NSW-PSID datasets. Follow-

[11] 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.* (2013).

[12] 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.

ing LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), Abadie and Imbens (2011), and Diamond and Sekhon (2013), we proceed, however, as if this assumption were satisfied.

We begin by discussing estimators which we use to study the impact of NSW on nonemployment. We consider estimators which belong to one of five main classes: standard parametric (regression-based), flexible parametric (Oaxaca–Blinder), kernel-based (kernel matching, local linear regression, and local logit), nearest-neighbor (NN) matching, and inverse probability weighting (IPW) estimators. 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).

In particular, we use as regression-based methods the linear probability model (LPM) as well as the logit, probit, and complementary log-log models. The complementary log-log model uses an asymmetric link function, which makes it more appropriate when the probability of success takes values close to zero or one, as is the case in our application.

We also follow Kline (2011) in using the Oaxaca–Blinder (OB) decomposition to estimate the ATT.[13] Since we consider a binary outcome, we use both linear and nonlinear OB methods. The linear OB decomposition is equivalent to the LPM but with the treatment indicator interacted with appropriately demeaned covariates. Similarly, nonlinear OB decompositions impose either a logit or a probit link function around the linear index, separately for both groups of interest (see, *e.g.*, Yun 2004, Fairlie 2005).

Turning to more standard treatment effect estimators, we consider several kernel-based methods, in particular kernel matching, local linear regression, and local logit. Kernel matching estimators play a prominent role in the program evaluation literature (see, *e.g.*, Heckman *et al.* 1997, Frölich 2004), and their asymptotic properties are established by Heckman *et al.* (1998). Similarly, local linear regression is studied by Fan (1992) and Heckman *et al.* (1998). Because our outcome is binary, we also consider local logit, as applied in Frölich and Melly (2010). Note that each of these estimators requires estimating the propensity score in the first step (based on the logit model) as well as choosing a bandwidth. For each of the methods, we select the bandwidth on the basis of leave-one-out cross-validation (as in Busso *et al.* 2009 and Huber *et al.* 2013) from a search grid $.005 \times 1.25^{g-1}$ for $g = 1, 2, \ldots, 15$, and repeat this process in each replication.[14]

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

[14] Note that the computation time is already quite large in the case of the NSW-PSID datasets, but it is completely prohibitive for NSW-CPS. Consequently, in the case of the NSW-CPS datasets, we calculate optimal bandwidths only once, for the original dataset, and use these values in our simulations.

We also apply several NN matching estimators, including both matching on covariates and on the estimated propensity score. Asymptotic properties for some of these estimators are derived by Abadie and Imbens (2006). Since these matching estimators are shown not to be $\sqrt{N}$-consistent in general, we also consider the bias-adjusted variant of both versions of matching (Abadie and Imbens, 2011). Like kernel-based methods, NN matching estimators require choosing a tuning parameter, $M$, the number of neighbors. We consider the workhorse case of $M = 1$ as well as $M = 2$ and $M = 4$, so we apply twelve NN matching estimators in total. We always match with replacement; if there are ties, all of the tied observations are used.

The last class of estimators includes two versions of inverse probability weighting (see, *e.g.*, Horvitz and Thompson 1952, Hirano *et al.* 2003, Wooldridge 2007) as well as two doubly robust estimators (see, *e.g.*, Robins *et al.* 1994, Wooldridge 2007, Uysal 2015, Słoczyński and Wooldridge 2016). We consider normalized reweighting, in which the weights are rescaled to sum to unity, and efficient reweighting, as proposed by Lunceford and Davidian (2004).[15] Our doubly robust estimators combine inverse probability weighting with a model for the conditional mean, and we consider both linear and logistic mean functions. The resulting estimators are consistent if at least one of the two models is correctly specified.

Moreover, for regression-based, Oaxaca–Blinder, and inverse probability weighting estimators we also consider a separate case in which we restrict our estimation procedures to those treated (or placebo treated) whose estimated propensity scores are larger than the minimum and smaller than the maximum estimated propensity score among the nontreated, *i.e.* to those who are located in the overlap region.[16] On the other hand, when we study the impact of NSW on earnings, the number of available estimators is smaller, since we cannot use those methods that require the outcome variable to be binary: logit, probit, and complementary log-log models (6 estimators in total); nonlinear OB methods (4 estimators in total); local logit (1 estimator); and the doubly robust estimator with a logistic mean function (2 estimators in total).

What follows, our final number of estimators for nonemployment (earnings) is equal

[15] Initially, we also considered unnormalized reweighting where the sum of weights is stochastic. However, in line with the results in Frölich (2004), the performance of this estimator was often extremely poor, to the extent that we treat this method as an outlier and leave it out of the analysis.

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

to 39 (26), including 8 (2) regression-based estimators, 6 (2) OB estimators, 5 (4) kernel-based estimators, 12 (12) NN matching estimators, and 8 (6) IPW estimators. We conduct our simulations in Stata and use several user-written commands in our estimation procedures: `locreg` (Frölich and Melly 2010), `nnmatch` (Abadie *et al.* 2004), `oaxaca` (Jann 2008), and `psmatch2` (Leuven and Sianesi 2003).

# 6 Results

This section provides a discussion of our empirical results. We begin by adding a few more remarks about some aspects of our simulation procedures.

The total number of simulation studies that we consider is sixty-four. Recall that we construct eight nonexperimental datasets, which we enumerate in Section 3, and we conduct eight empirical Monte Carlo studies for each of them. We vary three aspects of an EMCS procedure, namely the outcome variable, the set of control variables, and whether the study is designed as placebo or structured. There are two possible choices in each of these three cases, which gives a total of eight combinations.

Further details on our nonexperimental datasets are presented in the Appendix. For each dataset, we provide information on the "true effects" (or $\theta$) and associated standard errors. Moreover, for each dataset, we present summary statistics for the outcome and main control variables, separately for each treatment status. We also present two measures of overlap as well as a matrix of correlations between the main control variables, again conditional on treatment status. A "good" empirical Monte Carlo study, apart from having high internal validity, should be able to replicate these statistics reasonably well. Thus, the same statistics are also presented in the Web Appendix for each of the simulation studies. In general, the simulations replicate the "original datasets" well.

Further analysis proceeds as follows. In each of the nonexperimental datasets, we estimate the impact of NSW on both nonemployment and earnings using estimators discussed in Section 5. For each estimator, we consider two sets of control variables, "simple" and "balanced." Using our notation from Section 2, this gives us thirty-two sets of values of $\hat{\theta}_j^*$.[17] Because we already know the values of $\theta$, we also create thirty-two sets of "true biases," namely $b_j^* = \hat{\theta}_j^* - \theta$, and "absolute true biases," namely $\left|b_j^*\right| = \left|\hat{\theta}_j^* - \theta\right|$. These data are presented in the Web Appendix.

[17]The reason for having thirty-two sets of nonexperimental estimates—and not sixty-four—is very simple. Namely, we use eight nonexperimental datasets and—in each of them—we estimate the effect of NSW on two outcomes (nonemployment and earnings) using two specifications ("simple" and "balanced"). The fact that we also conduct two simulation studies (placebo and structured) for each such combination has no impact on the number of sets of nonexperimental estimates.

We also calculate mean estimates, or $\bar{\hat{\theta}}_j$, mean biases, or $\bar{b}_j = \bar{\hat{\theta}}_j - \bar{\bar{\theta}}$, and absolute mean biases, or $|\bar{b}_j| = \left|\bar{\hat{\theta}}_j - \bar{\bar{\theta}}\right|$, for each estimator in each of the sixty-four simulation studies.[18] The fact that a given study is referred to as "simple" or "balanced" determines the set of control variables in all estimation procedures. These data on biases in simulations are presented in the Web Appendix. Finally, this allows us to construct sixty-four datasets in which an estimator is the unit of observation. For each estimator, we observe $\hat{\theta}_j^*$ and $\bar{\hat{\theta}}_j$; $b_j^*$ and $\bar{b}_j$; and, finally, $\left|b_j^*\right|$ and $|\bar{b}_j|$. In each of these datasets we calculate the correlations between these measures (across estimators), analogous to the theoretical considerations in Section 2, and analyze the results. The disaggregated data on correlation coefficients in each of the sixty-four datasets are presented in the Web Appendix. In the body of the paper we focus on summary statistics for these correlations and we also analyze these data in a regression framework.

## Average Correlations

The first set of results concerns the average values of correlations across simulation studies. Our theoretical results suggest that there is no reason to expect a systematic relationship between the absolute biases in the sample of interest and the absolute mean biases in simulations. In column 3 of Table 1 we can see that on average across our EMCS designs this correlation is indeed close to zero. More precisely, it is equal to –0.054 and, in fact, different from zero at the 10% significance level. It turns out to be the case that this result is driven by placebo simulations. For the placebo design, the average correlation is also negative but larger in absolute value; namely, it is equal to –0.111 and different from zero at the 1% significance level. The average correlation for the structured design is equal to 0.003 and indistinguishable from zero. Although we do not offer any predictions on the average correlations between estimates and between biases, we report these results for completeness in columns 1 and 2 of Table 1.

As a robustness check for our results on absolute mean biases, we also calculate the average correlation between the "true biases" and the median biases in simulations as well as between the "absolute true biases" and the absolute median biases in simulations. This might be important if the results for mean biases are in some way affected by outliers. These results are presented in the Appendix; they are similar to those for mean biases, although we can no longer conclude that, for the placebo design, the average correlation between the absolute biases is significantly different from zero. In other words, we have less reasons to believe that placebo simulations have lower internal validity than

[18] As already discussed, the mean benchmark effect in simulations, $\bar{\bar{\theta}}$, is easily calculated for each study.

Table 1: Average Correlations in the Empirical Monte Carlo Studies

| | Estimates—Mean estimates | Biases—Mean biases | Absolute biases—Absolute mean biases |
|---|---|---|---|
| | (1) | (2) | (3) |
| | All simulations | | |
| Correlation | 0.063* | 0.063* | –0.054* |
| | (0.036) | (0.036) | (0.028) |
| Observations | 64 | 64 | 64 |
| | Placebo | | |
| Correlation | –0.015 | –0.015 | –0.111*** |
| | (0.051) | (0.051) | (0.035) |
| Observations | 32 | 32 | 32 |
| | Structured | | |
| Correlation | 0.141*** | 0.141*** | 0.003 |
| | (0.046) | (0.046) | (0.041) |
| Observations | 32 | 32 | 32 |

*Note:* The values in each cell represent mean correlation coefficients averaged across simulation studies. Standard errors are in parentheses. Tests of significance are two-tailed and the null hypothesis is that of a zero average correlation.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

structured in our application. Hence, our simulation results are in line with our earlier prediction that the internal validity of empirical Monte Carlo studies might be quite low. Neither of the EMCS designs appears to be helpful for selecting estimators in this context.

We also consider an additional robustness check, loosely inspired by Heckman and Hotz (1989), where we study whether empirical Monte Carlo studies, like specification tests in Heckman and Hotz (1989), are at least able to "eliminate the most unreliable and misleading estimators." We operationalize this idea in the following way. For each combination of a nonexperimental dataset, an outcome variable, and a set of control variables, we evaluate which estimators belong to the group of 20% of the estimators with highest "absolute true biases." These are the "worst estimators" we would like to avoid. Then, for each simulation study, we calculate what percentage of these estimators can be found in the top $x$% with lowest absolute mean biases, and we consider $x = 10$, $x = 25$, and $x = 50$. In this test, if $x$% of the worst estimators can be found in the top $x$% of estimators according to an EMCS, then this EMCS is as good as random in helping with estimator choice. If we can find less than $x$% of the worst estimators at the top, then an EMCS is helpful; if more than $x$%, then it is actively misleading. These results are presented in the Appendix. There is no test in which we would conclude that empirical Monte Carlo

studies are actually helpful. In one case, that of placebo simulations with $x = 50$, they might be evaluated as harmful.

## Determinants of Internal Validity

The second set of results concerns the determinants of the magnitudes of the correlation coefficients. For conciseness we omit the results on correlations between estimates, since these results are again numerically identical to the results on biases. We study this problem in a regression framework, using the magnitude of a given correlation coefficient as a dependent variable and characteristics of a given simulation study as independent variables. These results are presented in Table 2.

Turning to the reported estimates, we first examine the determinants of correlations between biases (column 1). We regress the correlation coefficient on all the characteristics of the original dataset and of the simulation study that we are able to control: whether we use the placebo or the structured design; whether we use the CPS or the PSID nonexperimental comparison group; whether we use earnings or nonemployment as the outcome variable; whether we use the Dehejia and Wahba (1999) or the Smith and Todd (2005) version of the NSW dataset; whether we use the "simple" or the "balanced" set of control variables; and whether we use the treatment or the control group from the NSW experiment. It turns out that two out of six coefficients are statistically significant. The placebo design does a worse job than the structured design in replicating biases. On the other hand, it seems easier to replicate the biases for the NSW-CPS datasets than for the NSW-PSID datasets. Moreover, column 2 reports coefficient estimates from a regression with the same independent variables, but with a different dependent variable: the correlation between absolute biases. As before, the estimated coefficient on placebo is negative and statistically significant; thus, at least in our study, the internal validity of placebo is lower than that of structured. It also turns out to be the case that it is more difficult to predict absolute biases using the Smith and Todd (2005) version of the NSW dataset. This is analogous to one of the results in Smith and Todd (2005), namely that it is more difficult to recover the experimental estimate of the impact of NSW using their version of the dataset, as compared with the version of Dehejia and Wahba (1999).

We also develop a further test of our theoretical predictions from Section 2. More precisely, we can now examine whether our rule of thumb for maximizing the correlation between the absolute biases, namely equation (12), is indeed capable of predicting the magnitude of this correlation. We construct a new variable, $\left|\bar{\tilde{\theta}}_{opt} - \bar{\tilde{\theta}}\right| = \left|\alpha_\theta + \beta_\theta \theta - \bar{\tilde{\theta}}\right|$, which measures the distance between the "optimal" value and the actual value of the

Table 2: Regression Analysis of the Simulation Results

| | Biases—Mean biases | Absolute biases—Absolute mean biases | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Placebo | –0.156** | –0.114** | –0.155*** | –0.211 |
| | (0.069) | (0.051) | (0.051) | (0.160) |
| CPS | 0.117* | –0.031 | –0.045 | –0.095 |
| | (0.069) | (0.051) | (0.049) | (0.072) |
| Earnings | 0.095 | –0.025 | –0.024 | –0.073 |
| | (0.069) | (0.051) | (0.050) | (0.078) |
| DW | 0.036 | 0.156*** | 0.132** | 0.170** |
| | (0.069) | (0.051) | (0.051) | (0.074) |
| Balanced | 0.027 | 0.006 | 0.033 | 0.029 |
| | (0.069) | (0.051) | (0.049) | (0.085) |
| Control | 0.025 | 0.036 | 0.047 | 0.098 |
| | (0.069) | (0.051) | (0.050) | (0.078) |
| Std. distance | | | –0.078** | –0.079* |
| | | | (0.035) | (0.046) |
| CPS × Placebo | | | | 0.099 |
| | | | | (0.102) |
| Earnings × Placebo | | | | 0.122 |
| | | | | (0.131) |
| DW × Placebo | | | | –0.067 |
| | | | | (0.109) |
| Balanced × Placebo | | | | 0.010 |
| | | | | (0.112) |
| Control × Placebo | | | | –0.094 |
| | | | | (0.109) |
| Std. distance × Placebo | | | | 0.040 |
| | | | | (0.112) |
| Constant | –0.009 | –0.068 | 0.013 | 0.020 |
| | (0.078) | (0.056) | (0.064) | (0.072) |
| Observations | 64 | 64 | 64 | 64 |
| $R^2$ | 0.156 | 0.211 | 0.262 | 0.311 |

*Note:* The dependent variables are the correlation between the bias in the sample of interest and the mean bias in simulations (column 1) and the correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations (columns 2–4). "Placebo" is an indicator variable that equals one if a given study is designed as placebo and zero otherwise. "CPS" is an indicator variable that equals one if a given study is based on the CPS comparison group and zero otherwise. "Earnings" is an indicator variable that equals one if `re78` is the outcome variable in a given study and zero otherwise. "DW" is an indicator variable that equals one if a given study is based on the Dehejia and Wahba (1999) version of the NSW data and zero otherwise. "Balanced" is an indicator variable that equals one if a given study uses the "balanced" set of control variables and zero otherwise. "Control" is an indicator variable that equals one if a given study is based on the control group from the NSW experiment and zero otherwise. "Std. distance" is the standardized distance between the "optimal" value and the actual value of the mean benchmark effect in simulations. Huber–White standard errors are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

mean benchmark effect in simulations. Intuitively, we would expect that, other things equal, the more distant we are from the correlation-maximizing value of the mean benchmark effect, the smaller is the correlation between the absolute biases. Because distance is measured in different units (dollars or percentage points) dependent on the outcome variable in a given simulation study (earnings or nonemployment), we begin by standardizing this new variable; more precisely, we divide raw values of $\bar{\bar{\theta}}_{opt} - \bar{\bar{\theta}}$ by its standard deviation, separately for studies of earnings and nonemployment, and then calculate absolute values. Finally, we add this "standardized distance" to our earlier set of independent variables and report the estimated coefficients in column 3. In line with our predictions, the coefficient on standardized distance is negative; it is also statistically significant. Thus, our rule of thumb does indeed have predictive power for the internal validity of empirical Monte Carlo studies.

Column 4 reports the final set of estimated coefficients, where we interact the indicator for placebo simulations with the remaining independent variables, including standardized distance. Thus, we examine whether the effects of these variables on our correlation of interest differ between placebo and structured simulations. The answer to this question is negative. Neither of the estimated coefficients is statistically significant, and quite clearly so, with all $t$-statistics for the interaction terms smaller than one.

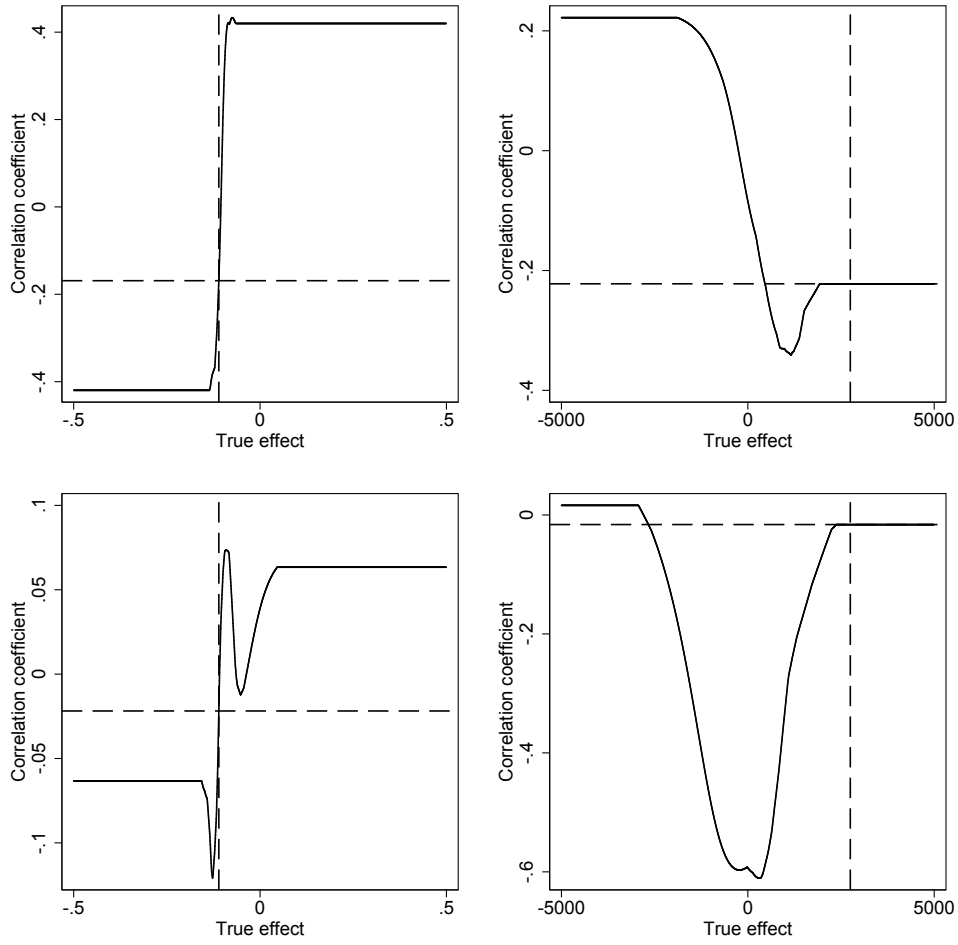
## Bounding the Effect of Interest

The third set of results concerns our earlier claim that it might sometimes be helpful to place some bounds on the population parameter of interest—in the hope that this might allow one to conclude that a given simulation study is at least informative under the condition that these bounds are correct. In other words, our results thus far suggest that one should not expect *a priori* that an empirical Monte Carlo study will be internally valid. However, for a simulation study that has already been conducted, we might perhaps be able to claim its internal validity by making additional assumptions *a posteriori*.

For each of our sixty-four simulation studies we calculate how the correlation between the absolute biases would change with $\theta$, holding $\bar{\bar{\theta}}$ fixed at its actual value. These figures are presented in the Web Appendix. Due to space restrictions, in the body of the paper we restrict our attention to only four cases which we present in Figure 5. The upper segments focus on two placebo simulations, while the lower segments focus on structured. The left-hand segments of Figure 5 present the dependence of our correlation of interest on $\theta$ for two studies of nonemployment; the right-hand segments—for two studies of earnings.

The upper left segment of Figure 5 is an example of a pessimistic scenario. Since this

Figure 5: The Dependence of $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ on $\theta$ in Four Simulation Studies




*Note:* All segments present the dependence of $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$ on $\theta$ in a particular simulation study, holding $\bar{\bar{\theta}}$ fixed at its actual value. The upper left segment corresponds to a placebo simulation study, based on the "DW treated/CPS" dataset, with `u78` as the outcome variable and the "simple" set of control variables. The upper right segment corresponds to a placebo simulation study, based on the "ST treated/CPS" dataset, with `re78` as the outcome variable and the "balanced" set of control variables. The lower left segment corresponds to a structured simulation study, based on the "DW treated/CPS" dataset, with `u78` as the outcome variable and the "balanced" set of control variables. The lower right segment corresponds to a structured simulation study, based on the "ST treated/PSID" dataset, with `re78` as the outcome variable and the "simple" set of control variables. Dashed lines represent the actual values of $\theta$ and $\text{Cor}(\left|\bar{b}_j\right|, \left|b_j^*\right|)$.

is a study of the effects of a training program on nonemployment, we might perhaps be willing to assume that this effect in nonpositive. Unfortunately, for negative values of $\theta$ both highly positive and highly negative values of the correlation between the absolute biases are possible. Moreover, these values are drastically different in a very small interval of values of $\theta$. In particular, if $\theta > -0.09$, the correlation is larger than 0.4; if $\theta < -0.13$,

the correlation is smaller than –0.4. In this case it would not be reasonable to bound $\theta$, since we observe very different values of our correlation of interest for plausible—and very similar—values of this parameter.

The lower left segment of Figure 5 presents a further example of a simulation study which is unlikely to be useful. As before, the assumption that the effects of training on nonemployment are negative would provide no information about the sign of our correlation coefficient of interest. Moreover, this correlation is uniformly very close to zero—never smaller than –0.117 and never larger than 0.073—and hence the predictive power of this simulation study is extremely low even in the best-case scenario.

A somewhat different situation is presented in the upper right segment of Figure 5. Since we now focus on the effects of training on earnings, it might be reasonable to assume that $\theta$ is nonnegative. If this is the case, however, our correlation of interest is guaranteed to be negative; more precisely, it will be smaller than –0.084 but not smaller than –0.341. At first, this might seem to be bad news—according to our definition, internal validity of this study is certainly low as long as $\theta$ is indeed nonnegative. However, this study can still be helpful in estimator choice—if only we decide to choose estimators which perform *badly* in this simulation study. If we did this, we would increase our chances of using an estimator with a small value of absolute true bias—although this help would be somewhat restricted by the fact that the magnitudes of these correlations are quite low.

Moreover, the lower right segment of Figure 5 is another example of a more optimistic outcome. Again, we focus on the effects of a training program on earnings and we might perhaps be willing to assume that this effect is positive. In this case our correlation of interest would again be guaranteed to be negative; it could be as strong as –0.611 but also much closer to zero. As before, this is a case where, if anything, one would want to choose estimators which perform *badly* in the simulations.

In other cases, we would expect that bounding the population parameter could sometimes allow us to conclude that our correlation of interest must be positive. Then, of course, we would choose estimators which perform well in a given EMCS. However, no such cases are available among the simulation studies that we have conducted.

# 7 Conclusions

In this paper we contribute to the recent literature on the finite-sample performance of estimators of average treatment effects under unconfoundedness. Influential papers by Huber *et al.* (2013) and Busso *et al.* (2014) suggest that an empirically motivated simulation exercise might be informative about the performance of estimators in a particular

context. In this paper we claim that such a high degree of internal validity of empirical Monte Carlo studies is, in fact, far from self-evident. We begin by developing a simple framework within which we demonstrate that there is little reason to expect our preferred measure of internal validity—the correlation between the absolute bias in the sample of interest and the absolute mean bias in simulations—to be systematically positive. We also show that this measure is dependent on the true value of the population parameter of interest. This is an important limitation of EMCS procedures, since were this object known, no simulation studies and no estimation procedures would be necessary.

We consider two different designs of an empirical Monte Carlo study. The first, which we term the structured design, is based on Busso *et al.* (2014). Here we generate new data which match particular features of the original dataset, and then generate outcomes using parameters estimated from the original data. We also consider the placebo design, as suggested by Huber *et al.* (2013). Here a sample of observations is drawn from the comparison data, and a placebo treatment is assigned using the propensity score from the full data. The treatment effect in the sample is therefore zero by construction.

Using the data on men from the influential papers by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), and Smith and Todd (2001, 2005), we construct an empirical test for both designs. Our results are quite negative—but they also corroborate our theoretical predictions. The average degree of internal validity of EMCS procedures is very low. In fact, we have no evidence that practitioners would be better off, *on average*, using an empirical Monte Carlo study than choosing estimators at random. However, we also demonstrate that if a simulation study has already been run then it is generally possible to assess how likely we are to learn something useful from its results. We discuss two actual cases of a simulation study which might, in fact, be helpful.

Consequently, we would generally discourage applied researchers who consider conducting an empirical Monte Carlo study. In our view, the expected gains are very limited. But, for clarity, we also note that we are far from dismissing empirical Monte Carlo studies in their entirety. The contributions of Huber *et al.* (2013) and Busso *et al.* (2014) are a very reasonable response to the limitations of the earlier literature, which had been assessing the performance of estimators using only data-generating processes that were entirely unlike what applied researchers encounter in practice. In this paper we merely state that the EMCS approach has more limitations than Huber *et al.* (2013) and Busso *et al.* (2014) might have hoped. Instead empirical researchers would be best advised to continue using different methods, as Busso *et al.* (2014) also suggest, as an important robustness check.

# 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.

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.

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

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

BREWER, M., T. F. CROSSLEY, AND R. JOYCE (2013): "Inference with Difference-in-Differences Revisited," IZA Discussion Paper no. 7742.

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.

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.

CAMPBELL, D. T. AND J. C. STANLEY (1963): "Experimental and Quasi-Experimental Designs for Research on Teaching," in *Handbook of Research on Teaching*, ed. by N. L. Gage, Rand McNally & Company, 171–246.

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.

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

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

FORTIN, N., T. LEMIEUX, AND S. FIRPO (2011): "Decomposition Methods in Economics," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter and D. Card, Elsevier, vol. 4, 1–102.

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 (2015): "The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation," IZA Discussion Paper no. 8756.

FRÖLICH, M. AND B. MELLY (2010): "Estimation of Quantile Treatment Effects with Stata," *Stata Journal*, 10, 423–457.

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

HECKMAN, J. J. AND 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.

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

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

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

HOROWITZ, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 5, 3159–3228.

HORVITZ, D. G. AND D. J. THOMPSON (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

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.

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

Punch, K. F. (2014): *Introduction to Social Research: Quantitative & Qualitative Approaches*, SAGE.

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

Słoczyński, T. and J. M. Wooldridge (2016): "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, forthcoming.

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.

UYSAL, S. D. (2015): "Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling," *Journal of Applied Econometrics*, 30, 763–786.

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

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

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

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

# Appendix

Table A1: "True Effects" in Nonexperimental Datasets

| Outcome variable | Version of NSW data | Subset of NSW data | Effect | Std. error |
|---|---|---|---|---|
| re78 | DW | treated | 1,794 | 671 |
| re78 | DW | control | 0 | 0 |
| re78 | ST | treated | 2,748 | 1,005 |
| re78 | ST | control | 0 | 0 |
| u78 | DW | treated | –0.1106 | 0.0434 |
| u78 | DW | control | 0 | 0 |
| u78 | ST | treated | –0.1744 | 0.0570 |
| u78 | ST | control | 0 | 0 |

*Note:* Values in this table are used as "true effects," or $\theta$, for our calculations of "true biases" and "absolute true biases" in nonexperimental datasets. "DW" refers to the Dehejia and Wahba (1999) version of the NSW data. "ST" refers to the Smith and Todd (2005) version of the NSW data. "Subset of NSW data" refers to whether we compare the treatment or the control group from the NSW experiment with the nonexperimental comparison groups, CPS-1 and PSID-1. If we use the treatment group, then $\theta$ is estimated, following LaLonde (1986), using the experimental data—as the coefficient on treatment in a univariate regression of the outcome variable on the treatment indicator. Huber–White standard errors are also presented. If we use the control group, then $\theta$ is zero by construction and is not subject to sampling error, as suggested by Smith and Todd (2005).

Table A2: Summary Statistics for the "DW control/CPS" Dataset

| | Comparison | | Control | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.136 | 0.343 | 0.354 | 0.479 |
| re78 | 14,847 | 9,647 | 4,555 | 5,484 |
| age | 33.23 | 11.05 | 25.05 | 7.058 |
| educ | 12.03 | 2.871 | 10.09 | 1.614 |
| married | 0.712 | 0.453 | 0.154 | 0.361 |
| black | 0.0735 | 0.261 | 0.827 | 0.379 |
| u74 | 0.120 | 0.325 | 0.750 | 0.434 |
| u75 | 0.109 | 0.312 | 0.685 | 0.466 |
| re74 | 14,017 | 9,570 | 2,107 | 5,688 |
| re75 | 13,651 | 9,270 | 1,267 | 3,103 |
| Observations | 15,992 | | 260 | |

*Note:* The comparison group in this dataset is "CPS-1," as constructed by LaLonde (1986). It is accompanied by the Dehejia and Wahba (1999) version of the control group from the NSW experiment.

Table A3: Summary Statistics for the "DW treated/CPS" Dataset

| | Comparison | | Treated | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.136 | 0.343 | 0.243 | 0.430 |
| re78 | 14,847 | 9,647 | 6,349 | 7,867 |
| age | 33.23 | 11.05 | 25.82 | 7.155 |
| educ | 12.03 | 2.871 | 10.35 | 2.011 |
| married | 0.712 | 0.453 | 0.189 | 0.393 |
| black | 0.0735 | 0.261 | 0.843 | 0.365 |
| u74 | 0.120 | 0.325 | 0.708 | 0.456 |
| u75 | 0.109 | 0.312 | 0.600 | 0.491 |
| re74 | 14,017 | 9,570 | 2,096 | 4,887 |
| re75 | 13,651 | 9,270 | 1,532 | 3,219 |
| Observations | 15,992 | | 185 | |

*Note:* The comparison group in this dataset is "CPS-1," as constructed by LaLonde (1986). It is accompanied by the Dehejia and Wahba (1999) version of the treatment group from the NSW experiment.

Table A4: Summary Statistics for the "ST control/CPS" Dataset

| | Comparison | | Control | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.136 | 0.343 | 0.387 | 0.489 |
| re78 | 14,847 | 9,647 | 4,609 | 6,032 |
| age | 33.23 | 11.05 | 26.01 | 7.108 |
| educ | 12.03 | 2.871 | 10.27 | 1.572 |
| married | 0.712 | 0.453 | 0.190 | 0.394 |
| black | 0.0735 | 0.261 | 0.817 | 0.388 |
| u74 | 0.120 | 0.325 | 0.542 | 0.500 |
| u75 | 0.109 | 0.312 | 0.472 | 0.501 |
| re74 | 14,017 | 9,570 | 3,858 | 7,254 |
| re75 | 13,651 | 9,270 | 2,277 | 3,919 |
| Observations | 15,992 | | 142 | |

*Note:* The comparison group in this dataset is "CPS-1," as constructed by LaLonde (1986). It is accompanied by the Smith and Todd (2005) version of the control group from the NSW experiment.

Table A5: Summary Statistics for the "ST treated/CPS" Dataset

| | Comparison | | Treated | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.136 | 0.343 | 0.213 | 0.411 |
| re78 | 14,847 | 9,647 | 7,357 | 9,027 |
| age | 33.23 | 11.05 | 25.37 | 6.251 |
| educ | 12.03 | 2.871 | 10.49 | 1.643 |
| married | 0.712 | 0.453 | 0.204 | 0.405 |
| black | 0.0735 | 0.261 | 0.824 | 0.383 |
| u74 | 0.120 | 0.325 | 0.500 | 0.502 |
| u75 | 0.109 | 0.312 | 0.324 | 0.470 |
| re74 | 14,017 | 9,570 | 3,590 | 5,971 |
| re75 | 13,651 | 9,270 | 2,596 | 3,872 |
| Observations | 15,992 | | 108 | |

*Note:* The comparison group in this dataset is "CPS-1," as constructed by LaLonde (1986). It is accompanied by the Smith and Todd (2005) version of the treatment group from the NSW experiment.

Table A6: Summary Statistics for the "DW control/PSID" Dataset

| | Comparison | | Control | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.115 | 0.319 | 0.354 | 0.479 |
| re78 | 21,554 | 15,555 | 4,555 | 5,484 |
| age | 34.85 | 10.44 | 25.05 | 7.058 |
| educ | 12.12 | 3.082 | 10.09 | 1.614 |
| married | 0.866 | 0.340 | 0.154 | 0.361 |
| black | 0.251 | 0.433 | 0.827 | 0.379 |
| u74 | 0.0863 | 0.281 | 0.750 | 0.434 |
| u75 | 0.100 | 0.300 | 0.685 | 0.466 |
| re74 | 19,429 | 13,407 | 2,107 | 5,688 |
| re75 | 19,063 | 13,597 | 1,267 | 3,103 |
| Observations | 2,490 | | 260 | |

*Note:* The comparison group in this dataset is "PSID-1," as constructed by LaLonde (1986). It is accompanied by the Dehejia and Wahba (1999) version of the control group from the NSW experiment.

Table A7: Summary Statistics for the "DW treated/PSID" Dataset

| | Comparison | | Treated | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.115 | 0.319 | 0.243 | 0.430 |
| re78 | 21,554 | 15,555 | 6,349 | 7,867 |
| age | 34.85 | 10.44 | 25.82 | 7.155 |
| educ | 12.12 | 3.082 | 10.35 | 2.011 |
| married | 0.866 | 0.340 | 0.189 | 0.393 |
| black | 0.251 | 0.433 | 0.843 | 0.365 |
| u74 | 0.0863 | 0.281 | 0.708 | 0.456 |
| u75 | 0.100 | 0.300 | 0.600 | 0.491 |
| re74 | 19,429 | 13,407 | 2,096 | 4,887 |
| re75 | 19,063 | 13,597 | 1,532 | 3,219 |
| Observations | 2,490 | | 185 | |

*Note:* The comparison group in this dataset is "PSID-1," as constructed by LaLonde (1986). It is accompanied by the Dehejia and Wahba (1999) version of the treatment group from the NSW experiment.

Table A8: Summary Statistics for the "ST control/PSID" Dataset

| | Comparison | | Control | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.115 | 0.319 | 0.387 | 0.489 |
| re78 | 21,554 | 15,555 | 4,609 | 6,032 |
| age | 34.85 | 10.44 | 26.01 | 7.108 |
| educ | 12.12 | 3.082 | 10.27 | 1.572 |
| married | 0.866 | 0.340 | 0.190 | 0.394 |
| black | 0.251 | 0.433 | 0.817 | 0.388 |
| u74 | 0.0863 | 0.281 | 0.542 | 0.500 |
| u75 | 0.100 | 0.300 | 0.472 | 0.501 |
| re74 | 19,429 | 13,407 | 3,858 | 7,254 |
| re75 | 19,063 | 13,597 | 2,277 | 3,919 |
| Observations | 2,490 | | 142 | |

*Note:* The comparison group in this dataset is "PSID-1," as constructed by LaLonde (1986). It is accompanied by the Smith and Todd (2005) version of the control group from the NSW experiment.

Table A9: Summary Statistics for the "ST treated/PSID" Dataset

| | Comparison | | Treated | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| u78 | 0.115 | 0.319 | 0.213 | 0.411 |
| re78 | 21,554 | 15,555 | 7,357 | 9,027 |
| age | 34.85 | 10.44 | 25.37 | 6.251 |
| educ | 12.12 | 3.082 | 10.49 | 1.643 |
| married | 0.866 | 0.340 | 0.204 | 0.405 |
| black | 0.251 | 0.433 | 0.824 | 0.383 |
| u74 | 0.0863 | 0.281 | 0.500 | 0.502 |
| u75 | 0.100 | 0.300 | 0.324 | 0.470 |
| re74 | 19,429 | 13,407 | 3,590 | 5,971 |
| re75 | 19,063 | 13,597 | 2,596 | 3,872 |
| Observations | 2,490 | | 108 | |

*Note:* The comparison group in this dataset is "PSID-1," as constructed by LaLonde (1986). It is accompanied by the Smith and Todd (2005) version of the treatment group from the NSW experiment.

Table A10: Correlation Matrix for the DW Treated Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | –0.008 | 1.000 | | | | | | |
| black | 0.053 | –0.037 | 1.000 | | | | | |
| married | 0.241 | 0.006 | –0.019 | 1.000 | | | | |
| re74 | –0.001 | 0.125 | 0.028 | 0.191 | 1.000 | | | |
| re75 | 0.071 | 0.005 | –0.030 | 0.290 | 0.633 | 1.000 | | |
| u74 | 0.153 | –0.073 | 0.018 | –0.115 | –0.670 | –0.533 | 1.000 | |
| u75 | 0.141 | –0.057 | 0.073 | –0.141 | –0.508 | –0.584 | 0.738 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables. "DW" refers to the Dehejia and Wahba (1999) version of the NSW data.

Table A11: Correlation Matrix for the DW Control Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | 0.044 | 1.000 | | | | | | |
| black | 0.109 | 0.113 | 1.000 | | | | | |
| married | 0.183 | 0.149 | 0.054 | 1.000 | | | | |
| re74 | –0.001 | –0.021 | –0.008 | 0.112 | 1.000 | | | |
| re75 | 0.036 | 0.040 | –0.077 | 0.248 | 0.675 | 1.000 | | |
| u74 | 0.069 | –0.106 | 0.041 | –0.049 | –0.643 | –0.588 | 1.000 | |
| u75 | 0.035 | –0.184 | 0.018 | –0.055 | –0.475 | –0.603 | 0.717 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables. "DW" refers to the Dehejia and Wahba (1999) version of the NSW data.

Table A12: Correlation Matrix for the ST Treated Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | 0.060 | 1.000 | | | | | | |
| black | 0.090 | –0.070 | 1.000 | | | | | |
| married | 0.181 | 0.129 | –0.008 | 1.000 | | | | |
| re74 | 0.042 | 0.161 | 0.068 | 0.239 | 1.000 | | | |
| re75 | 0.145 | –0.045 | –0.010 | 0.363 | 0.574 | 1.000 | | |
| u74 | 0.202 | –0.051 | –0.024 | –0.138 | –0.604 | –0.415 | 1.000 | |
| u75 | 0.239 | –0.014 | 0.060 | –0.154 | –0.391 | –0.466 | 0.613 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables. "ST" refers to the Smith and Todd (2005) version of the NSW data.

Table A13: Correlation Matrix for the ST Control Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | 0.069 | 1.000 | | | | | | |
| black | 0.060 | 0.106 | 1.000 | | | | | |
| married | 0.143 | 0.167 | 0.044 | 1.000 | | | | |
| re74 | –0.074 | –0.094 | 0.003 | 0.099 | 1.000 | | | |
| re75 | –0.023 | –0.007 | –0.094 | 0.281 | 0.631 | 1.000 | | |
| u74 | 0.233 | –0.065 | 0.040 | 0.013 | –0.581 | –0.504 | 1.000 | |
| u75 | 0.157 | –0.193 | –0.027 | –0.027 | –0.409 | –0.551 | 0.670 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables. "ST" refers to the Smith and Todd (2005) version of the NSW data.

Table A14: Correlation Matrix for the CPS Comparison Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | –0.128 | 1.000 | | | | | | |
| black | –0.014 | –0.102 | 1.000 | | | | | |
| married | 0.440 | –0.002 | –0.057 | 1.000 | | | | |
| re74 | 0.407 | 0.093 | –0.076 | 0.421 | 1.000 | | | |
| re75 | 0.350 | 0.121 | –0.082 | 0.395 | 0.870 | 1.000 | | |
| u74 | –0.105 | –0.023 | 0.018 | –0.179 | –0.540 | –0.485 | 1.000 | |
| u75 | –0.017 | 0.014 | 0.042 | –0.092 | –0.445 | –0.516 | 0.600 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables.

Table A15: Correlation Matrix for the PSID Comparison Units

| | age | educ | black | married | re74 | re75 | u74 | u75 |
|---|---|---|---|---|---|---|---|---|
| age | 1.000 | | | | | | | |
| educ | –0.236 | 1.000 | | | | | | |
| black | –0.038 | –0.324 | 1.000 | | | | | |
| married | 0.199 | –0.022 | –0.132 | 1.000 | | | | |
| re74 | 0.199 | 0.306 | –0.211 | 0.164 | 1.000 | | | |
| re75 | 0.169 | 0.325 | –0.216 | 0.148 | 0.841 | 1.000 | | |
| u74 | 0.168 | –0.036 | –0.059 | –0.001 | –0.446 | –0.334 | 1.000 | |
| u75 | 0.162 | –0.062 | –0.023 | –0.007 | –0.363 | –0.467 | 0.670 | 1.000 |

*Note:* The values in each cell represent the pairwise correlation coefficients between the main control variables.

Table A16: Overlap in Nonexperimental Datasets

| Comparison group | Version of NSW data | Control variables | Subset of NSW data | Measure 1 | Measure 2 |
|---|---|---|---|---|---|
| CPS | DW | balanced | control | 0.394 | 1.000 |
| CPS | DW | balanced | treated | 0.249 | 1.000 |
| CPS | DW | simple | control | 0.539 | 1.000 |
| CPS | DW | simple | treated | 0.457 | 1.000 |
| CPS | ST | balanced | control | 0.420 | 1.000 |
| CPS | ST | balanced | treated | 0.210 | 1.000 |
| CPS | ST | simple | control | 0.609 | 1.000 |
| CPS | ST | simple | treated | 0.441 | 1.000 |
| PSID | DW | balanced | control | 0.444 | 0.998 |
| PSID | DW | balanced | treated | 0.465 | 1.000 |
| PSID | DW | simple | control | 0.515 | 0.998 |
| PSID | DW | simple | treated | 0.574 | 0.999 |
| PSID | ST | balanced | control | 0.410 | 0.998 |
| PSID | ST | balanced | treated | 0.445 | 0.999 |
| PSID | ST | simple | control | 0.524 | 1.000 |
| PSID | ST | simple | treated | 0.575 | 0.999 |

*Note:* "DW" refers to the Dehejia and Wahba (1999) version of the NSW data. "ST" refers to the Smith and Todd (2005) version of the NSW data. "Simple" and "balanced" sets of control variables are explained in detail in footnote 8. "Subset of NSW data" refers to whether we compare the treatment or the control group from the NSW experiment with the nonexperimental comparison groups, CPS-1 and PSID-1. "Measure 1" is calculated as the proportion of all units whose estimated propensity scores are not larger than the smaller of the two subgroup-specific maximums and not smaller than the larger of the two subgroup-specific minimums of the estimated propensity score. "Measure 2" is calculated as the proportion of all units whose estimated propensity scores are not larger than the maximum and not smaller than the minimum estimated propensity score among the comparison units.

Table A17: Average Correlations in the Empirical Monte Carlo Studies: Median Biases

| | Biases—Median biases | Absolute biases—Absolute median biases |
|---|---|---|
| | All simulations | |
| Correlation | 0.109*** | –0.037 |
| | (0.034) | (0.025) |
| Observations | 64 | 64 |
| | Placebo | |
| Correlation | 0.090* | –0.052 |
| | (0.047) | (0.035) |
| Observations | 32 | 32 |
| | Structured | |
| Correlation | 0.128** | –0.022 |
| | (0.051) | (0.037) |
| Observations | 32 | 32 |

*Note:* The values in each cell represent mean correlation coefficients averaged across simulation studies. Standard errors are in parentheses. Tests of significance are two-tailed and the null hypothesis is that of a zero average correlation.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table A18: Eliminating the "Worst Estimators" in EMCS

| | Top 10% | Top 25% | Top 50% |
|---|---|---|---|
| | All simulations | | |
| Proportion | 0.103 | 0.263 | 0.588*** |
| | (0.016) | (0.023) | (0.032) |
| Observations | 64 | 64 | 64 |
| | Placebo | | |
| Proportion | 0.091 | 0.272 | 0.620*** |
| | (0.019) | (0.029) | (0.042) |
| Observations | 32 | 32 | 32 |
| | Structured | | |
| Proportion | 0.115 | 0.254 | 0.555 |
| | (0.026) | (0.037) | (0.048) |
| Observations | 32 | 32 | 32 |

*Note:* The values in each cell represent mean proportions of the worst estimators—defined as those that belong to the group of 20% of the estimators with highest absolute true biases—that belong to the top $x$% with lowest absolute mean biases in simulations, where $x = 10$, $x = 25$, or $x = 50$. In this test, if $x$% of the worst estimators belong to the top $x$% of estimators according to an EMCS, then this EMCS is as good as random in helping with estimator choice. If less than $x$% of the worst estimators belong at the top, then an EMCS is helpful; if more than $x$%, then it is actively misleading. Standard errors are in parentheses. Tests of significance are two-tailed and the null hypothesis is that of the average proportion equal to $x/100$.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.