How Bad is Unconfoundedness in Practice? Evidence from Randomised Controlled Trials with Imperfect Compliance*

Sylvain Chabé-Ferret† Jasmin Fliegner‡ Roland Rathelot§

January 31, 2018

Preliminary, Please do not quote

Abstract

Observational methods relying on the unconfoundedness assumption (e.g. OLS or matching) might be severely biased because of unobserved confounders. In this paper, we propose to measure this bias using randomised controlled trials with imperfect compliance. Contrary to most earlier attempts, our method does not require the collection of additional data on non-participants and does not suffer from the bias due to using different survey instruments for participants and non-participants. Our proposed approach can also accommodate encouragement designs where there are treated individuals in the control group. We also introduce a new decomposition of the bias of observational methods in two components: one due to unobserved confounders and the other to lack of common support. As a proof of concept, we run this decomposition on publicly-available data corresponding to 6 published papers, evaluating programmes in education, labor, micro-finance and health (most of them in developing countries) using both a simple approach conditioning linearly on all observed confounders, and an approach using machine learning to select the relevant covariates. In most cases, we find that the bias after conditioning on the observed covariates present in the datasets is as large as the bias before conditioning on anything. Our results suggest that the covariates we observe in this context are poor predictors of selection bias. We also find that the component of the bias of observational methods due to a failure of common support is generally small. We are hopeful to propose a framework in which future datasets could easily be added to enrich the analysis and to investigate which confounders matter most for a given type of programme and outcome.

Keywords: randomised experiments, matching, selection bias, public policy evaluation, conditional independence assumption.
JEL: C21, C26, D04.

*We would like to thank Helene Erkel-Rousse for supporting this project since its inception, students of the Econometrics of Program Evaluation class at TSE for analyzing the datasets examined in this paper prior to our own extensive analysis and participants at the IAST Tuesday workshop for fruitful discussions and comments. All errors are our own.

†TSE, Inra and Iast
‡TSE, BDPEMS, TU Berlin
§University of Warwick, CEPR

1 Introduction

Evaluating the impact a policy or programme has on its beneficiaries is key to policy makers. In the recent years, governments and administrations have increased their commitment to evaluate their actions. However, while randomised controlled trials (RCTs) are the gold standard for impact evaluation, they remain the exception rather than the norm. The push for evaluation has led to an increase in ex-post evaluations. In a few cases, a discontinuity in the assignment of the policy or the existence of an obvious instrument makes it possible to use quasi-experimental methods. In some other cases, past measures of the outcomes are available and the common-trend assumption looks reasonable enough to allow the use of differences-in-differences. Otherwise, practioners are left with comparing treated and untreated individuals, controlling for whatever available observable characteristics. The latter methods rely on the identification assumption called "conditional independence assumption" (CIA) or "unconfoundedness", which states that, conditional on this set of covariates, no other determinant of the outcome is correlated with the treatment status. The theoretical diagnosis about the inability of observational methods to tackle selection bias in general is clearly established. However, the increasing demand for evaluation by policy makers pushes researchers to understand how biased these methods are in practice. This paper has two contributions. First, we present a method to compare experimental results with those relying on unconfoundedness that applies to a large set of settings. Second, we illustrate our papers using the data from 6 recent papers evaluating programmes by RCT and computes the bias in all these cases.

The empirical literature investigating how biased observational methods are dates back to LaLonde (1986), who was the first to use the results of an RCT to assess the bias of non-experimental method – a linear regression in his case. Using an external data source as a control group, he finds that econometric methods fail to replicate the experimental results. LaLonde's seminal paper has been influential to the profession but only a handful of papers have followed up. Many of these papers use RCTs where randomization occurs after self-selection in the program. In this case, researchers have to collect additional data on non-participants to compute an observational estimator. Collecting data not included in the original experiment makes is costly and might generate biases due to the difference between the survey instruments used the experimental and non experimental samples. When researchers use the same survey instrument, they can also make mistakes when determining who is an eligible non-participant because it is difficult to check eligibility retrospectively.

We propose a method to measure the bias of observational methods that does not require collecting additional data. Our method relies on RCTs with imperfect compliance, i.e. there are both treated and untreated individuals in the group assigned to treatment (or control). In this case, self-selection occurs after random assignment of an offer or an encouragement to take the treatment. Within the group assigned to treatment, we can form an observational estimator by comparing the treated with the untreated individuals. In encouragement designs, we can also form an observational estimator with the group assigned to control because individuals assigned to this group are allowed to take up the treatment. We then compare the observational estimators to the experimental estimator of the treatment effect computed using Instrumental Variables (IV). In the eligibility design, the experimental estimator that we use is the Bloom estimator, which adjusts the intention to treat estimator by the proportion of participants in the treatment group. Under mild conditions, we show that the experimental and observational estimates coincide if all relevant confounders are observed and that the difference between experimental and observational estimates measures the bias of the observational method. In the encouragement design, the experimental estimate that we use is the Wald estimator. Comparing observational and experimental estimates requires care since the observational estimates are Treatment on the Treated parameters (TT) while the experimental estimate is a Local Average Treatment Effect (LATE). We show that, under mild conditions, combining the observational TTs of each arm in a Wald-like ratio yields the experimental LATE if all relevant confounders are observed.

We propose a new decomposition of the bias of observational methods between a component due to unobserved variables, that might decrease as we observe more covariates, and a component due to a failure of common support, that will not decrease even if we observe more covariates. Following Heckman et al. (1998a), we define a failure of common support as the impossibility to find non-participants with covariates similar to the participants. Recovering treatment effects outside of the common support is impossible with usual non-parametric observational methods. The observational estimator that we use is the Local Quadratic Regression Matching on the Propensity Score (LLRMPS) proposed by Heckman et al. (1998b). The specification of the propensity score is a key stage for the LLRMPS. In order to minimise human intervention, we use two alternative automatic specifications. In the first one, all available covariates enter linearly in the propensity score regression. In the second one, we also include higher-order polynomials and interactions between these covariates and use a lasso logit estimator to select the best specification.

Flolowing Belloni et al. (2014), we select as our final set of control covariates the ones that matter for selection and/or for outcomes using a double lasso procedure. On top of lasso for selecting the covariates, we implement automatic bandwidth selectors so that most tuning parameters in our analysis are automatically set up. We believe that automatizing covariates and bandwidth selection enables to avoid suspicion of specification searches that have plagued earlier debates on the properties of observational methods.

We apply our decomposition on 6 recently published papers with data publicly available online.[1] These papers are evaluations of education, labour, micro-finance and health programmes and all but one take place in developing countries. Our results show that, for most programmes, the bias after conditioning on the observed covariates present in the datasets is as large as the bias before conditioning on anything. Over all of our outcomes and treatments, the average of the absolute value of selection bias is equal to 18% of a standard deviation of outcomes, the bias of the linear matching estimator to 17.5% and the bias of the double lasso matching estimator to 16%. We also find that the component of the bias of observational methods due to a failure of common support is generally small. We see this paper as a proof of concept and hope to be able to apply our method to many more past and future RCTs with imperfect compliance.

Our paper contributes to the literature initiated by LaLonde (1986), Fraker and Maynard (1987), Heckman and Hotz (1989) and Heckman et al. (1998a) comparing the results of observational methods to that of RCTs. All of these papers and subsequent ones in the literature[2] use RCTs where randomization occurs after self-selection in the program (with the exception of Bléhaut and Rathelot (2014)). Our contribution to this literature is to propose a method that takes advantage of the increase in the number of RCTs with imperfect compliance and studies where authors provide their data openly. While all previous papers focus on one particular context, our method allows to use multiple studies. In the long run, the number of studies could become large enough to consider the bias as an outcome and understand which features are correlated with a higher or a lower bias. Our paper also extends the decomposition of selection bias proposed by Heckman et al. (1998a). As Heckman et al. (1998a), we distinguish between bias due to a failure of the common support assumption, observed confounders and unobserved confounders. Unlike Heckman et al. (1998a), we split the bias due to a failure of common support into

[1] Angrist et al. (2002, 2009); Behaghel et al. (2014); Burde and Linden (2013); Drexler et al. (2014); Oster and Thornton (2011).

[2] Agodini and Dynarski (2004); Arceneaux et al. (2006); Dehejia and Wahba (2002, 1999); Ferraro and Miranda (2014); Fraker and Maynard (1987); Friedlander and Robins (1995); Griffen and Todd (2017); Smith and Todd (2005).

two separate terms and show that only the second one contributes to the bias of the observational method. In medicine, Benson and Hartz (2000) and Concato et al. (2000) use meta-analysis to compare observational and experimental estimates of similar treatments, but obtained on different samples. Their conclusions are starkingly different from ours: observational methods seem largely vindicated. We interpret this discrepancy as meaning that selection bias is less severe in medicine, or that MDs are better at observing the relevant confounders, or that the meta-analysis approach to estimating the bias of observational studies is fraught with problems Kunz et al. (2000); Pocock and Elbourne (2000). In economics, Card et al. (2010, 2015) similarly find no differences in observational and experimental estimates of the effects of Job Training Programs (JTPs) in their meta-analysis. Recently, Eckles and Bakshy (2017) estimate the bias of observational methods in order to estimate peer effects on a social network using an RCT as a benchmark. They show that conditioning on rich information on activity on the network drives the bias of the observational method to zero. Gordon et al. (2017) show that observational methods cannot reproduce the results of 15 online experiments evaluating advertising campaigns. This paper is structured as follows: Section 2 presents the theoretical results underpinning the estimation of the bias of observational methods using RCTs with imperfect compliance. Section 3 details our estimation procedures. Section 4 presents the data that we use. Section 5 presents our main results and Section 6 concludes.

2 Theoretical results

The key to our approach is to use the same data to form an experimental and a non-experimental estimator of the treatment effect. RCTs with imperfect compliance offer that possibility under mild conditions that we make clear in this section.

R_i denotes the randomized offer or encouragement to take a program: it is equal to 1 when i is randomly invited or encouraged to take the treatment and zero otherwise. D_i denotes program participation: it takes value 1 when individual i decides to participate in the program and zero otherwise. D_i^r denotes the potential participation decision of individual i when $R_i = r$. We reserve the terms of treated and control to groups defined by R_i. We use the terms participants and non participants to refer to the groups defined by D_i. Y_i^d denotes the potential outcomes of individual i when $D_i = d$, $d \in \{0,1\}$. We assume that the usual SUTVA condition of absence of interactions among units holds.

In experiments with imperfect compliance, the actual program intake D_i differs from the

initial randomized treatment allocation R_i. For example, individuals in the treatment arm can choose not to take the program ($\Pr(D_i = 0|R_i = 1) > 0$). Among RCTs with imperfect compliance, it is useful to make a distinction between eligibility and encouragement designs. In an eligibility design, individuals allocated to the control group cannot participate in the program ($\Pr(D_i = 1|R_i = 0) = 0$). In an encouragement design, individuals allocated to the control group can still participate in the program ($\Pr(D_i = 1|R_i = 0) > 0$).

When compliance is imperfect, individuals can be split into four types. Compliers participate in the program when placed in the treatment arm ($D_i^1 = 1$) but do not participate when placed in the control arm ($D_i^0 = 0$). A shorthand notation for the compliers is $D_i^1 - D_i^0 = 1$. Always takers participate in the program both when placed in the treatment and control arms ($D_i^1 = D_i^0 = 1$). Never takers do not participate in the program when placed both in the treatment and control arms ($D_i^1 = D_i^0 = 0$). Defiers participate in the program when placed in the control arm, but do not when placed in the treatment arm ($D_i^1 - D_i^0 = -1$). In the remainder of the paper, we are going to assume that defiers do not exist.

Our approach consists in comparing the experimental estimate of the effect of program participation using R_i as an instrument for D_i to an observational estimate of the same effect that relies on unconfoundedness. In eligibility designs, under mild conditions, the experimental estimate is the effect of the program on the participant (TT): $\mathbb{E}[Y_i^1 - Y_i^0|D_i = 1]$. In encouragement designs, under mild conditions, the experimental estimate is the effect of the program on compliers, a.k.a. the Local Average Treatment Effect (LATE): $\mathbb{E}[Y_i^1 - Y_i^0|D_i^1 - D_i^0 = 1]$.

Key to our approach is the choice of an observational estimator that recovers either TT or LATE under the CIA. In the eligibility design, classical observational estimators such as matching applied on the treatment arm recover the TT under unconfoundedness. In the encouragement design, we show that, when unconfoundedness holds, a Wald-type weighted combination of the observational estimators on each treatment arm is equal to the LATE.

Finally, it is important to determine whether or not the bias of observational methods is due to unobserved covariates or to a failure of finding participants and non participants with the same values of the covariates. Following Heckman et al. (1998a), we say that this second part of the bias is due to a failure to have common support. We propose a definition of the bias due to the common support distinct from Heckman et al. (1998a),

though, since their definition encompasses two terms, only one of them being relevant for the bias of observational methods.

In the remainder of the section, we formally present our approach, first for the eligibility design, and then for the encouragement design.

2.1 Eligibility designs

They first key part of our approach is to use the treated group to form the observational estimator. The observational (or non-experimental) estimator NE_s can be built in the treated group by comparing the outcomes of participants and non participants conditional on a vector of observed characteristics X_i:

$$\begin{aligned} NE_s = \mathbb{E}[Y_i|D_i = 1, R_i = 1, S_i = 1] \\ - \mathbb{E}[\mathbb{E}[Y_i|X_i, D_i = 0, R_i = 1, S_i = 1]|D_i = 1, R_i = 1, S_i = 1]. \end{aligned} \tag{1}$$

S_i is the common support indicator: $S_i = \mathbb{1}[X_i : \Pr(D_i = 1|X_i, R_i = 1) < 1]$. NE_s is only well-defined on the common support, *i.e.* when there exists non-participants with the same values of X_i as the participants. The bias of the observational estimator is: $B_{ne} = NE_s - TT$. Under standard conditions summarized in Assumption 1 (X_i measures all the relevant confounders and there is no common support problem), the observational estimator recovers the treatment parameter of interest and B_{ne} is equal to zero.

Assumption 1 (Valid Non-Experimental Estimator)

(a) Unconfoundedness: $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i|X_i$

(b) Common Support: $0 < \Pr(D_i = 1|X_i) < 1$.

We are not imposing Assumption 1 in our analysis but we are testing its validity or measuring the extent to which it is violated.

The second key part of our approach is to use an experimental estimator to measure TT. The experimental estimator that we use for eligibility designs is computed by combining data from both treatment arms using a Bloom estimator:

$$E = \frac{\mathbb{E}[Y_i|R_i = 1] - \mathbb{E}[Y_i|R_i = 0]}{\Pr(D_i = 1|R_i = 1)}. \tag{2}$$

Under standard conditions summarized in Assumption 2 (Independence of assignment from potential outcomes, no direct effect of assignment on outcomes, existence of some

participants in the treatment group), the experimental estimator recovers the treatment effect of interest: $E = TT$. Our estimator of the bias of the observational method is $NE_s - E$, which is equal to B_{ne} under Assumption 2.

Assumption 2 (Valid Eligibility Design)

(i) Independence: $(Y_i^1, Y_i^0) \perp\!\!\!\perp R_i$

(ii) Exclusion restriction: $Y_i^{dr} = Y_i^d$, $d, r \in \{0, 1\}$

(iii) First Stage: $\Pr(D_i = 1 | R_i = 1) > 0$.

The validity of our method essentially rests on SUTVA and the conditions summarized in Assumption 2. Two of these conditions are key: SUTVA and Exclusion restriction. Indeed, Independence is guaranteed if randomization has been well-conducted and the sample size is large enough and First Stage can be directly tested by observing the proportion of participants in the treated group. We try to use only studies where SUTVA is likely to hold or has been tested and found to hold. We also use as our program participation variable the one that is the most proximate possible to the eligibility offer, so as to ensure that none of our non-participants has been exposed to any form of treatment. For example, when a training program covers several lectures, we define participation as coming to at least one lecture, not completing all lectures in the program.

We decompose our measure of the bias of the observational method into two separate components: B_u, that is due to unobserved confounders on the common support and B_{tts} that is due to a failure of the common support assumption:

$$B_{ne} = B_u + B_{tts}. \tag{3}$$

In order to form B_u and B_{tts}, we define E_s as the experimental estimator on the common support:

$$E_s = \frac{\mathbb{E}[Y_i | R_i = 1, S_i = 1] - \mathbb{E}[Y_i | R_i = 0, S_i = 1]}{\Pr(D_i = 1 | R_i = 1, S_i = 1)}. \tag{4}$$

We then have $B_u = NE_s - E_s$ and $B_{tts} = E_s - E$. B_u measures the failure of the observational estimator to account for all the relevant confounders on the common support. B_{tts} measures the bias due to the fact that the treatment effect estimated on the common support is not representative of the treatment effect outside of the common support. B_u might decrease as we condition for more covariates but B_{tts} cannot. Whether the bias of

the observational method stems mainly from B_u or B_{tts} determines whether there is some scope for improving the observational methods by observing more covariates.
Finally, as a presentation device, we contrast the bias of the observational method to the bias of a rough comparison between participants and non participants without adjusting for any covariates, the with/without comparison $WW = \mathbb{E}[Y_i|D_i = 1, R_i = 1] - \mathbb{E}[Y_i|D_i = 0, R_i = 1]$. The bias of the with/without comparison is $B = WW - TT$. B is generally called selection bias. We propose a four-way decomposition of B:

$$B = B_{ys} + B_x + \underbrace{B_u + B_{tts}}_{B_{ne}}, \tag{5}$$

where $B_{ys} = WW - WW_s$ and $B_x = WW_s - NE_s$ and WW_s is the with/without estimator computed on the common support ($WW_s = \mathbb{E}[Y_i|D_i = 1, R_i = 1, S_i = 1] - \mathbb{E}[Y_i|D_i = 0, R_i = 1, S_i = 1]$). Our decomposition complements the three-way decomposition of Heckman et al. (1998a) by splitting their original bias term due to the common support into two subparts: a part that the observational method solves (B_{ys}) and a part that the observational method does not solve (B_{tts}).

2.2 Encouragement designs

In an encouragement design, the experimental estimator that we can form estimates a $LATE$ and not a TT under mild assumptions. Also, since individuals can decide to participate in the program even when they are in the control group, we can form two separate observational estimators, one in the treated group (NE_s^1) and one in the control group (NE_s^0), with:

$$\begin{aligned} NE_s^d = &\mathbb{E}[Y_i|D_i = 1, R_i = d, S_i = 1] \\ &- \mathbb{E}[\mathbb{E}[Y_i|X_i, D_i = 0, R_i = d, S_i = 1]|D_i = 1, R_i = d, S_i = 1]. \end{aligned} \tag{6}$$

Our main contribution for encouragement designs is to show that a Wald-like combination of the observational estimators from each treatment arm recovers the $LATE$ under standard conditional independence conditions. By comparing the observationally-estimated $LATE$ and the experimentally-estimated $LATE$, we can infer the bias of the observational method for the $LATE$. It might seem frustrating not to be able to recover the bias of the observational method for the TT parameter, but the experimentally-induced variation in an encouragement design only reveals the causal effect for compliers, not for

the all group of participants.
The experimental estimator that we use in an encouragement design is the Wald estimator:

$$E = \frac{\mathbb{E}[Y_i|R_i = 1] - \mathbb{E}[Y_i|R_i = 0]}{\Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0)}. \tag{7}$$

Under standard conditions stated in Assumption 3, E estimates a $LATE$. More formally, $LATE \equiv \mathbb{E}[Y_i^1 - Y_i^0|D_i^1 - D_i^0 = 1]$.

Assumption 3 (Valid Encouragement Design) *(a) Independence:* $(Y_i^1, Y_i^0, D_i^1, D_i^0) \perp\!\!\!\perp R_i$

(b) Exclusion restriction: $Y_i^{dr} = Y_i^d$, $d, r \in \{0, 1\}$

(c) First Stage: $\Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0) > 0$

(d) Monotonicity: $\Pr(D_i^1 - D_i^0 = -1) = 0$.

Compared to the eligiblity design, the main additional condition that we are going to maintain in our analysis is Monotonicity. In general, in our applications, Monotonicity seems a natural assumption.
Our observational – or non-experimental – estimator for encouragement designs is:

$$NE_s = \frac{NE_s^1 \Pr(D_i = 1|R_i = 1, S_i = 1) - NE_s^0 \Pr(D_i = 1|R_i = 0, S_i = 1)}{\Pr(D_i = 1|R_i = 1, S_i = 1) - \Pr(D_i = 1|R_i = 0, S_i = 1)}. \tag{8}$$

We use the same name, NE_s, that we have used for the observational estimator in the eligibility design. We tolerate this abuse of notation for two reasons. First, with this notation, the definition of the different bias terms in the decomposition of B is the same in both encouragement and eligibility designs. Second, the observational estimator in the eligibility design is a particular case of the observational estimator that we have defined in the encouragement design where $\Pr(D_i = 1|R_i = 0, S_i = 1) = 0$ and, by convention, $NE_s^0 = 0$ (whereas it is actually not well-defined in the eligibility design). We believe the simplicity of not having too many notations trumps the possible confusion from having the same quantities defined in different ways.
Under standard conditional independence conditions in both the treated and control groups as summarized in Assumption 4, our observational estimator estimates the LATE. This result is summarized in Theorem 1.

Assumption 4 (Valid Non-Experimental Estimator) *For* $d \in \{0, 1\}$*:*

(a) Conditional Independence: $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i | X_i, R_i = d$

(b) Common Support: $0 < \Pr(D_i = 1 | X_i, R_i = d) < 1$

Theorem 1 *Under Assumptions 3 and 4,* $NE_s = E_s = E = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1]$*, and as a consequence,* $B_{ne} = B_u = B_{tts} = 0$*.*

Proof: Classical results imply that the Wald estimators E and E_s recover $\mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1]$ and $\mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1]$ respectively under Assumption 3. Adding Common Support, we have $E = E_s$. Under Assumption 4, classical results show that $NE_s^d = \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = d] = \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, R_i = d]$. What remains to be shown is that $NE_s = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1]$ under Assumptions 3 and 4:

$$\begin{aligned} NE_s^1 &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = 1] \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i^1 = D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) \\ &\quad + \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1), \end{aligned}$$

where the second equality comes from Independence and Monotonicity. Now:

$$\begin{aligned} \Pr(D_i^1 = D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) &= \frac{\Pr(D_i^1 = D_i^0 = 1 \wedge D_i = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\ &= \frac{\Pr(D_i^1 = D_i^0 = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\ &= \frac{\Pr(D_i = 1 | S_i = 1, R_i = 0)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)}, \end{aligned}$$

where the first equality comes from Bayes rule, the second equality from the fact that $D_i^1 = D_i^0 = 1$ imply $D_i = 1$ and the third equality from Monotonicity and First Stage. Using the same approach, we have:

$$\begin{aligned} \Pr(D_i^1 - D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) &= \frac{\Pr(D_i^1 - D_i^0 = 1 \wedge D_i = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\ &= \frac{\Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)}, \end{aligned}$$

where the first equality uses Bayes rule and the second equality uses the fact that $D_i^1 - D_i^0 = 1$ implies $D_i = 1$ when $R_i = 1$. Under Monotonicity and Conditional Independence,

we also have:

$$\begin{aligned} NE_s^0 &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = 0] \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1]. \end{aligned}$$

Combining the formulas for NE_s^1 and NE_s^0, the numerator of the NE_s estimator in equation 8 is:

$$\begin{aligned} & NE_s^1 \Pr(D_i = 1 | R_i = 1, S_i = 1) - NE_s^0 \Pr(D_i = 1 | R_i = 0, S_i = 1) \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1) \\ &\quad + \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i = 1 | S_i = 1, R_i = 0) \\ &\quad - \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i = 1 | S_i = 1, R_i = 0) \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1). \end{aligned}$$

Finally, Monotonicity and First Stage imply that:

$$\Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1) = \Pr(D_i = 1 | S_i = 1, R_i = 1) - \Pr(D_i = 1 | S_i = 1, R_i = 0),$$

which proves the result. ■

The intuition behind Theorem 1 can be summarized as follows: if all confounders are observed, the observational estimator applied to the treated group estimates a weighted average of the effect of the treatment on the always takers and on the compliers whereas the observational estimator applied to the control group estimates the treatment effect on the always takers. When taking an appropriately reweighted difference between these two estimates, we recover the effect on the compliers, a.k.a. the $LATE$.

Our proposed measure of the bias of the observational estimator in an encouragement design is thus $B_{ne} = NE_s - E$, where NE_s is a Wald-like ratio of the observational estimators computed on each treatment arm and E is the Wald estimator. In order to compute a similar decomposition of selection bias as the one we derive for eligibility designs, we define E_s to be the Wald estimator on the common support, WW^1 and WW^0 to be the with/without comparisons on each treatment arm, WW to be their Wald-like ratio and WW_s to be the Wald-like ratio of WW_s^1 and WW_s^0, the with/without comparisons taken on the common support.

3 Estimation

3.1 Eligibility designs

The estimators of the theoretical quantities are:

$$\hat{WW} = \frac{1}{\sum_{i=1}^{N} D_i R_i} \sum_{i=1}^{N} D_i R_i Y_i - \frac{1}{\sum_{i=1}^{N} (1-D_i) R_i} \sum_{i=1}^{N} (1-D_i) R_i Y_i$$

$$\hat{WW}_s = \frac{1}{\sum_{i=1}^{N} D_i R_i \hat{S}_i} \sum_{i=1}^{N} D_i R_i \hat{S}_i Y_i - \frac{1}{\sum_{i=1}^{N} (1-D_i) R_i \hat{S}_i} \sum_{i=1}^{N} (1-D_i) R_i \hat{S}_i Y_i$$

$$\hat{NE}_s = \frac{1}{\sum_{i=1}^{N} D_i R_i \hat{S}_i} \sum_{i=1}^{N} D_i R_i \hat{S}_i (Y_i - \hat{\mathbb{E}}[Y_i | D_i = 0, \hat{P}_i, \hat{S}_i = 1])$$

$$\hat{E} = \frac{\frac{1}{\sum_{i=1}^{N} R_i} \sum_{i=1}^{N} R_i Y_i - \frac{1}{\sum_{i=1}^{N} (1-R_i)} \sum_{i=1}^{N} (1-R_i) Y_i}{\frac{\sum_{i=1}^{N} D_i R_i}{\sum_{i=1}^{N} R_i}}$$

$$\hat{E}_s = \frac{\frac{1}{\sum_{i=1}^{N} R_i \hat{S}_i} \sum_{i=1}^{N} R_i \hat{S}_i Y_i - \frac{1}{\sum_{i=1}^{N} (1-R_i) \hat{S}_i} \sum_{i=1}^{N} (1-R_i) \hat{S}_i Y_i}{\frac{\sum_{i=1}^{N} D_i R_i \hat{S}_i}{\sum_{i=1}^{N} R_i \hat{S}_i}}$$

$$\hat{S}_i = \mathbb{1}[P_i : \hat{f}_1(\hat{P}_i) > c_q \text{ and } \hat{f}_0(\hat{P}_i) > c_q]$$

where $\hat{P}_i$ is the propensity score;[3] $\hat{\mathbb{E}}[Y_i | D_i = 0, \hat{P}_i]$ is the local quadratic regression estimator of $\mathbb{E}[Y_i | D_i = 0, P_i]$;[4] $\hat{f}_d(\hat{P}_i)$ is the density of the distribution of $\hat{P}_i$ conditional on $D_i = d$;[5] c_q is the threshold density level under which an observation is trimmed;[6] q is the trimming level.

In practice, we estimate $\hat{E}$ and $\hat{E}^s$ using a 2SLS regression of Y_i on D_i with the covariates X_i entering linearly using R_i and X_i as instruments.

[3] $\hat{P}_i$ is the predicted value of a probit or logit regression of D_i on X_i conditional on $R_i = 1$.

[4] $\hat{\mathbb{E}}[Y_i | D_i = 0, \hat{P}_i]$ is estimated using a weighted least squares regression of Y_j on $(\hat{P}_i - \hat{P}_j)$ and $(\hat{P}_i - \mathbf{P}_j)^2$ on the untreated sample with $R_j = 1$. The bandwidth is chosen using cross validation. We use a quartic kernel.

[5] $\hat{f}_d(\hat{P}_i)$ is estimated using a kernel density estimator with bandwidth chosen using Silverman's rule of thumb.

[6] c_q is the q^{th} quantile of the distribution of $\hat{f}_1(\hat{P}_i)$ and $\hat{f}_0(\hat{P}_i)$.

3.2 Encouragement designs

The estimation of $\hat{WW}$ and $\hat{WW}_s$ is the same as in the eligibility design. The other quantities are estimated as follows:

$$\hat{NE}_s^d = \frac{1}{\sum_{i=1}^N \mathbb{1}[R_i = d] D_i \hat{S}_i} \sum_{i=1}^N \mathbb{1}[R_i = d] D_i \hat{S}_i (Y_i - \hat{\mathbb{E}}[Y_i | D_i = 0, \hat{P}_i^d, \hat{S}_i = 1, R_i = d])$$

$$\hat{NE}_s = \frac{\hat{NE}_s^1 \frac{\sum_{i=1}^N D_i R_i \hat{S}_i}{\sum_{i=1}^N R_i \hat{S}_i} - \hat{NE}_s^0 \frac{\sum_{i=1}^N D_i (1-R_i) \hat{S}_i}{\sum_{i=1}^N (1-R_i) \hat{S}_i}}{\frac{\sum_{i=1}^N D_i R_i \hat{S}_i}{\sum_{i=1}^N R_i \hat{S}_i} - \frac{\sum_{i=1}^N D_i (1-R_i) \hat{S}_i}{\sum_{i=1}^N (1-R_i) \hat{S}_i}}$$

$$\hat{E} = \frac{\frac{1}{\sum_{i=1}^N R_i} \sum_{i=1}^N R_i Y_i - \frac{1}{\sum_{i=1}^N (1-R_i)} \sum_{i=1}^N (1-R_i) Y_i}{\frac{\sum_{i=1}^N D_i R_i}{\sum_{i=1}^N R_i} - \frac{\sum_{i=1}^N D_i (1-R_i)}{\sum_{i=1}^N (1-R_i)}}$$

$$\hat{E}_s = \frac{\frac{1}{\sum_{i=1}^N R_i \hat{S}_i} \sum_{i=1}^N R_i \hat{S}_i Y_i - \frac{1}{\sum_{i=1}^N (1-R_i) \hat{S}_i} \sum_{i=1}^N (1-R_i) \hat{S}_i Y_i}{\frac{\sum_{i=1}^N D_i R_i \hat{S}_i}{\sum_{i=1}^N R_i \hat{S}_i} - \frac{\sum_{i=1}^N D_i (1-R_i) \hat{S}_i}{\sum_{i=1}^N (1-R_i) \hat{S}_i}}$$

$$\hat{S}_i = \mathbb{1}[\forall d \in \{0,1\}, \hat{P}_i^d : \hat{f}_1^d(\hat{P}_i^d) > c_q^d \text{ and } \hat{f}_0^d(\hat{P}_i^d) > c_q^d]$$

$\hat{P}_i^d$ is the propensity score on the arm with $R_i = d$. $\hat{P}_i^d$ is the predicted value of a probit regression of D_i on the observed covariates X_i. $\hat{S}_i$ is the intersection of the common supports on both treatment arms. It is computed as follows:

1. The predicted propensity score in each treatment arm is computed for all observations, whatever their treatment arm.

2. The densities of the propensity score conditional on $D_i = 1$ and $D_i = 0$ are computed for all points in the sample, using only the treated observations in each treatment arm separately as sources of observations for computing the densities (the untreated and the points in the other treatment arms act only as grid points, not as observation points).

3. The density cutoff level in terms of propensity score is computed separately on each treatment arm using only the treated observations on this arm.

4. The common support is computed for all the observations in the sample that have a density of the propensity score above the threshold.

5. The intersection of the common supports is simply made of all the observations that belong to the common support on both arms.

3.3 Covariate selection

We use two approaches for selecting the covariates that enter in the estimation of the propensity score:

Linear We simply enter the covariates linearly.

LassoYD We use a double post lasso estimator, using as covariates the ones selected by lasso among linear, quadratic and interaction terms both in the probit regression and in a regression of Y_i on the covariates. This approach is advised by Belloni et al. (2014) as more efficient at selecting the right set of covariates.

4 Data

We use data from 6 published papers using RCTs with imperfect compliance and with data made public online Angrist et al. (2002, 2009); Behaghel et al. (2014); Burde and Linden (2013); Drexler et al. (2014); Oster and Thornton (2011). We describe the setting of each paper and the available data in what follows.

Behaghel data

The Behaghel et al. (2014) paper analyses the effect of three different counseling programs for job seekers in four french administrative regions on employment outcomes. There are three randomization arms (assignment from january 2007 - december 2007): job seekers are either assigned to a public intense counseling program (CVE), a private intense counseling program (OPP) or to the control group which corresponds to the standard public counseling program. The assignment probabilities vary locally and over time horizons. Hence, the original paper weighs the observations to correct for this. The outcomes are assessed at different time horizons: after 3, 6, 9 and 12 months. Control variables are baseline characteristics and include rich information on education, age, skill level, gender, marital status, nationality, reason for unemployment, experience and wage target (see table 2 in the original paper). The original experiment is conceived as an eligibility design. Individuals were offered the possibility to take the offered new treatment in each treatment arm, but could decide to stock with the standard control treatment. Actually, only 32% of the individuals assigned to the public arm eventually entered the public treatment and 43% of the individuals assigned to the private treatment eventually entered the private treatment (Table 4 in the paper). Unfortunately, there were crossovers and the eligibility design was violated. Individuals mainly switched from the public and

standard arms to the private one, seemingly because of pressures from private providers to have more applicants. The scope of these violations remains very limited though. Only 2.8% (resp. 1.5%) of those assigned to the standard (resp. public) track ended up with a private provider. Only 0.4% (resp. 0.5%) of those assigned to the standard (resp. private) track ended up in the public program. The authors consider these crossovers to be small enough so that they can neglect them and still interpret their estimates as TT parameters. Another way would have been to derive bounds on the TT. These crossovers pose a severe problem to our approach. If we consider this program as a true encouragement design, then we have a very low number of treated for the two treatments in the control arm and a very imprecise estimator, maybe even infeasible. We choose to consider this experiment as a true eligibility design and thus assign all the members of the control arm to the no-treatment condition. If anything, this approach slightly favors the observational estimator.

Drexler data

The Drexler et al. (2014) paper has two treatment arms: standard accounting (ACC) and rule of thumb (ROT). This is an eligibility design, with a control arm where no one receives the offer of the two treatments. The original paper is an intention to treat analysis, but acknowledges huge self-selection into the treatments (Table A2 in the paper). We choose to measure treatment uptake by attending at least one class. Treatment uptake is 45.8% in the ACC arm and 43.1% in the ROT arm (Table A3 in the paper). The main outcomes are a business practices index (Practices), the existence of any reporting errors (Mistakes) and Sales. Control variables are baseline characteristics such as business type, loan size, participation in a savings accounts and pre-treatment outcomes when available.

Burde data

The Burde and Linden (2013) paper estimates the impact of randomly building schools in Afghan villages on children's enrollment and test scores. We use the second part of the experiment as an encouragement design: assuming enrollment is the only way through which schools impact test scores, we look at the effect of enrollment on test scores. One way this assumption might be false is if some children substitute the low-quality village school for a high quality but distant government school. It is in principle testable, if we have the type of school in which children were enrolled before the experiment. The LATE estimation is actually not at the center of the article but the result is reported p.36 ("If we assume that the change in test scores is due only to additional enrollment, we estimate that, for those girls caused to enroll in a formal school, test scores increased by 1.28 standard deviations in the fall."). The procedure is detailed in footnote 21, as well

as instances in which the exclusion restriction would be violated. It is also mentioned as an important result in the introduction (p.28: "Test scores for those girls caused to attend school as a result of the experiment increase by 1.28 standard deviations due to enrollment.") Outcomes are thus test scores in the Fall of 2007 and Spring 2008. The encouragement is the assignment of a school in one's village. The treatment is actual school enrollment in 2007 or in 2008. Control variables are the child's age, gender and whether she is the child of the household head. Household-level controls are the distance to the nearest formal school, the number of years the family has lived in the village, ethnicity, the number of people in the household, whether the household farms, the size of the land and the number of sheep that the household owns and the age and years of education of the household head.

Oster data

In the Oster and Thornton (2011), there is only one treatment: offering a menstrual cup to a girl. This is an eligibility design, where control girls are not offered the cup. The original paper is mostly an intention to treat analysis, but includes also a 2SLS analysis in the online appendix. We measure treatment uptake by usage of the cup. Usage of the cup is 60% after 6 months (p.97 in the paper; or see their 2009 NBER paper for more details). Outcomes examined in the paper are (see Table 2 in the paper and Tables 3 and 6 in the Online Appendix): Attendance, measured by official levels, time use diary and random on-site visits; Time spent in school and time spent doing various chores (including laundry); Test scores. Control variables are age, grade, mother's education, work for pay, father's Hindu ethnicity, menses at baseline, baseline exam score and income brackets. One specificity of the paper is its heavy reliance on individual fixed effects, in order to capture the "usual" attendance rate. Our observational method does not allow for the inclusion of individual fixed effects. In this analysis, we use the average attendance rate as outcome. Because we do not observe periods before the treatment, pre-treatment attendance is built using all days, not only period days.

Angrist31 data

There are three treatment arms in the Angrist et al. (2009) paper: Student Support Program (SSP), which offers access to peer-advising and Facilitated Study Groups; Student Fellowship Program (SFP) that offers merit scholarships for students whose grade improve over the year; and both SSP and SFP (SFSP). This is an eligibility design, with a control arm where no one receives the offer of the three treatments. The original paper is mostly an intention to treat analysis, but includes also a 2SLS analysis, especially for the sample of women (Table 8 in the paper). We measure treatment uptake by signing

up for the program. Treatment uptake is 52% in the SSP arm, 86% in the ROT arm and 76% in the SFSP arm (Table 3 in the paper). Outcomes examined in the paper are (see Table 2 in the paper): Fall grades, but only for the subset of students that have taken a full course in the Fall (1,255) (grade.20059.fall); First and second year GPA (both for all students that persist up to the end of the first year (1,399) and for the ones that have taken a full fall class) (GPA.year1 and GPA.year2); On probation or withdrew in both years (prob.year1 and prob.year2); Good standing in both years (goodstanding.year1 and goodstanding.year2); Credits earned in both years (credits.earned1 and credits.earned2). Control variables are sex, mother tongue, high school grade, number of credits enrolled, parents' education and measures of academic motivation and procrastination.

Angrist66 data

In the Angrist et al. (2002) paper, the authors study the consequences of a Colombian program offering vouchers covering the costs of private secondary schools (the PACEs program) on the educational outcomes of students. The authors use the randomized allocation of vouchers as an instrumental variable for actual voucher uptake. Indeed, only 90% of the students offered a voucher actually redeemed it while around 24% of the students in the control group secured a grant by other means. In our application, receiving a scholarship is the treatment, the randomized treatment is the voucher offer and outcomes are: Highest grade completed at the time of the survey (scyfnsh); In school at the time of the survey (inschl); Number of repetitions at the time of the survey (nrept); Finished 8th grade (at the time of the survey) (finish8); Married or living with companion at the time of the survey (married). The control variables are age, sex, strata of residence, Barrio of residence, month of interview, interview in person, interview using new survey, father's and mother's age, highest grade and whether they earn the minimum wage.

5 Results

5.1 Results for the individual studies

Behaghel data The results from the Behaghel dataset are shown in Figures 1a and 1b. For the CVE treatment, we find a large negative selection bias for employment 3 months after the treatment, and zero selection bias thereafter. For the OPP treatment, we find a negative selection bias at 3 months and a positive one for later outcomes. Both for CVE and OPP, observational methods do not decrease selection bias.

Drexler data The results for the Drexler dataset are presented in Figures 1c and 1d. In

the ACC treatment, selection bias is close to zero, as are the biases of the observational estimators. In the ROT treatment, selection bias is negative for practices and sales and positive for mistakes. Observational methods fail to undo much of this bias.

Burde data Figures 1e and 1f show our results from the Burde dataset. We see positive selection bias for test scores measured in 2007. Both observational methods seem to devrease selection bias, although they both provide rather imprecise estimates.

Oster data Figures 2a and 2b show the results from the Oster dataset. We do not see selection bias except for the Laundry-Period outcome. Observational methods do not correct for this bias. Observational methods are also strikingly imprecise, especially the lasso one.

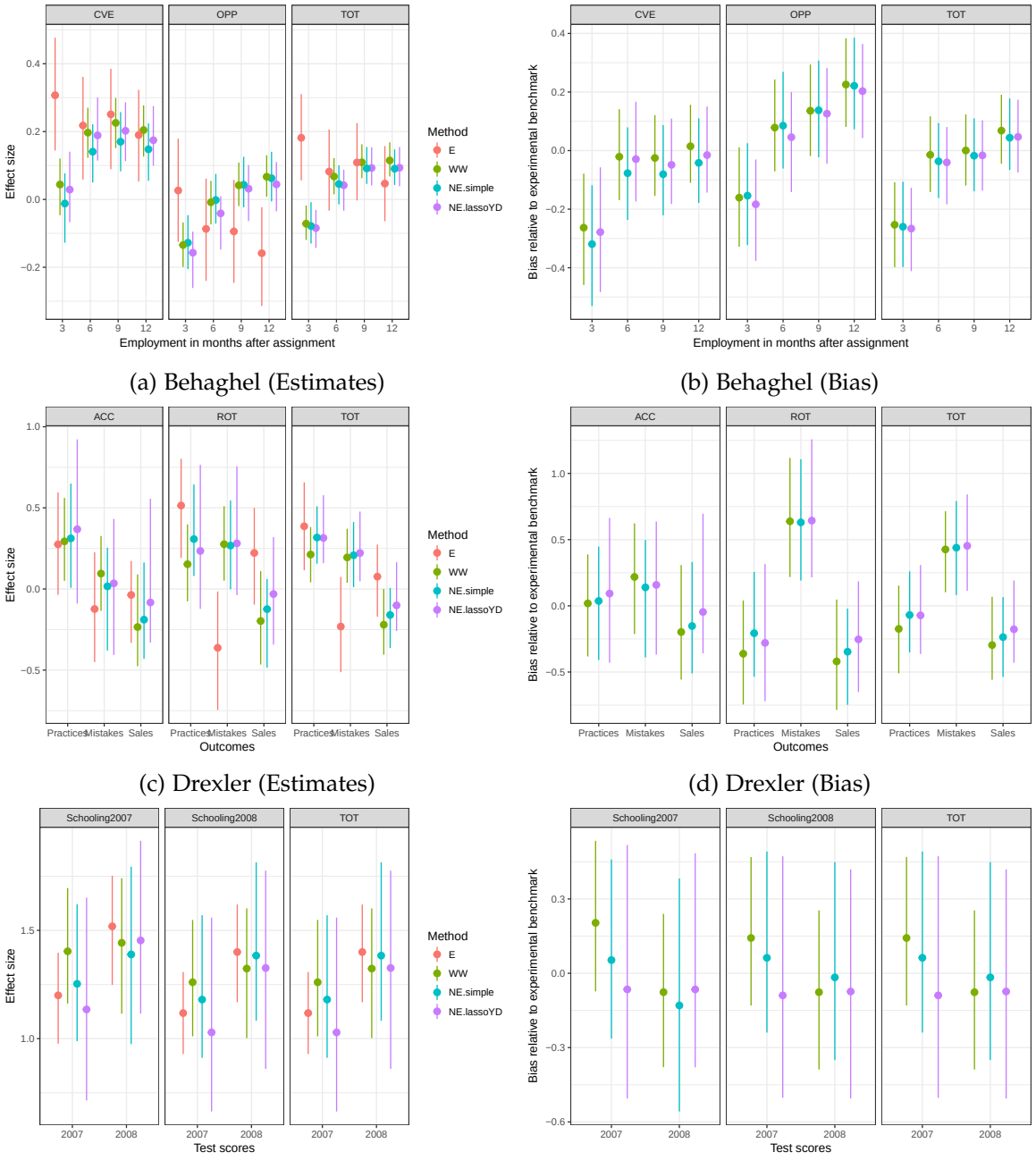
Angrist31 data The results for the Angrist31 dataset are shown in Figures 2c, 2d, 2e and 2f. For the SFP treatment, we see positive selection bias for the credits.earned.1 and prob.year.1 outcomes. For the SFSP treatment, we see positive selection bias for credits.earned.2 and negative selection bias for grades.20059.fall. Observational methods do not decrease these biases. For the SSP and SP treatments, we see positive selection bias for credits.earned.1, credits.earned.2, goodstanding.year.1, goodstanding.year.2, grades.20059.fall and prob.year.1. Observational methods undo the bias for the first four outcomes but do not affect the bias of the last two.

Angrist66 data The results for the Angrist66 dataset are presented in Figures 2g and 2h. Selection bias is positive for the finish8 and inschl outcomes. Observational estimates seem to help a little to decrease this bias.

5.2 Overall results

When combining the estimates of the bias of observational methods for all of the outcomes and treatments in our 6 datasets, we can start to have a broad idea of the performances of observational methods.

Figure 3a shows the distribution of the absolute value of selection bias. It shows that the distribution of the bias of the simple linear observational estimator is slightly more concentrated close to zero than selection bias whereas the distribution of the bias of the lassoYD estimator exhibits a clear spike close to zero. Unfortunately, both observational methods also exhibit a number of large biases similar to that of the naive with/without estimator. It thus seems that observational methods and the lassoYD approach in particular are good at decreasing small selection bias but do not seem to be able to undo selection bias when it is large.



(a) Behaghel (Estimates)

(b) Behaghel (Bias)

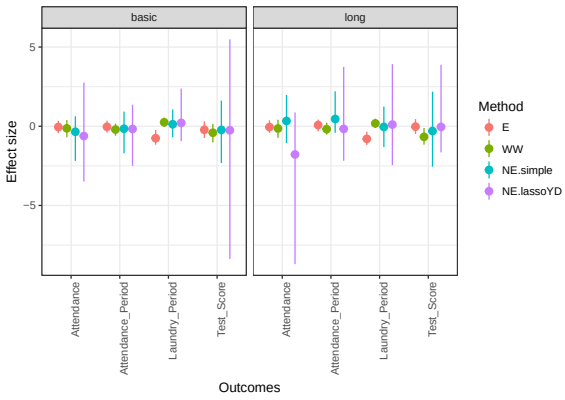
(c) Drexler (Estimates)

(d) Drexler (Bias)

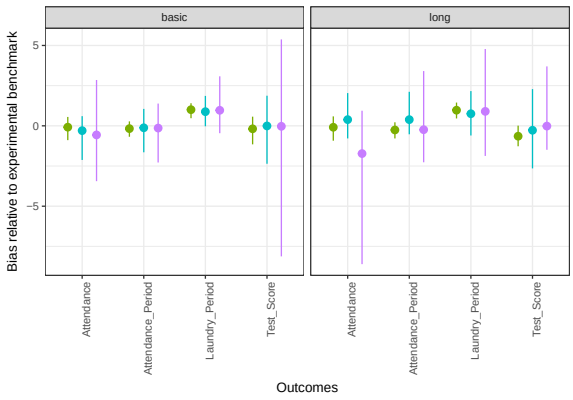
(e) Burde (Estimates)

(f) Burde (Bias)

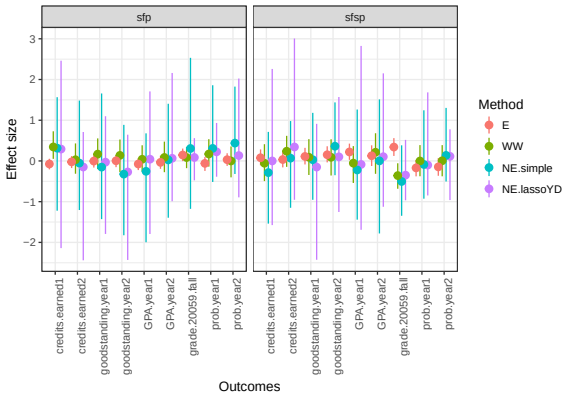
Figure 1: Experimental and non experimental estimates, confidence intervals estimated with 500 bootstrap simulations



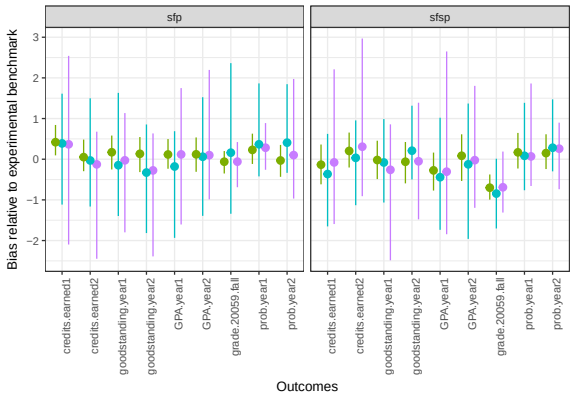
(a) Oster (Estimates)



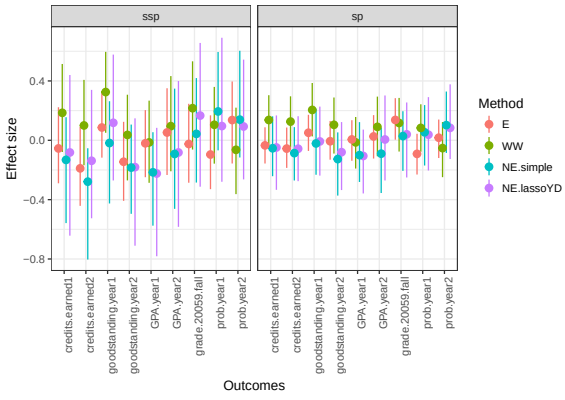
(b) Oster (Bias)



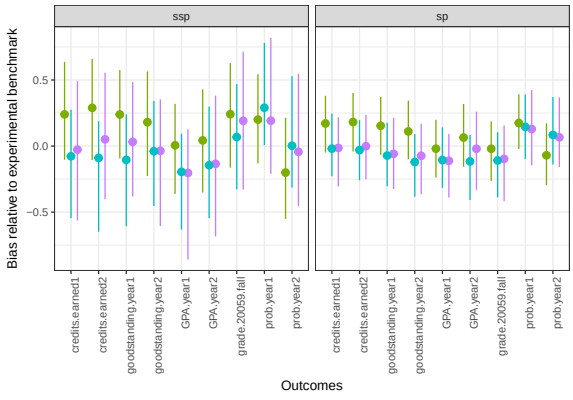
(c) Angrist31 (Estimates)



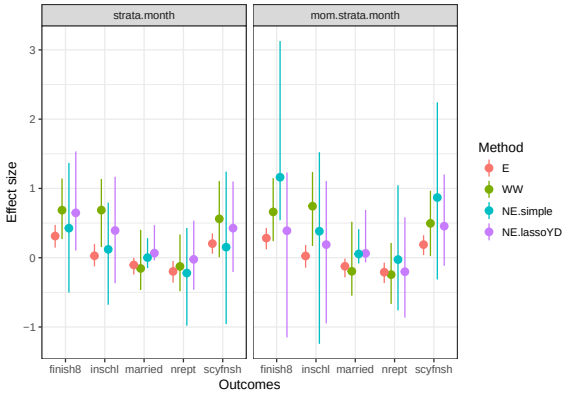
(d) Angrist31 (Bias)



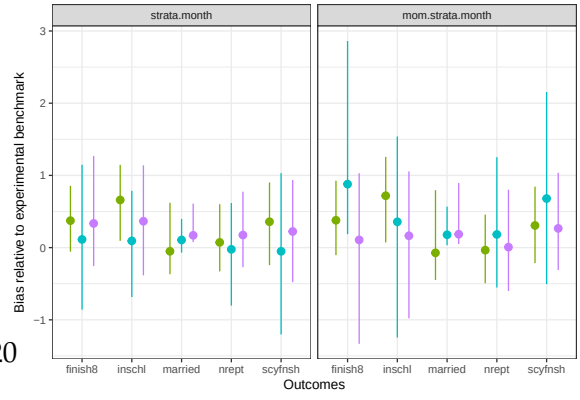
(e) Angrist31 (Estimates)



(f) Angrist31 (Bias)

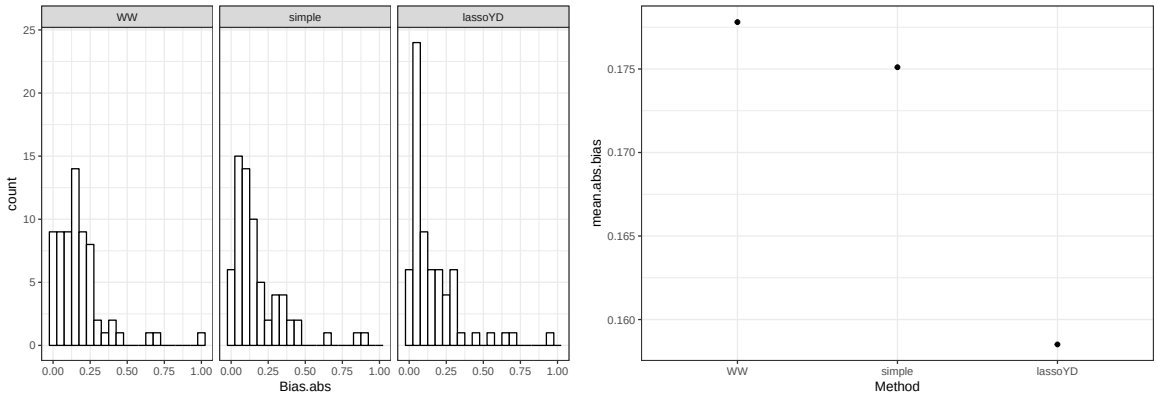


(g) Angrist66 (Estimates)



(h) Angrist66 (Bias)

Figure 2: Experimental and non experimental estimates, figure 1 continued



(a) Distribution (b) Mean absolute value of bias

Figure 3: Distribution and means of the absolute values of selection bias and of the bias of observational methods

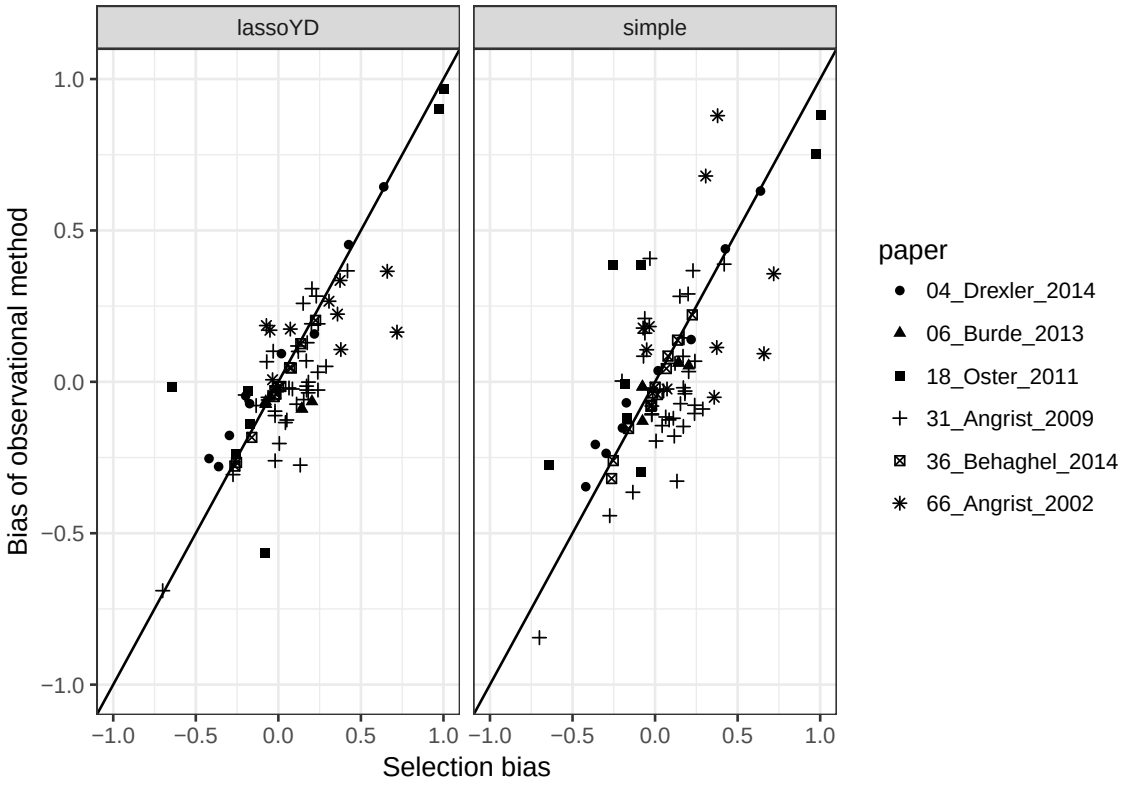
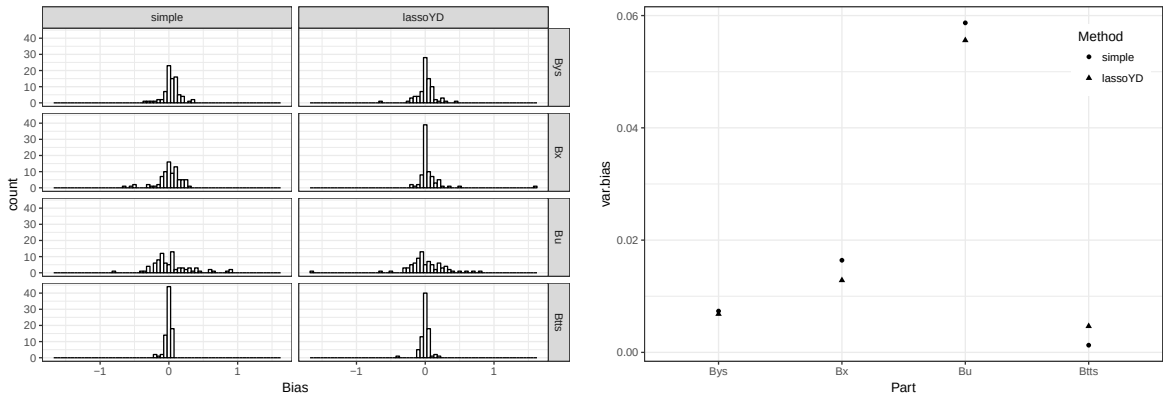


Figure 4: Performance of the non-experimental estimator



(a) Distribution (zoomed)

(b) Variance

Figure 5: Components of selection bias

Figure 3b confirms this result by showing that the mean absolute value of bias decreases from 18% of a standard deviation of outcomes for the naive with/without estimator, to 17.5% for the simple linear observational method and to 16% for the lassoYD observational method, still a very large value.

Figure 4 shows the correlation between selection bias and the bias of observational methods. The first striking result is that most observations lie on the 45 degree line, indicating that observational methods fail to capture selection bias. With perfect observational methods, observations would lie on the horizontal 0 line, which is not really the case.

Finally, Figure 5 shows the results of the decomposition of selection bias in four components. Figure 5a suggests that the component due to unobserved confounders is the most important one for explaining the bias of observational methods. Figure 5b confirms that this is the case by showing that the variance of this component largely dominates that of the common support component.

6 Conclusion

While this paper focuses on 6 RCTs, we are hopeful to propose a framework in which future datasets could easily be added to enrich the analysis and to investigate which observed confounders matter most for a given type of programme and outcome. Equipped with such knowledge, researchers could decide which observational method to use, or alternatively decide to run an RCT, decide to collect critical covariates, correct the results of observational methods ex-post and have a more informed assessment of the results stemming from observational methods.

References

Agodini, R. and M. Dynarski (2004): "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *The Review of Economics and Statistics*, 86, 180–194.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, 92, 1535–1558.

Angrist, J., D. Lang, and P. Oreopoulos (2009): "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 1, 136–63.

Arceneaux, K., A. S. Gerber, and D. P. Green (2006): "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis*, 14, 37 – 62.

Behaghel, L., B. Crépon, and M. Gurgand (2014): "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment," *American Economic Journal: Applied Economics*, 6, 142–74.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608.

Benson, K. and A. J. Hartz (2000): "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine*, 342, 1878–1886, pMID: 10861324.

Bléhaut, M. and R. Rathelot (2014): "Expérimentation contrôlée contre appariement : le cas dun dispositif daccompagnement de jeunes diplômés demandeurs demploi," *Economie & Prévision*, 204-205, 163–181.

Burde, D. and L. L. Linden (2013): "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools," *American Economic Journal: Applied Economics*, 5, 27–40.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, 120, F452–F477.

——— (2015): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations," Working Paper 21431, National Bureau of Economic Research.

Concato, J., N. Shah, and R. I. Horwitz (2000): "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *New England Journal of Medicine*, 342, 1887–1892, pMID: 10861325.

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," *The Review of Economics and Statistics*, 84, 151–161.

Drexler, A., G. Fischer, and A. Schoar (2014): "Keeping It Simple: Financial Literacy and Rules of Thumb," *American Economic Journal: Applied Economics*, 6, 1–31.

Eckles, D. and E. Bakshy (2017): "Bias and high-dimensional adjustment in observational studies of peer effects," *ArXiv e-prints*.

Ferraro, P. J. and J. J. Miranda (2014): "The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark," *Journal of Economic Behavior & Organization*, 107, 344 – 365.

Fraker, T. and R. Maynard (1987): "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *The Journal of Human Resources*, 22, 194–227.

Friedlander, D. and P. K. Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *The American Economic Review*, 85, 923–937.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2017): "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *SSRN*.

Griffen, A. S. and P. E. Todd (2017): "Assessing the Performance of Nonexperimental Estimators for Evaluating Head Start," *Journal of Labor Economics*, 35, S7–S63.

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, J. A. Smith, and P. E. Todd (1998a): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1099.

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

Kunz, R., K. S. Khan, and H.-H. Neumayer (2000): "Observational Studies and Randomized Trials," *New England Journal of Medicine*, 343, 1194–1197, pMID: 11041757.

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

Oster, E. and R. Thornton (2011): "Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation," *American Economic Journal: Applied Economics*, 3, 91–100.

Pocock, S. J. and D. R. Elbourne (2000): "Randomized Trials or Observational Tribulations?" *New England Journal of Medicine*, 342, 1907–1909, pMID: 10861329.

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