# Using Experiments to Correct for Selection in Observational Studies[*]

Susan Athey[†] Raj Chetty[‡] Guido W. Imbens[§]

First Version, August 2019; Current version September 2020

**Abstract**

In the social sciences there has been an increase in interest in randomized experiments to estimate causal effects, partly because their internal validity tends to be high, but they are often small and contain information on only a few variables. At the same time, as part of the big data revolution, large, detailed, and representative, administrative data sets have become more widely available. However, the credibility of estimates of causal effects based on such data sets alone can be low. In this paper, we develop statistical methods for systematically combining experimental and observational data to improve the credibility of estimates of the causal effects. We focus on a setting with a binary treatment where we are interested in the effect on a primary outcome that we only observe in the observational sample. Both the observational and experimental samples contain data about a treatment, observable individual characteristics, and a secondary (often short term) outcome. To estimate the effect of a treatment on the primary outcome, while accounting for the potential confounding in the observational sample, we propose a method that makes use of estimates of the relationship between the treatment and the secondary outcome from the experimental sample. We interpret differences in the estimated causal effects on the secondary outcome between the two samples as evidence of unobserved confounders in the observational sample, and develop control function methods for using those differences to adjust the estimates of the treatment effects on the primary outcome. We illustrate these ideas by combining data on class size and third grade test scores from the Project STAR experiment with observational data on class size and both third and eighth grade test scores from the New York school system.

**Keywords: Causality, Experiments, Observational Studies, Long Term Outcomes**

[*] We are grateful for discussions with Nathan Kallus, Xiaojie Mao, Dylan Small, and for comments from seminar participants at the University of Pennsylvania and Berkeley. We also want to acknowledge excellent research assistance from Kevin Chen. This research was funded through the Sloan Foundation, Schmidt Futures, and ONR grant N00014-17-1-2131.

[†] Graduate School of Business, Stanford University, and NBER, athey@stanford.edu.

[‡] Department of Economics, Harvard University, and NBER, chetty@harvard.edu.

[§] Graduate School of Business, and Department of Economics, Stanford University, and NBER, imbens@stanford.edu.

# 1 Introduction

There has been an influential movement in empirical studies in economics towards relying more on experimental as opposed to observational data to estimate causal effects (*e.g.*, Duflo et al. [2007], Angrist and Pischke [2010]). The internal validity of randomized experiments tends to be high, and their analyses are relatively straightforward (Athey and Imbens [2017]). However, due to the practical challenges involved in running experiments (*e.g.,* Glennerster and Takavarasha [2013]), there are often drawbacks to using experimental data. Experiments are often limited in sample size, in the richness of the information collected, as well as in representativeness, raising concerns about external validity. At the same time, as part of the big data revolution, large, detailed, and by their nature often representative, administrative data sets have become more widely available (*e.g.*, Chetty [2009]). However, it is challenging to use such datasets to estimate causal effects because observational studies often lack internal validity. In this paper, we develop statistical methods for systematically combining experimental and observational data in an attempt to leverage the strengths of both types of data. We focus on a canonical case where both experimental and observational data contain information about individual treatment assignments and a secondary (e.g. short term) outcome (where the datasets contain different individuals), but only the observational data contains information about the primary (often long term) outcome of interest.

We illustrate our methods combining data from the New York school system (the "observational sample") and from Project STAR (the "experimental sample", see Krueger and Whitmore [2001] for an earlier analysis). Our goal is to estimate the effect of class size on 8th grade test scores in New York (what we call the "primary outcome"). However, these 8th grade test scores are not available in the Project STAR data; instead, the experimental sample includes test scores only through the 3rd grade (what we call the "secondary outcome"). See Table 1 for the average outcomes in each sample by treatment status. We find that, even after adjusting for observed pre-treatment student characteristics, the estimated effect of class size on third grade scores is quite different in Project STAR than it is in the observational data from New York. For the experimental sample from Project STAR we see that there is a substantial positive effect of the small class size, an increase of 0.181 in 3rd grade scores. On the other hand, in the observational sample from New York we see a substantial negative relationship between the treatment and

Table 1: Average Outcomes and Standard Errors by Sample and Treatment Status

| | Project STAR | New York | |
|---|---|---|---|
| | 3rd Grade Score<br>(secondary outcome) | 3rd Grade Score<br>(secondary outcome) | 8th Grade Score<br>(primary outcome) |
| Mean Controls (regular class size) | 0.011<br>(0.015) | 0.157<br>(0.001) | 0.155<br>(0.001) |
| Mean Treated (small class size) | 0.193<br>(0.025) | 0.070<br>(0.001) | −0.028<br>(0.002) |
| Difference | 0.181<br>(0.029) | −0.087<br>(0.002) | −0.183<br>(0.002) |

test scores in both 3rd and 8th grade, -0.087 for 3rd grade scores and -0.183 for 8th grade scores.

In this paper we interpret the difference in 3rd grade score estimates, 0.181 and -0.087 as the result of unobserved confounders present in the observational sample. We use the distribution of outcomes in the experimental sample to disentangle the contribution of the unobserved confounders in the observational sample. Under the strong assumption, latent unconfoundedness, that the unobserved confounders for the primary and secondary outcome are the same, we can use this to obtain credible estimates of the causal effects on the primary outcome.

We can interpret this difference in the 3rd grade tests results, 0.181 in Project STAR versus -0.087 in New York, in two ways. One interpretation is that the difference (even after adjusting for pre-treatment variables) is due to differences between the two populations, so that it reflects lack of external validity of the Project STAR sample. A second interpretation is that it reflects lack of internal validity or non-random selection into the treatment in New York, in other words, the presence of unobserved confounders in the New York sample. (In practice, it is of course possible that both complications are present.) In this paper we focus on the latter explanation, and maintain the assumption that the experimental dataset has both internal and external validity; that is, we assume that after adjusting for pre-treatment variables, the underlying populations in the experimental and observational datasets are comparable, even if treatments are assigned differently. Under this maintained assumption, the 0.181 is the preferred estimate of the causal effect on small classes on 3rd grade scores. Given that interpretation the negative correlation between the treatment and 3rd grade outcomes in New York must be due to unobserved confounding, for example, sorting of students who are likely to test poorly into schools with low class sizes. As a result, it appears implausible to interpret the correlation between 8th grade scores and class size in New York as causal.

The main question we address in this paper is how we can adjust the 8th grade results for New York in Table 1 in the light of the experimental Project STAR 3rd grade results and the New York 3rd grade results, under the assumption that the difference in 3rd grade results is due to endogenous selection into the treatment or lack of internal validity in New York. Because 8th and 3rd scores may be measured on different scales, we cannot simply subtract the difference in 3rd grade effects in the two samples, in a difference-in-differences approach. Instead our approaches uses the relation between 3rd grade scores and class size in New York (which is a combination of the causal effect and the selection effect) and the relation between the 3rd grade scores and class size in Project Star (which is causal) to extract the selection component. We then adjust for this selection component to remove the endogeneity in the relation between 8th grade scores and class size in New York.

Formally, we consider a set up with two datasets, the experimental sample and the observational sample. For each unit in the observational dataset, we observe pre-treatment variables, a binary treatment assignment, the primary outcome, and a (vector-valued) secondary outcome. For each unit in the experimental dataset, we observe the same variables as in the observational dataset, except that we do not observe the primary outcome. Table 1 illustrates this observational scheme. This observation scheme is also studied in Rosenman et al. [2018, 2020], Kallus and Mao [2020]. Rosenman et al. [2018] focuses on the problem where assignment is unconfounded in both samples. Kallus and Mao [2020] considers the case where assignment in the combined sample is unconfounded, but not in each of the samples separately. Rosenman et al. [2020] allow for unobserved confounders in the observational sample and consider shrinkage estimators. Kallus et al. [2018] focus on a different case where the same variables are observed in the two samples, but as in our set up, unconfoundedness does not hold in the observational sample. Mealli and Pacini [2013] focuses on an instrumental variables setting where the presence of multiple outcomes can improve estimates.

This set up in this paper differs from the surrogate set up in Athey et al. [2019] where in the

Table 1. Observation Scheme: ✓ is observed, ? is missing

| Units | Sample<br>$G_i$ | Treatment<br>$W_i$ | Primary<br>Outcome<br>$Y_i^{\text{P}}$ | Secondary<br>Outcome<br>$Y_i^{\text{S}}$ | Pretreatment<br>Variables<br>$X_i$ |
|---|---|---|---|---|---|
| 1 to $N_{\text{E}}$ | E | ✓ | ? | ✓ | ✓ |
| $N_{\text{E}}+1$ to $N_{\text{E}}+N_{\text{O}}$ | O | ✓ | ✓ | ✓ | ✓ |

observational study the treatment indicator $W_i$ is not observed. The fact that in the observational sample the treatment indicator is observed allows us to relax the surrogacy assumptions Athey et al. [2019] exploit, in particular the assumption that the only effect of the treatment on the primary outcome is through the secondary, intermediate, outcome. Typically, in our setting, the primary outcome is a long-term outcome such as eventual educational attainment, long-term wages, or mortality, while the secondary outcome may be a multi-dimensional vector of shorter-term outcomes that are associated with the long-term outcome. The object of interest is a low-dimensional estimand, for example, the average causal effect of the treatment on the primary outcome. The role of the secondary outcome and the pretreatment variables is to aid in the effort of credibly estimating the average causal effect on the primary outcome.

Our approach makes use of three maintained assumptions. First, the sample of units in the observational dataset is representative of the population of interest. This assumption essentially defines the estimand. However, the problem is that treatment is not randomly assigned in the observational data. Second, the treatment in the experimental study was randomly assigned, ensuring that the experimental study has internal validity. We can easily generalize this to the case where the maintained assumption is that treatment assignment in the second sample is unconfounded given a set of pretreatment variables (Rosenbaum and Rubin [1983], Imbens and Rubin [2015]). In our application this assumption is satisfied by design. However, because the primary outcome is not observed in this data set, we cannot estimate the average effect of interest on the experimental sample alone. Third, we assume that the pretreatment variables capture the differences between the populations that the observational and experimental sample were drawn from, so that conditional on these pretreatment variables, estimates of the treatment effect in the experimental sample have external validity (Shadish et al. [2002], Hotz et al. [2005]). However, these three maintained assumptions, external validity for the observational sample, and internal validity and conditional external validity for the experimental sample, are not sufficient for identification of the average causal effect of the treatment on the primary outcome.

Our first and main contribution to this problem is to formulate a novel assumption, which we call "latent unconfoundedness." In combination with the three maintained assumptions, this latent unconfoundedness assumption allows for point-identification of the average causal effect of the treatment on the primary outcome in the observational study, without generating testable implications. The latent unconfoundedness assumption implies that the the unobserved confounders in the observational sample that confound treatment assignment and the secondary outcome are the same unobserved confounders that affect both treatment assignment and the primary outcome. Moreover, observing the secondary outcome allows for the extraction of this confounder by exploiting the presence of the experimental data. Formally the assumption links (without the need for functional form assumptions) the biases in treatment-control differences in the secondary outcome (which can be estimated given the presence of the experimental data) to the biases in treatment-control comparisons in the primary outcome (which the experimental data are silent about) using a control function approach (Heckman [1979], Heckman and Robb [1985], Imbens and Newey [2009], Wooldridge). This approach also bears some similarity to the Changes-In-Changes approach in Athey and Imbens [2006]. For a unit in the observational sample the control function is essentially the rank of the secondary outcome in the distribution of secondary outcomes in the experimental sample with the same treatment. The method makes use of the fact that under our maintained assumptions, systematic differences in the estimated effect of the treatment between the experimental and observational sample are attributed to violations of unconfoundedness in the observational data.

In our second contribution, we propose three different approaches to estimation of the average treatment effect under the maintained assumptions in combination with latent uconfoundedness. The three approaches consist of (*i*) imputating the missing primary outcome in the experimental sample, (*ii*) weighting of the units in the observational sample to remove biases, and (*iii*) control function methods. Our analyses show how the presence of the experimental data can be systematically exploited to relax the assumption of unconfoundedness that is common in observational studies.

In our third contribution we apply the new methods to obtain estimates of the effect of small class sizes on 8th grade test scores in New York. The combination of the New York data with the experimental Project STAR data leads to a positive estimates of the effect of small classes. In contrast, an analysis using only the New York data, and assuming unconfoundedness, leads to implausible negative estimates.

# 2 Two Examples

To lay out the conceptual issues at the heart of the current paper we considerin this section two simple examples in some detail. These two examples allow us to introduce the identifying assumptions and estimation strategies that are the main contribution of this paper.

## 2.1 Set Up

The basic set up is the same in both examples. Using the potential outcome set up developed for observational studies by Rubin [1974] (see Imbens and Rubin [2015] for a textbook discussion), let the pair of potential outcomes for this outcome for unit $i$ be denoted by $Y_i^{\rm P}(0)$ and $Y_i^{\rm P}(1)$, where the superscript "P" stands for "Primary". In many applications this is a long term outcome. In our application this is 8th grade scores. The treatment received by unit $i$ will be denoted by $W_i \in \{0,1\}$, an indicator for small class size in our application. There is also a secondary outcome, possibly a short term outcome, with the pair of potential outcomes for unit $i$ denoted by $Y_i^{\rm S}(0)$ and $Y_i^{\rm S}(1)$, where the superscript "S" stands for "Secondary". In our application this is a 3rd grade score. In the two examples both the primary and secondary outcomes are scalars, but in applications the secondary outcome is likely to be vector-valued. The realized values for the primary and secondary outcomes are $Y_i^{\rm P} \equiv Y_i^{\rm P}(W_i)$ and $Y_i^{\rm S} \equiv Y_i^{\rm S}(W_i)$. We may also observe pretreatment variables, denoted by $X_i$

We are interested in the average treatment effect on the primary outcome,

$$\tau^{\rm P} \equiv \mathbb{E}\left[Y_i^{\rm P}(1) - Y_i^{\rm P}(0)\right], \tag{2.1}$$

although other estimands such as the average effect on the treated can be accomodated in this set up. The average effect on the secondary outcome, $\tau^{\rm S} \equiv \mathbb{E}\left[Y_i^{\rm S}(1) - Y_i^{\rm S}(0)\right]$, is, for the purpose of the current study, not of intrinsic interest.

We have two samples to draw on for estimation of $\tau^{\mathrm{P}}$. In that sense the set up connects to the literature on combining data sets, *e.g.,* Hotz et al. [2005], Pearl et al. [2014], Ridder and Moffitt [2007]. The first sample is from an observational study. It is a random sample from the population of interest. For all units in this observational sample we observe the quadruple $(W_i, Y_i^{\mathrm{S}}, Y_i^{\mathrm{P}}, X_i)$, The second sample is a possibly selective sample from the same population, with the assignment completely random. For all units in this experimental sample we observe the triple $(W_i, Y_i^{\mathrm{S}}, X_i)$, but not the primary outcome. The motivation for considering this setting is that it is often expensive to conduct randomized experiments, and it may not be feasible to observe the primary outcome in the experiment.

Let $G_i \in \{\mathrm{E}, \mathrm{O}\}$, be the indicator for the subpopulation or group a unit is drawn from. Then we can think of the combined sample as a random sample of size $N$ from an artificial super-population for which we observe the quadruple $(W_i, G_i, Y_i^{\mathrm{S}}, Y_i^{\mathrm{P}} \mathbf{1}_{G_i=\mathrm{O}}, X_i)$, where $\mathbf{1}_{G_i=\mathrm{O}}$ is a binary indicator, equal to 1 if $G_i = \mathrm{O}$ and equal to 0 if $G_i = \mathrm{E}$.

## 2.2 A Binary Outcome Example

For the purpose of the first example in this section, we assume both the secondary and primary outcome are binary, $Y_i^{\mathrm{P}}(w), Y_i^{\mathrm{S}}(w) \in \{0, 1\}$ for $w \in \{0, 1\}$. For expositional reasons we also assume in this section that there are no pretreatment variables.

For all outcome types $t \in \{\mathrm{S}, \mathrm{P}\}$, all groups $g \in \{\mathrm{E}, \mathrm{O}\}$, and all treatment levels $w \in \{0, 1\}$ the sample averages and sample sizes, define

$$\overline{Y}_w^{t,g} \equiv \frac{1}{N_w^g} \sum_{i=1}^N Y_i^t \mathbf{1}_{G_i=g, W_i=w}, \qquad \text{and} \quad N_w^g \equiv \sum_{i=1}^N \mathbf{1}_{G_i=g, W_i=w}.$$

Assuming that $N_0^{\mathrm{O}}$, $N_1^{\mathrm{O}}$, $N_0^{\mathrm{E}}$, and $N_1^{\mathrm{E}}$ are all positive, six of the eight average outcomes $\overline{Y}_w^{t,g}$ are well-defined and can be calculated from the data, $\overline{Y}_0^{\mathrm{P,O}}$, $\overline{Y}_1^{\mathrm{P,O}}$, $\overline{Y}_0^{\mathrm{S,O}}$, $\overline{Y}_1^{\mathrm{S,O}}$, $\overline{Y}_0^{\mathrm{S,E}}$, and $\overline{Y}_1^{\mathrm{S,E}}$. The remaining two, $\overline{Y}_0^{\mathrm{P,E}}$ and $\overline{Y}_1^{\mathrm{P,E}}$ are not well-defined because we do not observe the primary outcome in the experimental sample.

### 2.2.1 Using the Two Samples Separately

Let us first consider estimation of $\tau^{\mathrm{P}}$ and $\tau^{\mathrm{S}}$ using one sample at a time. Using only the experimental sample there is no way to estimate the average treatment effect on the primary outcome, because this sample does not contain any information on the primary outcome. We can estimate the average effect on the secondary outcome, using the experimental sample, as

$$\hat{\tau}^{\mathrm{S,E}} = \overline{Y}_1^{\mathrm{S,E}} - \overline{Y}_0^{\mathrm{S,E}}.$$

This estimator $\hat{\tau}^{\mathrm{S,E}}$ would be unbiased for $\tau^{\mathrm{S}}$ if the experimental sample had external validity and could be considered a random sample from the population of interest (formally, if $G_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$, what Hotz et al. [2005] call location unconfoundedness). Without that assumption this estimator would not necessarily be unbiased.

Using only the observational sample the natural estimator for the average causal effect on the primary and secondary outcomes would be

$$\hat{\tau}^{\mathrm{P,O}} = \overline{Y}_1^{\mathrm{P,O}} - \overline{Y}_0^{\mathrm{P,O}}, \qquad \text{and} \;\; \hat{\tau}^{\mathrm{S,O}} = \overline{Y}_1^{\mathrm{S,O}} - \overline{Y}_0^{\mathrm{S,O}},$$

respectively. For these estimators to be consistent for the average causal effect of the treatment on the primary and secondary outcomes we would need something like unconfoundedness, which, in the absence of pretreatment variables, corresponds to:

$$W_i \;\perp\!\!\!\perp\; \Big(Y_i^{\mathrm{S}}(0), Y_i^{\mathrm{S}}(0), Y_i^{\mathrm{P}}(0), Y_i^{\mathrm{P}}(0)\Big) \;\Big|\; G_i = \mathrm{O}.$$

With only the observational sample, there is not much in terms of alternatives for obtaining point estimates of the average treatment effect on the primary and secondary outcomes.

### 2.2.2 Combining the Two Samples

Now consider estimation of $\tau^{\mathrm{S}}$ in the presence of both experimental and observational samples. In this case we have two distinct estimators for $\tau^{\mathrm{S}}$, namely $\hat{\tau}^{\mathrm{S,E}}$ and $\hat{\tau}^{\mathrm{S,O}}$. If we find no difference between $\hat{\tau}^{\mathrm{S,E}}$ and $\hat{\tau}^{\mathrm{S,O}}$, or at least no statistically significant difference, then both would appear to be reasonable estimates. We might improve the precision of either estimator by combining them efficiently (*e.g.,* Rosenman et al. [2018, 2020], Kallus and Mao [2020]). However, if we find a substantial and statistically significant difference between the two, we can infer that either the assignment in the observational sample is not random (unconfoundedness does not hold), or the experimental sample is not a random sample from the population of interest (no external validity). In that case there may still be efficiency gains in combining the data, as discussed in Rosenman et al. [2020]. However, in large samples the two estimators will be converging to different limits. There is no information in the data to determine whether unconfoundedness in the observational sample, or external validity in the experimental sample, is violated. Note that these assumptions are of a very different nature. The researcher has to use *a priori* arguments to choose between $\hat{\tau}^{\mathrm{S,E}}$ and $\hat{\tau}^{\mathrm{S,O}}$ and the corresponding assumptions. Choosing for $\hat{\tau}^{\mathrm{S,E}}$ would imply being less concerned with the external validity of the experimental sample, whereas the choice for $\hat{\tau}^{\mathrm{S,O}}$ would imply that the internal validity of the observational study would be viewed as less of a concern. In many cases researchers have argued for the primacy of internal validity over external validity (*e.g.*, Shadish et al. [2002], Imbens [2010]), though others have argued against that perspective (*e.g.*, Manski [2013], Deaton [2010]). If one prefers $\hat{\tau}^{\mathrm{S,O}}$ (downplaying the concerns about internal validity of the observational sample), the natural estimator for $\tau^{\mathrm{P}}$ is $\hat{\tau}^{\mathrm{P,O}}$, and there is little use for the experimental sample. However, if the researcher prefers $\hat{\tau}^{\mathrm{S,E}}$, the question arises how to estimate $\tau^{\mathrm{P}}$.

The current paper is concerned with this question. We take the position that there is an *a priori* preference for $\hat{\tau}^{\mathrm{S,E}}$ over $\hat{\tau}^{\mathrm{S,O}}$, possibly after accounting for differences in covariate distributions to deal with some of the external validity concerns (Hotz et al. [2005]). Then we address the main question of how to adjust the estimator $\hat{\tau}^{\mathrm{P,O}}$ to take into account the difference between $\hat{\tau}^{\mathrm{S,O}}$ and $\hat{\tau}^{\mathrm{S,E}}$, Conceptually there are multiple natural ways of doing so. We discuss three of these ways. The first one is based on imputation of the missing primary outcomes in the experimental sample. The second one is based on weighting the units in the observational sample. The third one is based on a control function approach. In this simple nonparametric case with binary outcomes the three approaches lead to identical point estimates .

### 2.2.3 Imputation

The first approach is to take a missing data perspective on the primary $Y_i^{\mathrm{P}}$ in the experimental sample and impute these missing values using the observational sample. Consider unit $i$ in the experimental sample with $W_i = w$ and $Y_i^{\mathrm{S}} = y^{\mathrm{S}}$. Assuming the missing data on $Y_i^{\mathrm{P}}$ are missing at random (*e.g.,* Rubin [1976], Little and Rubin [2019], Rubin [2004]) suggests using the distribution of $Y_i^{\mathrm{P}}$ among units in the observational sample with $W_i = w$ and $Y_i^{\mathrm{S}} = y^{\mathrm{S}}$ to impute the missing values. If we are interested in estimating the average effect, we can just use the average value of $Y_i^{\mathrm{P}}$ in this subsample as the imputed value. Denote this average value for all values of $y^{\mathrm{S}}$ and $w$ by:

$$\overline{Y}^{\mathrm{P,O}}_{w,y^{\mathrm{S}}} = \sum_{i=1}^{N} \mathbf{1}_{G_i=\mathrm{O},W_i=w,Y_i^{\mathrm{S}}=y^{\mathrm{S}}} Y_i^{\mathrm{P}} \Big/ \sum_{i=1}^{N} \mathbf{1}_{G_i=\mathrm{O},W_i=w,Y_i^{\mathrm{S}}=y^{\mathrm{S}}}.$$

The imputed value for $Y_i^{\mathrm{P}}$ for unit $i$ in the experimental sample is then the average of $Y_j^{\mathrm{P}}$ in the observational sample over all units $j$ with the same treatment level, $W_j = W_i$, and the same value for the secondary outcome, $Y_j^{\mathrm{S}} = Y_i^{\mathrm{S}}$:

$$\hat{Y}_i^{\mathrm{P}} = \overline{Y}^{\mathrm{P,O}}_{W_i,Y_i^{\mathrm{S}}}.$$

Then the imputation estimator for $\tau^{\mathrm{P}}$ is the difference in average imputed values in the experimental sample by treatment status, leading to the first estimator for $\tau^{\mathrm{P}}$:

$$\hat{\tau}^{\mathrm{P,imp}} \equiv \frac{1}{N_1^{\mathrm{E}}} \sum_{i:G_i=\mathrm{E}} W_i \overline{Y}^{\mathrm{P,O}}_{W_i,Y_i^{\mathrm{S}}} - \frac{1}{N_0^{\mathrm{E}}} \sum_{i:G_i=\mathrm{E}} (1-W_i) \overline{Y}^{\mathrm{P,O}}_{W_i,Y_i^{\mathrm{S}}}. \tag{2.2}$$

### 2.2.4 Weighting

The second approach to using the experimental secondary outcomes is to reweight the observational sample where we do observe the primary outcome. Consider the $N_w^{\mathrm{O}}$ units in the observational sample with $W_i = w$. The fraction of those treated units in the observational study with secondary outcome $Y_i^{\mathrm{S}} = 1$ is $\hat{p}_w^{\mathrm{S,O}} = \sum_{i:G_i=\mathrm{O},W_i=w} Y_i^{\mathrm{S}} / \sum_{i:G_i=\mathrm{O},W_i=w} 1$. The experimental study tells us this fraction would have been approximately $\hat{p}_w^{\mathrm{S,E}} = \sum_{i:G_i=\mathrm{E},W_i=w} Y_i^{\mathrm{S}} / \sum_{i:G_i=\mathrm{E},W_i=w} 1$, had the treatment been randomly assigned. The comparison $\hat{p}_w^{\mathrm{S,O}}$ versus $\hat{p}_w^{\mathrm{S,E}}$ reflects on the possible violation of the unconfoundedness assumption in the observational sample. We can give these units a weight $\lambda_{w,1} = \hat{p}_w^{\mathrm{S,E}} / \hat{p}_w^{\mathrm{S,O}}$ to adjust for the bias stemming from such violations. Similarly, units with treatment $W_i = w$ and $Y_i^{\mathrm{S}} = 0$ would be given a weight $\lambda_{w,0} = (1-\hat{p}_w^{\mathrm{S,E}})/(1-\hat{p}_w^{\mathrm{S,O}})$. We then use these to estimate the average effect on the primary outcome as the difference of weighted averages of the treated and control outcomes in the observational sample, leading to the second estimator:

$$\hat{\tau}^{\mathrm{P,weight}} \equiv \frac{\sum_{G_i=\mathrm{O}} \lambda_{1,Y_i^{\mathrm{S}}} W_i Y_i^{\mathrm{P}}}{\sum_{G_i=\mathrm{O}} \lambda_{1,Y_i^{\mathrm{S}}} W_i} - \frac{\sum_{G_i=\mathrm{O}} \lambda_{0,Y_i^{\mathrm{S}}} (1-W_i) Y_i^{\mathrm{P}}}{\sum_{G_i=\mathrm{O}} \lambda_{0,Y_i^{\mathrm{S}}} (1-W_i)}. \tag{2.3}$$

Simple algebra shows that the two estimators for $\tau$ are identical, $\hat{\tau}^{\mathrm{P,imp}} = \hat{\tau}^{\mathrm{P,weight}}$. This algebraic result relies on the outcome model and the weights being fully nonparametric in this simple example with the secondary outcome taking on only two values. In settings with the secondary outcome continuous, the two approaches will generally give different answers in finite samples.

## 2.3 A Control Function Approach in a Linear Model Setting

In our second example we consider a simple linear model that exhibits most clearly some of the key features of the approach in the current paper. Specifically, suppose we have a linear model for the secondary potential outcomes with a constant treatment effect:

$$Y_i^{\mathrm{S}}(0) = X_i^\top \gamma^{\mathrm{S}} + \alpha_i^{\mathrm{S}}, \qquad Y_i^{\mathrm{S}}(1) = Y_i^{\mathrm{S}}(0) + \tau^{\mathrm{S}}.$$

This models holds for both the experimental and observational samples. The properties of the unobserved component $\alpha_i^{\mathrm{S}}$ are key, and they may differ in the two samples. In the experimental sample the randomization guarantees that we have the following conditional independence:

$$W_i \ \perp\!\!\!\perp \ \alpha_i^{\mathrm{S}} \ \Big| \ X_i, G_i = \mathrm{E}.$$

In fact the randomization implies even stronger conditions, $W_i \perp\!\!\!\perp \alpha_i^{\mathrm{S}}, X_i | G_i = \mathrm{E}$, but we do not need those here. In the observational study we do not in general have the same conditional independence:

$$W_i \ \not\!\perp\!\!\!\perp \ \alpha_i^{\mathrm{S}} \ \Big| \ X_i, G_i = \mathrm{O}.$$

This randomization in the experimental sample implies that we can estimate the parameters of the model for the secondary outcome, $\tau^{\mathrm{S}}$ and $\gamma^{\mathrm{S}}$, by least squares regression of $Y_i^{\mathrm{S}}$ on $W_i$ and $X_i$ using only the data from the experimental sample. In other words, the conditional mean of $Y_i^{\mathrm{S}}$ given $W_i$ and $X_i$ has a causal interpretation as a function of $W_i$ in the experimental sample, but not in the observational sample.

Now consider the primary outcome. We specify a similar linear model for the primary outcome, but allow the coefficients to be different from those of the model for the secondary outcome,

$$Y_i^{\mathrm{P}}(0) = X_i^\top \gamma^{\mathrm{P}} + \alpha_i^{\mathrm{P}}, \qquad Y_i^{\mathrm{P}}(1) = Y_i^{\mathrm{P}}(0) + \tau^{\mathrm{P}}.$$

Again the concern is that in the observational sample the unobserved component might be correlated with the treatment:

$$W_i \ \not\!\perp\!\!\!\perp \ \alpha_i^{\mathrm{P}} \ \Big| \ X_i, G_i = \mathrm{O}.$$

Such a correlation would imply that a linear regression of $Y_i^{\mathrm{P}}$ on $W_i$ and $X_i$ using the data from the observational sample would not be consistent for the causal effect $\tau^{\mathrm{P}}$ because of endogeneity of $W_i$. Now a key assumption is that there is a relationship between the short term and long term unobserved components $\alpha_i^{\mathrm{P}}$ and $\alpha_i^{\mathrm{P}}$ that allows us to remove the endogeneity bias in the long term relationship using the difference between the short term results for the experimental and observational data using a control function approach (Heckman and Robb [1985], Imbens and Newey [2009], Kline and Walters [2019]). The key assumption that links the endogeneity problems for the primary and secondary outcomes is

$$\alpha_i^{\mathrm{P}} = \delta \alpha_i^{\mathrm{S}} + \varepsilon_i^{\mathrm{P}}, \qquad \text{with } \; W_i \;\perp\!\!\!\perp\; \varepsilon_i^{\mathrm{P}} \;\Big|\; X_i, \alpha_i^{\mathrm{S}}, G_i = \mathrm{O}. \tag{2.4}$$

Later we relax this assumption to remove the functional form dependence, but for the moment let us focus on this version with linearity and additivity. The key is that the residual for the primary outcome, $\alpha_i^{\mathrm{P}}$, is related to the residual for the secondary outcome, $\alpha_i^{\mathrm{S}}$, with the remainder, $\varepsilon_i^{\mathrm{P}} = \alpha_i^{\mathrm{P}} - \mathbb{E}[\alpha_i^{\mathrm{P}}|\alpha_i^{\mathrm{S}}]$ unrelated to the treatment.

Let us show in some detail how this assumptions aids in the identification of $\tau^{\mathrm{P}}$ in this linear example. First, we can estimate $\tau^{\mathrm{S}}$ and $\gamma^{\mathrm{S}}$ from the experimental sample by linear regression. Denote these least squares estimates by $\hat{\tau}^{\mathrm{S}}$ and $\hat{\gamma}^{\mathrm{S}}$. Then we can estimate the residual $\alpha_i^{\mathrm{S}}$ for the units in the observational sample as

$$\hat{\alpha}_i^{\mathrm{S}} = Y_i^{\mathrm{S}} - W_i \hat{\tau}^{\mathrm{S}} - X_i^{\top} \hat{\gamma}^{\mathrm{S}}. \tag{2.5}$$

If this model is correct, and if the assignment to treatment in the observational sample were random, and finally, if the observational and experimental samples were randomly drawn from the same population, the population value of these residuals $\alpha_i^{\mathrm{S}}$ would have mean zero and be uncorrelated with the treatment indicator in the observational sample. The presence of non-zero association of this residuals and the treatment is exploited to adjust the estimates of the treatment effect on the primary outcome. We can do so by including this residual as a control variable in the least squares regression with the long term outcome as the dependent variable, using the observational data. The key insight is that we can use the linear representation in (2.4) to write the long te outcome as:

$$Y_i^{\mathrm{P}} = W_i \tau + X_i^{\top} \gamma + \delta \alpha_i^{\mathrm{S}} + \varepsilon_i^{\mathrm{P}}, \qquad \text{with } \; W_i \perp\!\!\!\perp \varepsilon_i^{\mathrm{P}} \;\Big|\; X_i, \alpha_i^{\mathrm{S}}, G_i = \mathrm{O}. \tag{2.6}$$

Therefore this regression will lead to a consistent estimator for $\tau^{\mathrm{P}}$ under the current assumptions.

To further develop intuition for the control function approach, consider the example where the primary and secondary outcomes are eight and third grade test scores, and the treatment is class size. Using the experimental sample we estimate the the average effect of the class size on third grade scores. We then calculate the residuals in the observational study. We may find that the residuals are larger on average for the treated individuals than for the control individuals. This suggests that the treatment assignment in the observational sample was correlated with the third grade potential outcomes, with individuals with high values for the potential outcomes more likely to be in the treatment group. We then use that information to compare eighth grade scores for individuals with the same residuals, so we adjust for the original non-random selection into treatment.

## 2.4 The Connection Between the Imputation and Control Function Approaches

The control function approach to dealing with the endogeneity of the treatment in the observational study we used in the second example may appear at first sight to be conceptually quite different from the weighting and imputation approaches in the first example. In fact the two approaches are closely related. Consider the imputation of $Y_i^{\mathrm{P}}$ for a unit in the experimental sample given the linear model. Substituting for $\alpha_i^{\mathrm{S}}$ using Equation (2.5) into Equation (2.6) implies that we can write for the primary outcome in the observational sample:

$$Y_i^{\mathrm{P}} = W_i\beta + X_i^{\top}\lambda + \delta Y_i^{\mathrm{S}} + \varepsilon_i,$$

where

$$\beta = \tau^{\mathrm{P}} - \delta\tau^{\mathrm{S}}, \qquad \text{and} \quad \lambda = \gamma^{\mathrm{P}} - \delta\gamma^{\mathrm{S}}.$$

Hence the imputed value for $Y_i^{\mathrm{P}}$ in the experimental sample using the estimated parameters from the observational sample leads to (ignoring estimating error)

$$\hat{Y}_i^{\mathrm{P}} = W_i\beta + X_i^{\top}\lambda + \delta Y_i^{\mathrm{S}}.$$

Using the omitted variable bias formula it is easy to see that regressing this imputed value $\hat{Y}_i^{\mathrm{P}}$ on $W_i$ and $X_i$ (but omitting $Y_i^{\mathrm{S}}$), in the experimental sample, leads to

$$\hat{Y}_i^{\mathrm{P}} = W_i\bar{\beta} + X_i^{\top}\bar{\lambda} + \varepsilon_i, \qquad \text{with} \quad \bar{\beta} = \beta + \delta\beta^{\mathrm{S}} = \tau^{\mathrm{P}}.$$

Thus, the coefficient on $W_i$ in this regression of the imputed primary outcome on the treatment and the pretreatment variables is consistent for the causal effect of the treatment on the primary outcome.

# 3 The General Case

The two examples in the preceeding section convey much of the intuition for our approach:. The key assumption in the linear case is (2.4), which connects the bias in causal estimates for the primary and secondary outcomesin the observational sample. Making such assumptions allows us improve upon estimates for $\tau^{\mathrm{P}}$ based on the observational sample alone. What we do in this section is generalize the first example to the case where (*i*) the secondary and primary outcomes may be continuous and (*ii*) the secondary outcome may be vector-valued, and (*iii*) where pre-treatment variables are present. We also generalize the control function approach to (*i*) the nonlinear case, and (*ii*) the case with multiple secondary outcomes in order to allow the critical assumptions to be weakened. We present the formal assumptions that justify the weighting and imputation estimators, and present their general forms and how they relate to the control function approach.

We are interested in causal estimands defined for the population of interest. At a general level such estimands include simple average treatment effects, but more generally also the average effect of a policy that assigns the treatment to individuals in this population on the basis of covariates (*e.g.*, Manski [2004], Dehejia [2005], Hirano and Porter [2009], Athey and Wager [2017], Zhou et al. [2018]).

Define

$$\tau_g^t \equiv \mathbb{E}\left[ Y_i^t(1) - Y_i^t(0) \middle| G_i = g \right], \tag{3.1}$$

is the average effect of the treatment on outcome $t \in \{\mathrm{S}, \mathrm{P}\}$ for group $g \in \{\mathrm{O}, \mathrm{E}\}$. The superscripts on the estimands denote the outcome, and subscripts denote the population. The primary estimand we focus on in this paper is the average effect of the treatment on the long term outcome in the observational study population:

$$\tau \equiv \tau_{\mathrm{O}}^{\mathrm{P}} \equiv \mathbb{E}\left[ Y_i^{\mathrm{P}}(1) - Y_i^{\mathrm{P}}(0) \middle| G_i = \mathrm{O} \right], \tag{3.2}$$

where we drop the subscript and superscript to simplify the notation.

## 3.1 Three Maintained Assumptions

There are three key features of our set up. First, we are interested in the population that the units in the observational study were drawn from. That is, the observational study has external validity.

**Assumption 1.** (External Validity of the Observational Study) *The observational sample is a random sample of the population of interest.*

At some level this can be thought of as simply defining the estimand in terms of the population distribution underlying the observational sample.

Second, we maintain throughout the paper the assumption that the treatment in the experimental sample is unconfounded.

**Assumption 2.** (Internal Validity of the Experimental Sample) *For $w = 0, 1$,*

$$W_i \perp\!\!\!\perp \Big(Y_i^{\mathrm{P}}(w), Y_i^{\mathrm{S}}(w)\Big) \Big| X_i, G_i = \mathrm{E}. \tag{3.3}$$

Kallus and Mao [2020] make a different assumption here,

$$W_i \perp\!\!\!\perp \Big(Y_i^{\mathrm{P}}(w), Y_i^{\mathrm{S}}(w)\Big) \Big| X_i, \tag{3.4}$$

where unconfoundedness holds in the combined sample, rather than in the experimental sample. Assumption (3.4) does not imply our assumption (3.3), or the other way around. In our application, with assignment in the Project STAR experimental sample completely randomized, our assumption is satisfied by design, and in general (3.4) would not hold. In other settings, for example where the data are sampled from a single population rather than two separate populations, (3.4) may be more appropriate than our assumption.

However, the external validity of the experimental study is not guaranteed. Instead we assume that conditional on the pretreatment variables we have external validity (Hotz et al. [2005]):

**Assumption 3.** (Conditional External Validity) *The experimental study has conditional external validity if*

$$G_i \perp\!\!\!\perp \Big(Y_i^{\mathrm{P}}(0), Y_i^{\mathrm{P}}(1), Y_i^{\mathrm{S}}(0), Y_i^{\mathrm{S}}(1)\Big) \Big| X_i. \tag{3.5}$$

This assumption implies that if we find systematic differences between in differences in average outcomes by treatment status conditional on covariates between the experimental and observational sample, these differences must arise from violations of unconfoundedness for the observational sample.

The first result is that these three maintained assumptions are in general not sufficient for point-identification of the average effect of interest. Of course this does not mean that these assumptions do not have any identifying power. They do in fact affect the identified sets in the spirit of the work by (Manski [1990]).

**Lemma 1.** *The combination of Assumptions 1-3 is not sufficient for point-identification of* $\tau^{\mathrm{P}}$.

The proof for this result is given in the appendix.

## 3.2 Unconfoundedness for the Observational Sample

Next, let us consider the assumption that assignment in the observational study is unconfounded.

**Assumption 4.** (Unconfoundedness in the Observational Sample)
*For* $w = 0, 1$,

$$W_i \;\perp\!\!\!\perp\; \Big(Y_i^{\mathrm{S}}(w), Y_i^{\mathrm{P}}(w)\Big) \;\Big|\; X_i, G_i = \mathrm{O}, \tag{3.6}$$

This assumption is made, for example, in Rosenman et al. [2018]. This assumption is sufficient for identification of $\tau$, but it is stronger than necessary. Intuitively it implies that we do not need the experimental sample for identification because under unconfoundedness the observational sample is sufficient for identification of the average treatment effect. However, the experimental sample may still be useful for precision. The precise version of the unconfoundedness assumption here is slightly different from that in, say, Rosenbaum and Rubin [1983] where it is assumed that $W_i$ is independent of the full set of $(Y_i^{\mathrm{P}}(w), Y_i^{\mathrm{S}}(w))_{w\in\{0,1\}}$. It is what is referred to in Imbens [2000] as "weak unconfoundedness." This issue will come up later.

**Lemma 2.** *The set of Assumptions 2-4 has a testable implication:*

$$G_i \;\perp\!\!\!\perp\; Y_i^{\mathrm{S}} \;\Big|\; X_i, W_i. \tag{3.7}$$

We can also use this result to assess whether a particular set of pre-treatment variables is sufficient for unconfoundedness. Here we are interested in finding a set of pretreatment variables $X_i$ such that

$$G_i \ \perp Y_i^{\mathrm{S}} \ \Big| \ X_i = x, W_i, \tag{3.8}$$

holds.

## 3.3 Latent Unconfoundedness

Suppose that we reject the conditional independence in Lemma 2, so that we know that the full set of maintained assumptions, 2-4, does not hold. If we maintain unconfoundedness in the experimental sample, it must be that either conditional external validity in the experimental study, or unconfoundedness in the observational study must be violated. If we interpret such a finding as evidence against conditional external validity, and are willing to maintain unconfoundedness of the treatment assignment in the observational study, we should simply put aside the experimental data set and focus on estimates based on solely on the observational study. In many cases, however, we may wish to maintain conditional external validity and interpret a finding that (3.7) does not hold as evidence that unconfoundedness does not hold for the observational study. Here we explore methods for using the difference between the estimates of the causal effects for the experimental study (which we know to be internally valid) and the estimates for the observational study (which need not be internally valid) to adjust long term estimates for the observational study.

The idea, although not the implementation, is somewhat similar to that in a Difference-In-Differences (Card [1990], Card and Krueger [1994], Angrist and Pischke [2008]) set up where the initial (pre-treatment) differences between a treatment and control group are used to adjust post-treatment differences between the treatment and control group. More specifically, it relates to the Changes-In-Changes approach in Athey and Imbens [2006] where functional form assumptions are avoided. Here initial differences in treatment effects between an experimental and observational study are used to adjust subsequent treatment effects for the observational study.

They key additional assumption that links the biases, in the observational study, between adjusted comparisons for the primary and secondary outcomes, is the following.

**Assumption 5.** (Latent Unconfoundedness)
*For $w \in \{0,1\}$,*

$$W_i \perp\!\!\!\perp Y_i^{\mathrm{P}}(w) \;\Big|\; X_i, Y_i^{\mathrm{S}}(w), G_i = \mathrm{O}. \tag{3.9}$$

This assumption is both novel as well as critical in the current discussion, so let us offer some remarks.

**Remark 1.** *Compared to a regular unconfoundedness assumption, we add the variable $Y_i^{\mathrm{S}}(w)$ to the conditioning set. At first this may appear to be an innocuous addition. However, following the standard approach to exploiting unconfoundedness assumptions, we see that this is not the case. Typically we use such an assumption to create subpopulations defined by the conditioning variables, and then compare treated and control units. To be specific, suppose we wish to estimate $\mathbb{E}[Y_i^{\mathrm{P}}(1)|G_i = \mathrm{O}]$. We would first estimate the conditional expectation $\mathbb{E}[Y_i^{\mathrm{P}}(1)|Y_i^{\mathrm{S}}(1) = y^{\mathrm{S}}, W_i = 1, X_i = x, G_i = \mathrm{O}]$. Then, however, we would need to average this over the marginal distribution of $(Y_i^{\mathrm{S}}(1), X_i)$ in the observational sample, but in this observational sample we only see draws from the conditional distribution of $(Y_i^{\mathrm{S}}(1), X_i)$ given $W_i = 1$, and this is not the same distribution because of the failure of unconfoundedness in the observational sample. To address this, we need to exploit the presence of the experimental sample.*

To highlight the link to the control function literature (Heckman [1979], Heckman and Robb [1985], Imbens and Newey [2009], Wooldridge, Athey and Imbens [2006], Kline and Walters [2019], Mogstad et al. [2018], Mogstad and Torgovitsky [2018], Wooldridge [2015]), let us model the primary and secondary outcomes as

$$Y_i^{\mathrm{P}}(w) = h^{\mathrm{P}}(w, \nu_i, X_i), \qquad \text{and } \; Y_i^{\mathrm{S}}(w) = h^{\mathrm{S}}(w, \eta_i, X_i),$$

with $h^{\mathrm{S}}(w, \eta, x)$ strictly monotone in $\eta$. Now we can write the latent unconfoundedness assumption as

$$W_i \perp\!\!\!\perp \nu_i \;\Big|\; X_i, \eta_i, G_i = \mathrm{O}.$$

Although it is not generally true that $W_i \perp\!\!\!\perp \nu_i | X_i, G_i = \mathrm{O}$, adding $\eta_i$ to the conditioning set restores the exogeneity of $W_i$ in the observational sample.

It is useful to contrast this with a control function in a non-parametric instrumental variables setting (*e.g.,* Imbens and Newey [2009]), where the two models are

$$Y_i^{\mathrm{P}}(w) = h^{\mathrm{P}}(w, \nu_i, X_i), \qquad \text{and} \;\; W_i(z) = r(z, \eta_i, X_i),$$

with $r(z, \eta, x)$ strictly monotone in $\eta$. The key assumption here is that

$$W_i \;\perp\!\!\!\perp\; \nu_i \;\Big|\; X_i, \eta_i.$$

The model relating the outcome of interest and the endogenous regressor is essentially the same in the two settings, $Y_i^{\mathrm{P}}(w) = h^{\mathrm{P}}(w, \nu_i, X_i)$. In both cases we address the endogeneity by conditioning on an additional variable, the control variable $\eta_i$. This control variable is estimated using an auxiliary model. This auxilliary model differs between the set up in the current paper and the instrumental variables setting. In the instrumental variables setting we model the relation between the endogenous regressor and an additional variable, the instrument, and deriving the control variable from that relation. In the current setting we model the relation between the secondary outcome and the endogenous regressor and deriving the control variable from that relation. In both cases the auxiliary model has a strict monotonicity assumption. This shows some of the limitations of the approach: the unobserved confounder $\eta$ cannot have a dimension higher than that of the secondary outcome.

Formally, adding Assumption 5 (latent unconfoundedness) to Assumptions 1-3 allows us to point-identify the average effect of interest.

**Theorem 1.** *Suppose that Assumptions 1-3 and 5 hold, so that the experimental study is unconfounded and has conditional external validity, and the observational study has latent unconfoundedness. Then $\tau_{\mathrm{O}}^{\mathrm{P}}$, the average effect of the treatment on the primary outcome in the observational study is point-identified.*

## 3.4 Missing At Random

There is an interesting and close connection between Assumptions 1-3 and 5 and the Missing-At-Random (MAR) assumption in the missing data literature (Rubin [1976], Little and Rubin [2019], Rubin [2004]).

**Lemma 3.** *Suppose that Assumptions 1-3 and 5 hold. Then:*

$$G_i \;\perp\!\!\!\perp\; Y_i^{\mathrm{P}} \;\Big|\; W_i, X_i, Y_i^{\mathrm{S}}. \tag{3.10}$$

Because $G_i = \mathrm{E}$ is equivalent to an indicator that $Y_i^{\mathrm{P}}$ missing, and because $W_i$, $X_i$, and $Y_i^{\mathrm{S}}$ are observed for all individuals in the sample, the conditional independence in (3.10) is equivalent to a MAR assumption. The result does not go the other way around. The MAR assumption by itself has no testable implications, but the combination of Assumptions Assumptions 1-3 and 5 does imply some inequality restrictions on the joint distribution of the observed variables. Kallus and Mao [2020] starts with a Missing-At-Random assumption, and uses that in combination with an unconfoundedness assumption on the full sample to identify the average effect of the treatment for the full sample.

# 4 Estimation and Inference

In this section we extend the same three estimation strategies we discussed in the examples in Section 2, imputation, weighting, and control function methods, to the general case.

## 4.1 The Imputation Approach

First, consider the imputation approach. Estimate the conditional mean of the primary outcome given the secondary outcome, treatment and pre-treatment variables in the observational sample:

$$\kappa(w, x, y, \mathrm{O}) \equiv \mathbb{E}\left[Y_i^{\mathrm{P}} \middle| W_i = w, X_i = x, Y_i^{\mathrm{S}} = y, G_i = \mathrm{O}\right].$$

Then impute for the units in the experimental sample the primary outcome as $\hat{Y}_i^{\mathrm{P}} = \hat{\kappa}(W_i, X_i, Y_i^{\mathrm{S}}, \mathrm{O})$. Then use the standard program evaluation methods to adjust for differences in covariates if necessary. If in the experimental sample the treatment is completely random, we would estimate the average treatmet effect in the experimental sample as

$$\hat{\tau}^{\mathrm{imp,E}} = \frac{1}{N_1^{\mathrm{E}}} \sum_{i:P_i=\mathrm{E}} W_i \kappa(1, X_i, Y_i^{\mathrm{S}}, \mathrm{O}) - \frac{1}{N_1^{\mathrm{E}}} \sum_{i:P_i=\mathrm{E}} (1 - W_i)\kappa(0, X_i, Y_i^{\mathrm{S}}, \mathrm{O}).$$

However, we wish to estimate the average effect in the observational sample, which may have a different distribution of the pre-treatment variables. This requires one additional layer of adjustment that depends on the pre-treatment variables. Define

$$r(x) = \mathrm{pr}(G_i = \mathrm{O} | X_i = x).$$

Then we weight the units by the ratio $r(X_i)/(1 - r(X_i))$:

$$\hat{\tau}^{\mathrm{imp}} = \frac{\sum_{i:P_i=\mathrm{E}} W_i \kappa(1, X_i, Y_i^{\mathrm{S}}, \mathrm{O}) r(X_i)/(1 - r(X_i)}{\sum_{i:P_i=\mathrm{E}} W_i r(X_i)/(1 - r(X_i)}$$
$$- \frac{\sum_{i:P_i=\mathrm{E}} (1 - W_i) \kappa(0, X_i, Y_i^{\mathrm{S}}, \mathrm{O}) r(X_i)/(1 - r(X_i))}{\sum_{i:P_i=\mathrm{E}} (1 - W_i) r(X_i)/(1 - r(X_i))}.$$

## 4.2 The Weighting Approach

Second, consider the weighting approach. Estimate the distribution of $(Y_i^{\mathrm{S}}, W_i)$ in the observational and experimental sample as

$$f_{W,Y^{\mathrm{S}}|X,P}(w, y^{\mathrm{S}} | x, p),$$

for all $x \in \mathbb{X}$ and $p \in \{\mathrm{E}, \mathrm{O}\}$. Then construct the weights for all units in the observational sample as a function of $(W_i, X_i, Y_i^{\mathrm{S}})$:

$$\lambda_i = \frac{f_{W,Y^{\mathrm{S}}|X,P}(W_i, Y_i^{\mathrm{S}} | X_i, \mathrm{E})}{f_{W,Y^{\mathrm{S}}|X,P}(W_i, Y_i^{\mathrm{S}} | X_i, \mathrm{O})}.$$

These weights adjust for the differences between the observational and experimental sample.

Assuming we have completely random assignment in the experimental sample, we estimate the average treatment effect as

$$\hat{\tau}^{\mathrm{weight}} = \frac{\sum_{i:P_i=\mathrm{O}} Y_i W_i \lambda_i}{\sum_{i:P_i=\mathrm{O}} (1 - W_i) \lambda_i} - \frac{\sum_{i:P_i=\mathrm{O}} (1 - W_i) \lambda_i}{\sum_{i:P_i=\mathrm{O}} W_i \lambda_i}.$$

If instead we have unconfounded treatment assignment in the experimental sample, we need the weights that adjust for the non-randomness in the experimental sample. By the maintained assumptions this requires only adjusting for the differences in pre-treatment variables. Let the propensity score be

$$e(x, g) = \mathrm{pr}(W_i = 1 | X_i = x, G_i = g).$$

This leads to

$$\hat{\tau}^{\mathrm{weight}} = \frac{\sum_{i:P_i=\mathrm{O}} Y_i W_i \lambda_i / e(X_i, \mathrm{E})}{\sum_{i:P_i=\mathrm{O}} (1 - W_i) \lambda_i / e(X_i, \mathrm{E})} - \frac{\sum_{i:P_i=\mathrm{O}} (1 - W_i) \lambda_i / (1 - e(X_i, \mathrm{E}))}{\sum_{i:P_i=\mathrm{O}} W_i \lambda_i / (1 - e(X_i, \mathrm{E}))}.$$

### 4.3 The Control Function Approach

Finally, the control function approach. First estimate the conditional distribution of the secondary outcome given treatment and pre-treatment variables in both samples:

$$F_{Y^{\mathrm{S}}|W,X,G}(y^{\mathrm{S}}|w,x,g) = \mathrm{pr}(Y_i^{\mathrm{S}} \leq y^{\mathrm{S}}|W_i = w, X_i = x, G_i = g).$$

Then calculate the control variable for each unit in the observational sample as

$$\hat{\eta}_i = \hat{F}_{Y^{\mathrm{S}}|W,X,G}(Y_i^{\mathrm{S}}|W_i, X_i, \mathrm{E}).$$

Next, estimate the conditional mean of the primary outcome in the observational sample given treatment status, control variable, and pre-treatment variables:

$$\gamma(w,h,x) \equiv \mathbb{E}\left[ Y_i^{\mathrm{P}} \middle| W_i = w, \eta_i = h, X_i = x, G_i = \mathrm{O}\right].$$

Finally, estimate the average treatment effect $\tau$ as

$$\hat{\tau}^{\mathrm{cf}} = \frac{1}{N_1^{\mathrm{E}}} \sum_{i:G_i=\mathrm{E}} W_i \hat{\gamma}(1, \hat{\eta}_i, X_i) - \frac{1}{N_0^{\mathrm{E}}} \sum_{i:G_i=\mathrm{E}} (1-W_i)\hat{\gamma}(1, \hat{\eta}_i, X_i).$$

# 5 An Application

To illustrate the ideas in this paper we analyze data on the effect of class size on educational outcomes. We use the data from the Project STAR experiment on class size, where we observe for all children whether they are in a regular or small class. As the short term outcome we use the third grade score. We also observe the pre-treatment variables gender, whether the student gets a free lunch, and ethnicity. For the observational data we use data from the New York school system. We observe the same variables, but also eighth grade scores.

In Table 2 we report the results from several ordinary least squares regressions. The first two columns show the results from a regression of the short term outcome on the treatment, separately for the two samples, and controlling for the pretreatment variables. In the experimental sample in Project STAR we find a positive effect of small class size of 0.157. In the observational sample the least squares estimate is negative, -0.048. This suggests that there are unmeasured confounders. If we regress the eighth grade scores on the treatment in the observational sample we still get a negative estimate, -0.074. Now we follow the control function approach and include in that linear regression the control function, that is, the estimated residual:

$$\hat{\alpha}_i^{\mathrm{S}} = Y_i^{\mathrm{S}} - W_i\hat{\beta}^{\mathrm{S}} - X_i^{\top}\hat{\gamma}_i^{\mathrm{S}}.$$

Including this in the regression gives a coefficient of 0.640 with a standard error of 0.001. It changes the coefficient on the treatment to 0.061, now much more in line with what one would expect given the causal effect of a small class size on the third grade scores in the experimental sample. We can also do this through the imputation approach. First we use the regression of the long term outcome on short term outcome and covariates to predict the long term outcome for the observations in the experimental sample. Then we regress the predicted value on the treatment, leading to the same estimate of 0.061.

Table 2:

| | Secondary | Secondary | Primary | Primary | Imputed Primary |
|---|---|---|---|---|---|
| $W_i$ | 0.157<br>(0.028) | $-0.048$<br>(0.002) | $-0.074$<br>(0.003) | 0.061<br>(0.018) | 0.061<br>(0.018) |
| $\hat{\alpha}_i^S$ | | | | 0.640<br>(0.001) | |
| $N$ | $6,027$ | $1,131,339$ | $498,597$ | $498,597$ | $6,027$ |
| $R^2$ | 0.130 | 0.060 | 0.040 | 0.420 | 0.170 |
| Sample | Experimental | Observational | Observational | Observational | Experimental |
| Covariates | Yes | Yes | Yes | Yes | Yes |

We also investigate if the surrogacy assumption holds here. Estimating a regression of the primary outcome on the secondary outcome and the treatment indicator (and including pre-treatment variables), leads to a coefficient on the treatment indicator of -0.039, with a standard error of 0.002. Thus, it appears that third grade scores are not a valid surrogate for eighth grade scores. Thus, the effect of class size is not fully captured by third grade scores, but also arises through other channels. Thus, accounting for the latent confounder is important.

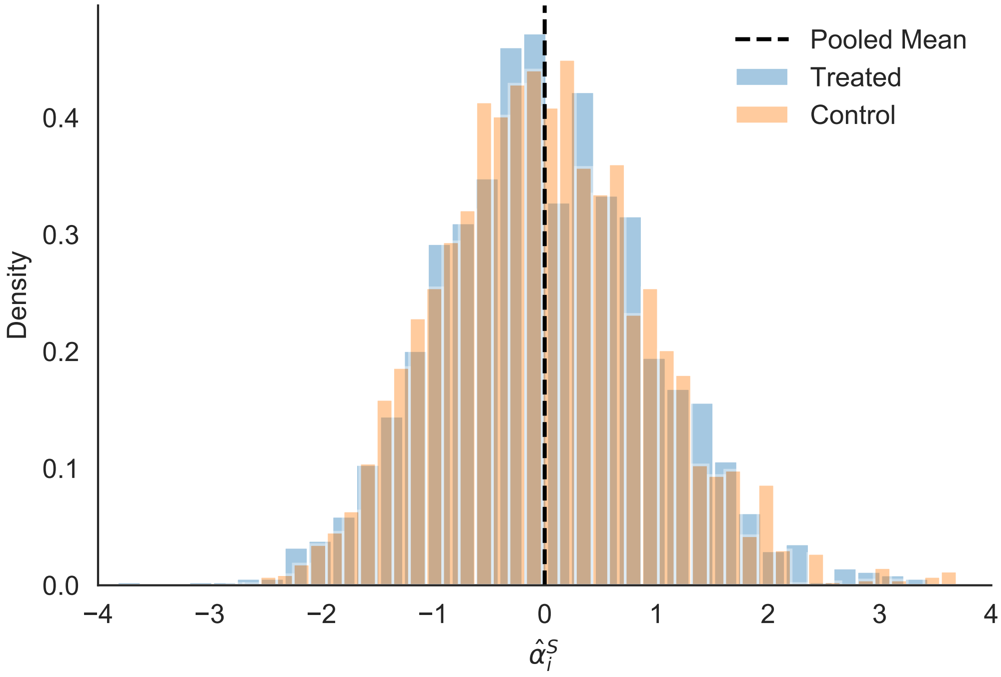

Figure 1: $\alpha_i^S$ in the experimental sample by treated and control

# 6 Conclusion

In this paper, we develop new statistical methods for systematically combining experimental and observational data in an attempt to leverage the internal validity of the experimental studies and the external validity and high precision of the observational studies. We do so in a setting where the experimental sample contains information on a secondary outcome and the observational study contains information on both primary and secondary outcomes. We articular a new and critical assumption that allows us to link the biases in comparisons in the observational study between primary and secondary outcome exploing the bias-free information on the secondary outcome from the experimental data. We illustrate these new results by combining data from the Project STAR experiment with observational data from the New York school system. We find that the biases in the observational study are substantial, but that the adjustment procedure based on the experimental data leads to more plausible results.

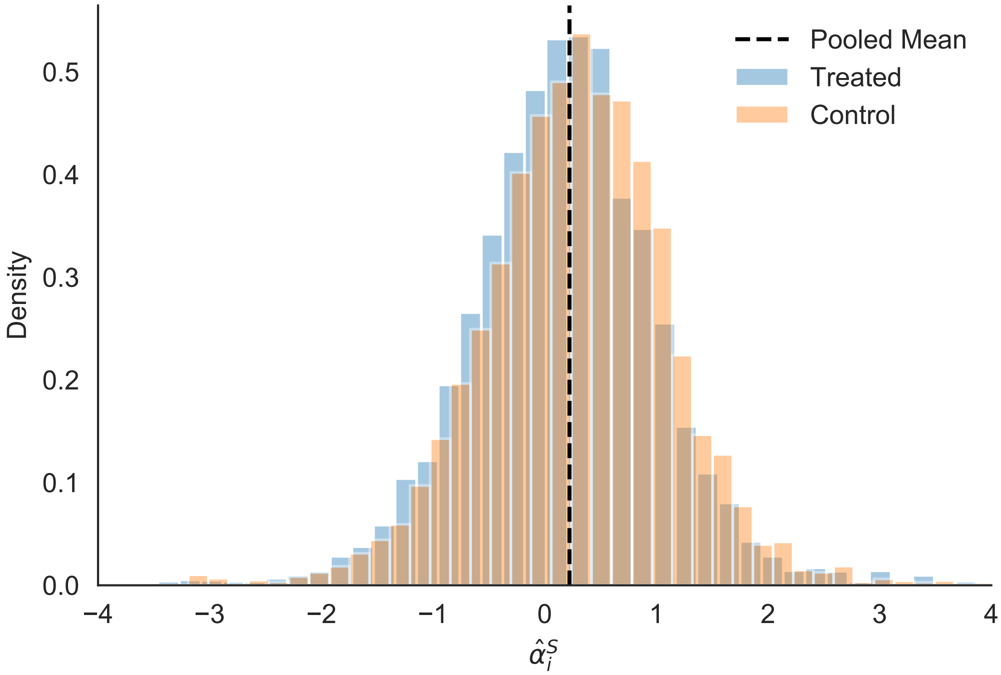

Figure 2: $\alpha_i^S$ in the observational sample by treated and control

Adams et al. [2006], D'Agostino et al. [2006], Abadie and Imbens [2016], Alonso et al. [2006]

## Appendix: Proofs of Results

**Proof of Lemma 1:** To prove this result we show that we cannot infer from the joint distribution of $(W_i, X_i, G_i, Y_i^{\mathrm{S}}, Y_i^{\mathrm{P}} 1_{G_i=\mathrm{O}})$, in combination with the assumptions, the distribution of $Y_i^{\mathrm{P}}(1)$ conditional on $X_i$ and $G_i = \mathrm{E}$. This distribution can be written as

$$f_{Y^{\mathrm{P}}(1)|X,G=\mathrm{E}}(y|x) = f_{Y^{\mathrm{P}}(1)|X,G=\mathrm{E},W=1}(y|x)p(W=1|X=x,G=\mathrm{E})$$

$$+f_{Y^{\mathrm{P}}(1)|X,G=\mathrm{E},W=0}(y|x)p(W=0|X=x,G=\mathrm{E}).$$

The data are not informative about the distribution of $Y_i^{\mathrm{P}}(1)$ given $W_i = 0$, $X_i$ and $G_i = \mathrm{E}$. Assumption 3 implies that this distribution is the same as the distribution of $Y_i^{\mathrm{P}}(1)$ given $W_i = 0$, $X_i$ and $G_i = \mathrm{O}$, but the data are not informative about that either. □

**Proof of Lemma 2:** To prove the result we show that

$$G_i \perp\!\!\!\perp Y_i^{\mathrm{S}}(1) \;\Big|\; X_i = x, W_i = 1.$$

We can factor the conditional distribution of $(Y_i^{\mathrm{S}}(1), G_i)$ given $X_i$ and $W_i = 1$ as

$$f(Y^{\mathrm{S}}(1), G|X, W=1) = f(Y^{\mathrm{S}}(1)|G, X, W)f(G|X, W=1).$$

By the unconfoundedness assumptions, Assumptions 2 and 4 it follows that this is equal to

$$f(Y^{\mathrm{S}}(1)|G, X)f(G|X, W=1).$$

By Conditional External Validity (Assumption 3) this is equal to

$$f(Y^{\mathrm{S}}(1)|X)f(G|X, W=1).$$

By Assumptions 2 and 4 this is equal to

$$f(Y^{\mathrm{S}}(1)|X, W=1)f(G|X, W=1),$$

which implies the conditional independence we set out to prove. □

**Proof of Theorem 1:**[1] To be clear here, we index the expectations operator by the random variable that the expectation is taken over. By definition

$$\tau_{\mathrm{O}}^{\mathrm{P}} = \mathbb{E}_{Y_i^{\mathrm{P}}(1),Y_i^{\mathrm{P}}(0)}\left[Y_i^{\mathrm{P}}(1) - Y_i^{\mathrm{P}}(0)\middle|\, G_i = \mathrm{O}\right] = \mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, G_i = \mathrm{O}\right] - \mathbb{E}_{Y_i^{\mathrm{P}}(0)}\left[Y_i^{\mathrm{P}}(0)\middle|\, G_i = \mathrm{O}\right].$$

We focus on identification of the first term, which by iterated expectations can be written as

$$\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, G_i = \mathrm{O}\right] = \mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, X_i, G_i = \mathrm{O}\right]\middle|\, G_i = \mathrm{O}\right]. \tag{A.1}$$

Identification of the second term follows by the same argument. By Conditional External Validity (Assumption 3), we can write the inner expectation as

$$\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, X_i, G_i = \mathrm{O}\right] = \mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, X_i, G_i = \mathrm{E}\right],$$

[1] We are grateful to Nathan Kallus and Xiaojie Mao for pointing out a mistake in an earlier version of the proof of this theorem.

so that (A.1) is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.2}$$

By iterated expectations this is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, Y_i^{\mathrm{S}}(1), X_i, G_i=\mathrm{E}\right]\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.3}$$

By Conditional External Validity (Assumption 3), this is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, Y_i^{\mathrm{S}}(1), X_i, G_i=\mathrm{O}\right]\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.4}$$

By Latent Unconfoundedness (Assumption 5) this is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}(1)\middle|\, Y_i^{\mathrm{S}}(1), W_i=1, X_i, G_i=\mathrm{O}\right]\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.5}$$

By the definitions $Y_i^{\mathrm{P}}=Y_i^{\mathrm{P}}(W_i)$ and $Y_i^{\mathrm{S}}=Y_i^{\mathrm{S}}(W_i)$ this is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[\mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}\middle|\, Y_i^{\mathrm{S}}, W_i=1, X_i, G_i=\mathrm{O}\right]\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.6}$$

Define

$$h(y^{\mathrm{S}}, x)\equiv \mathbb{E}_{Y_i^{\mathrm{P}}(1)}\left[Y_i^{\mathrm{P}}\middle|\, Y_i^{\mathrm{S}}=y^{\mathrm{S}}, W_i=1, X_i=x, G_i=\mathrm{O}\right],$$

so that (A.6) is

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[h(Y_i^{\mathrm{S}}(1), X_i)\middle|\, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.7}$$

Note that $h(y^{\mathrm{S}}, x)$ is directly identified from the observational sample.
Because of the unconfoundedness in the experimental sample (Assumption 2), (A.7) is equal to

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}(1)}\left[h(Y_i^{\mathrm{S}}(1), X_i)\middle|\, W_i=1, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.8}$$

By the definition of $Y_i^{\mathrm{S}}=Y_i^{\mathrm{S}}(W_i)$, and because the conditional distribution of $Y_i^{\mathrm{S}}(1)$ conditional on $W_i=1, X_i, G_i=\mathrm{O}$ is the same as the conditional distribution of $Y_i^{\mathrm{S}}$ conditional on $W_i=1, X_i, G_i=\mathrm{O}$, we can change the random variable that the expectation is taken over and write this as

$$\mathbb{E}_{X_i}\left[\mathbb{E}_{Y_i^{\mathrm{S}}}\left[h(Y_i^{\mathrm{S}}, X_i)\middle|\, W_i=1, X_i, G_i=\mathrm{E}\right]\middle|\, G_i=\mathrm{O}\right]. \tag{A.9}$$

The inner expectation

$$k(x)\equiv \mathbb{E}_{Y_i^{\mathrm{S}}}\left[h(Y_i^{\mathrm{S}}, X_i)\middle|\, W_i=1, X_i=x, G_i=\mathrm{E}\right],$$

is identified from the experimental sample. The expectation

$$\mathbb{E}[k(X_i)|G_i=\mathrm{O}],$$

is identified from the observational sample, which completes the proof. □

# References

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. Econometrica, 84(2):781–807, 2016.

Kenneth F Adams, Arthur Schatzkin, Tamara B Harris, Victor Kipnis, Traci Mouw, Rachel Ballard-Barbash, Albert Hollenbeck, and Michael F Leitzmann. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. New England Journal of Medicine, 355(8):763–778, 2006.

Ariel Alonso, Geert Molenberghs, Helena Geys, Marc Buyse, and Tony Vangeneugden. A unifying approach for surrogate marker validation based on prentice's criteria. Statistics in medicine, 25(2):205–221, 2006.

Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, 2008.

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of economic perspectives, 24(2):3–30, 2010.

Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. Econometrica, 74(2):431–497, 2006.

Susan Athey and Guido W Imbens. The econometrics of randomized experiments. Handbook of Economic Field Experiments, 1:73–140, 2017.

Susan Athey and Stefan Wager. Efficient policy learning. arXiv preprint arXiv:1702.02896, 2017.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

David Card. The impact of the mariel boatlift on the miami labor market. Industrial and Labor Relation, 43(2):245–257, 1990.

David Card and Alan Krueger. Minimum wages and employment: Case study of the fast-food industry in new jersey and pennsylvania. American Economic Review, 84(4):772–793, 1994.

Raj Chetty. Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. Annual Review of Economics, pages 451–488, 2009.

Ralph B D'Agostino, Michael J Campbell, and Joel B Greenhouse. Surrogate markers: back to the future: Special papers for the 25th anniversary of statistics in medicine. Statistics in medicine, 25(2):181–182, 2006.

Angus Deaton. Instruments, randomization, and learning about development. Journal of Economic Literature, 48(2):424–455, 2010.

Rajeev H Dehejia. Program evaluation as a decision problem. Journal of Econometrics, 125(1): 141–173, 2005.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. Handbook of development economics, 4:3895–3962, 2007.

Rachel Glennerster and Kudzai Takavarasha. Running randomized evaluations: A practical guide. Princeton University Press, 2013.

James Heckman and R. Robb. Alternative methods for evaluating the impact of interventions. Longitudinal analysis of labor market data, pages 156–245, 1985.

James J Heckman. Sample selection bias as a specification error. Econometrica, 47(1):153–161, 1979.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. Econometrica, 77(5):1683–1701, 2009.

V Joseph Hotz, Guido W Imbens, and Julie H Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics, 125(1): 241–270, 2005.

Guido Imbens. The role of the propensity score in estimating dose–response functions. Biometrika, 87(0):706–710, 2000.

Guido Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). Journal of Economic Literature, pages 399–423, 2010.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. Econometrica, 77(5):1481–1512, 2009.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv:2003.12408, 2020.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In Advances in Neural Information Processing Systems, pages 10888–10897, 2018.

Patrick Kline and Christopher R Walters. On heckits, late, and numerical equivalence. Econometrica, 87(2):677–696, 2019.

Alan B Krueger and Diane M Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. The Economic Journal, 111(468):1–28, 2001.

Roderick JA Little and Donald B Rubin. Statistical analysis with missing data, volume 793. Wiley, 2019.

Charles F Manski. Nonparametric bounds on treatment effects. The American Economic Review, 80(2):319–323, 1990.

Charles F Manski. Statistical treatment rules for heterogeneous populations. Econometrica, 72 (4):1221–1246, 2004.

Charles F Manski. Public policy in an uncertain world: analysis and decisions. Harvard University Press, 2013.

Fabrizia Mealli and Barbara Pacini. Using secondary outcomes and covariates to sharpen inference in instrumental variable settings. Journal of the American Statistical Association, 108: 1120–1131, 2013.

Magne Mogstad and Alexander Torgovitsky. Identification and extrapolation of causal effects with instrumental variables. Annual Review of Economics, 10:577–613, 2018.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. Econometrica, 86(5):1589–1619, 2018.

Judea Pearl, Elias Bareinboim, et al. External validity: From do-calculus to transportability across populations. Statistical Science, 29(4):579–595, 2014.

Geert Ridder and Robert Moffitt. The econometrics of data combination. Handbook of econometrics, 6:5469–5547, 2007.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Evan Rosenman, Art B Owen, Michael Baiocchi, and Hailey Banack. Propensity score methods for merging observational and experimental datasets. arXiv preprint arXiv:1804.07863, 2018.

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. arXiv preprint arXiv:2002.06708, 2020.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5):688, 1974.

Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

Donald B Rubin. Multiple imputation for nonresponse in surveys, volume 81. John Wiley & Sons, 2004.

William R Shadish, Thomas D Cook, and Donald T Campbell. Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company, 2002.

Jeffrey M Wooldridge. Control function methods in applied econometrics. Journal of Human Resources, 50(2):420–445, 2015.

J.M. Wooldridge. Econometric Analysis of Cross Section and Panel Data. The MIT Press. ISBN 9780262232586.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. arXiv preprint arXiv:1810.04778, 2018.