The Experimental Selection Correction Estimator: Using Experiments to Remove Biases in Observational Estimates*

Susan Athey† Raj Chetty‡ Guido W. Imbens§

First Version: August 2019; Current version: May 2025

Abstract

Researchers increasingly have access to two types of data: (i) large observational datasets where treatment (*e.g.*, class size) is not randomized but several primary outcomes (*e.g.*, graduation rates) and secondary outcomes (*e.g.*, test scores) are observed and (ii) experimental data in which treatment is randomized but only secondary outcomes are observed. We develop a new method to estimate treatment effects on primary outcomes in such settings. We use the difference between the secondary outcome and its predicted value based on the experimental treatment effect to measure selection bias in the observational data. Controlling for this estimate of selection bias yields an unbiased estimate of the treatment effect on the primary outcome under a new assumption that we term *latent unconfoundedness*, which requires that the same confounders affect the primary and secondary outcomes. Latent unconfoundedness weakens the assumptions underlying commonly used surrogate estimators. We apply our estimator to identify the effect of third grade class size on students' outcomes. Estimated impacts on test scores using OLS regressions in observational school district data have the opposite sign of estimates from the Tennessee STAR experiment. In contrast, selection-corrected estimates in the observational data replicate the experimental estimates. Our estimator reveals that reducing class sizes by 25% increases high school graduation rates by 0.7 percentage points. Controlling for observables does not change the OLS estimates, demonstrating that experimental selection correction can remove biases that cannot be addressed with standard controls.

Keywords: causality, surrogates, observational studies, long-term outcomes, control functions

*An earlier version was circulated under the title "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes." We thank Kevin Chen and Yechan Park for excellent research assistance. We are grateful for comments from Gary Chamberlain, Nathan Kallus, Xiaojie Mao, David Ritzwoller, Jonah Rockoff, Dylan Small, and numerous seminar participants. This research was funded through the Chan-Zuckerberg Initiative, Sloan Foundation, Schmidt Futures, ONR grant N00014-17-1-2131 and N00014-19-1-2468, Harvard University, and a gift from Amazon.

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

‡Department of Economics, Harvard University, Opportunity Insights, and NBER, chetty@g.harvard.edu.

§Graduate School of Business, Stanford University and NBER, imbens@stanford.edu.

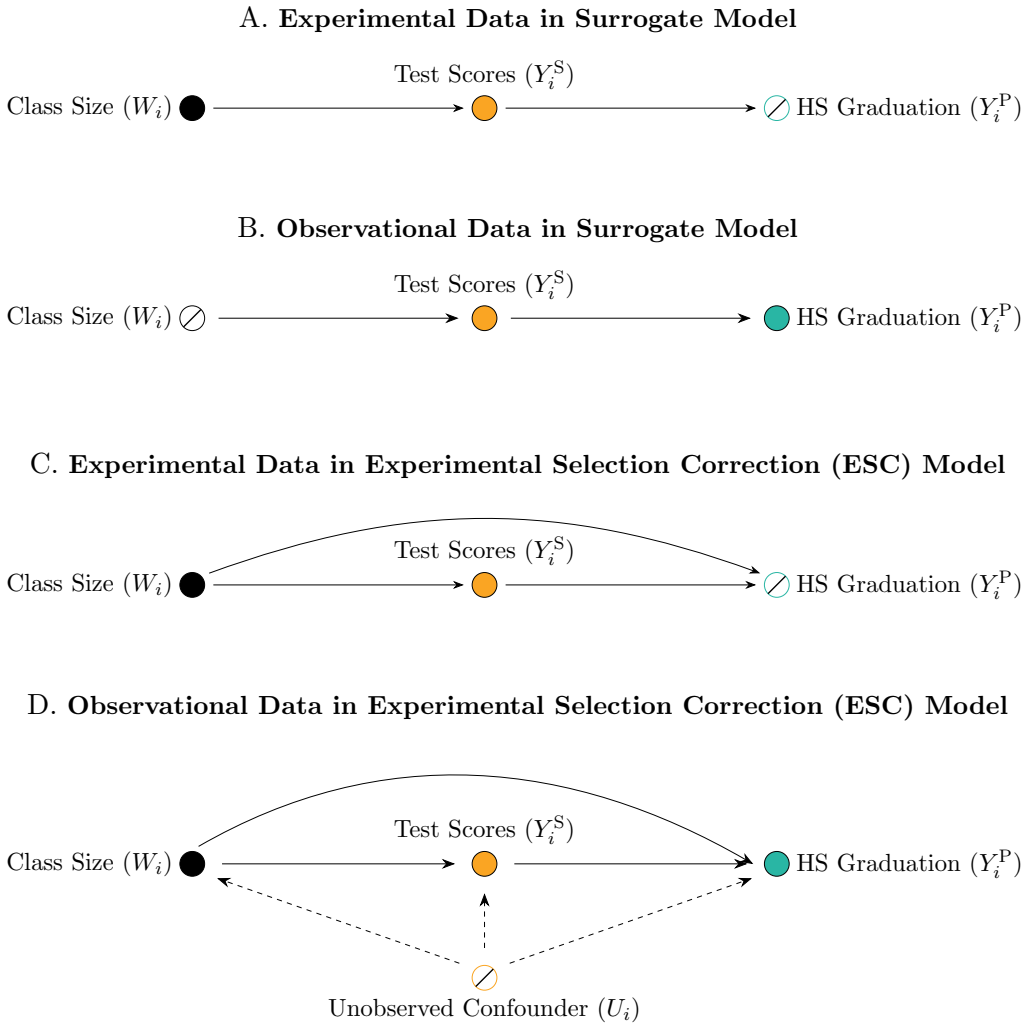
1 Introduction

As observational data become more widely available, researchers seeking to estimate treatment effects increasingly have access to two types of data: (i) large observational datasets where treatments and a broad range of outcomes are observed, but treatment is not randomized and (ii) smaller experimental datasets where treatment is randomly assigned, but only a subset of outcomes are observed. For example, in the context of education, many analysts have been interested in identifying the causal effects of classroom sizes in elementary school on high school graduation rates. Observational data with information on class sizes and graduation rates are now widely available from school districts' administrative records. But causal inference using these data is challenging because of selection biases arising from non-random assignment to classrooms. Causal inference is more straightforward in experimental data – such as the widely studied Project STAR class size experiment (e.g., Krueger [1999]) – but experimental datasets often do not contain information on outcomes such as graduation rates because they are observed with long lags.

The most common method of identifying the causal effect of a treatment (*e.g.*, class size reduction) on the primary outcome of interest (*e.g.*, graduation rates) in such settings is to use secondary intermediate outcomes that are observed in the experimental data (*e.g.*, test scores) as statistical surrogates [Prentice, 1989, Athey et al., 2019]. The surrogate approach, illustrated in Figures 1a-b below, uses the observational dataset to estimate the relationship between the primary outcome (Y_i^{P}) and secondary outcome (Y_i^{S}), and then estimates the impact of the treatment of interest (W_i) on Y_i^{P} based on that relationship. Under the surrogacy assumptions that (i) W_i only affects Y_i^{P} through its impact on Y_i^{S} and (ii) there are no unobserved confounders that affect the relationship between Y_i^{S} and Y_i^{P} in the observational sample, this approach provides an unbiased estimate of the effect of W_i on Y_i^{P}.

The surrogate approach has been applied in many fields, from economics to product testing to public health [Alonso et al., 2006, Adams et al., 2006, D'Agostino et al., 2006, Gupta et al., 2019]. Yet there remains concern that the surrogacy assumptions may be violated in these settings. For example, test scores are a widely used surrogate in labor economics, but researchers have identified other pathways through which childhood interventions affect long-term outcomes outside test scores, such as non-cognitive skills [Heckman et al., 2006, Chetty et al., 2011].

Figure 1: Comparison of Experimental Selection Correction and Surrogate Models



Notes: This figure depicts the assumptions and informational structures underlying the surrogate and Experimental Selection Correction (ESC) estimators using directed acyclic graphs. Solid circles denote variables observed in the data; empty circles with lines denote unobserved variables. Panels A and B depict the experimental and observational data in the surrogate approach, while Panels C and D show the same for the ESC estimator. In the experimental data (Panels A and C), class size (W_i) is randomly assigned, ensuring that there is no unobserved confounder when estimating the treatment effect on test scores (Y_i^S). However, the primary outcome of interest, high school graduation (Y_i^P), is unobserved in the experimental sample. The ESC estimator requires that W_i is observed in the observational sample, whereas the surrogate estimator does not. The identifying assumptions underlying the surrogate estimator are that (i) any effect of W_i on the primary outcome Y_i^P operates exclusively through the secondary outcome Y_i^S (Panels A and B) and (ii) there are no unobserved confounders that affect the relationship between Y_i^S and Y_i^P in the observational sample (Panel C). The ESC estimator relaxes these assumptions by (i) permitting a direct effect of W_i on Y_i^P (Panels C and D) and (ii) allowing for an unobserved confounder (U_i) that influences both class size and graduation outcomes in the observational data (Panel D).

In this paper, we develop an "Experimental Selection Correction" (ESC) estimator that identifies the effect of W_i on Y_i^{P} even when the surrogacy assumptions are violated, as illustrated in Panels C and D of Figure 1. Our estimator relies on more information than the surrogacy approach: it requires that the observational dataset contains information not just on the primary and secondary outcomes, Y_i^{S} and Y_i^{P}, but also on treatment W_i (with variation in treatment across observations). With this additional information, we show how one can identify the effect of W_i on Y_i^{P} under strictly weaker assumptions than those required for the surrogacy approach.

To illustrate the general information scheme we analyze, consider a setting with two datasets: (i) the Project STAR experimental data, where class size is randomized and we observe test scores (Y_i^{S}) but not high school graduation rates (Y_i^{P}), and (ii) observational data from the New York City school district, in which class size is observed but not randomized (and hence likely to be correlated with both observed and unobserved characteristics) and we observe both test scores and graduation rates.

In the STAR experimental data, we can estimate the treatment effect of small class size (W_i) on 3rd grade test scores by regressing test scores on an indicator for being assigned to a small class (with 7 fewer students on average) in 3rd grade. Column 1 of Table 1 shows that being assigned to a small class in 3rd grade increases students' end-of-3rd-grade test scores by 0.19 standard deviations (SD). We cannot, however, estimate the effect of class size on high school graduation in the STAR data, because we do not observe graduation in the STAR sample.[1]

In the observational NYC sample, estimating an analogous OLS regression of test scores on an indicator for being assigned to a small class yields an estimate of -0.12 SD (s.e. 0.01, Column 2 of Table 1). Children in smaller classes are also 1.76 percentage points (s.e. 0.29) less likely to graduate from high school. These negative estimates of the causal effect of class size reductions are implausible both in the light of the positive experimental Project STAR estimates and based on *a priori* beliefs. Of course, the OLS estimates may be confounded because class size is not randomly assigned in NYC. For example, students with needs for additional educational support may be assigned to smaller classes. Our goal is to obtain an unconfounded estimate of W_i on Y_i^{P}, *i.e.*, to fill in the lower left box in Table 1.

[1]Researchers attempted to follow the STAR students longitudinally, but were only able to collect information on high school graduation for 43% of students, whose characteristics are not representative of the experimental sample as a whole. This underscores the challenges of tracking primary outcomes in experiments and motivates the approach we take here of combining observational administrative records and experimental data.

Table 1: Estimated Effects of Small Class Assignment in STAR vs. NYC Data

Sample:	Exp. (STAR)	Obs. (New York)	Exp. + Obs. (STAR + NYC)
Estimator:	OLS	OLS	Exp. Selection Correction (ESC)
Outcome			
3rd Grade Test Score	0.19	−0.12	0.19
	(0.04)	(0.01)	(0.04)
(secondary outcome)			
HS Graduation	?	−1.76	0.69
		(0.29)	(0.34)
(primary outcome)			

Notes: This table reports point estimates (with standard errors in parentheses) of the effect of assignment to a small class on end-of-3rd-grade test scores and high school graduation rates. Each cell reports estimates from a separate model. Columns 1 and 2 present ordinary least squares (OLS) estimates using the experimental STAR sample and the observational New York sample, respectively. Column 3 combines both experimental and observational data using the Experimental Selection Correction (ESC) estimator. The missing entry denoted by "?" for High School Graduation in the STAR sample reflects the absence of information on graduation in the STAR dataset. The specification in Column 1 includes school fixed effects (since all STAR students are in the same cohort), while Columns 2 and 3 include both school and cohort fixed effects.

We obtain an unconfounded estimate of the effect of W_i on Y_i^{P} in the observational data by using the difference in the distribution of test scores (Y_i^{S}) conditional on treatment in the observational and experimental samples to adjust for selection. In linear models, the ESC estimator can be implemented in three straightforward steps (see Appendix for code). First, we estimate the effect of class size on test scores (τ^{S}) in the experimental sample using a linear regression, as in Column 1 of Table 1.[2] Second, for all students in the observational sample, we calculate the difference between the secondary outcome (test score) and the predicted test score based on the student's class size (the residual $\alpha_i^{\mathrm{S}} = Y_i^{\mathrm{S}} - \tau^{\mathrm{S}} W_i$), with the parameter τ^{S} in the prediction model coming from the experimental sample. Finally, we regress the primary outcome (graduation rates) on treatment (class size) in the observational data, controlling for the residual α_i^{S}.

We show that this control function approach identifies the causal effect of W_i on Y_i^{P} under three assumptions: (i) random assignment (or unconfoundedness) in the experimental sample;

[2]We focus on the case where treatment is randomly assigned in an experiment, but any quasi-experimental research design that yields an unbiased estimate of the treatment effect on the secondary outcome τ^{S} can be used to implement the ESC estimator.

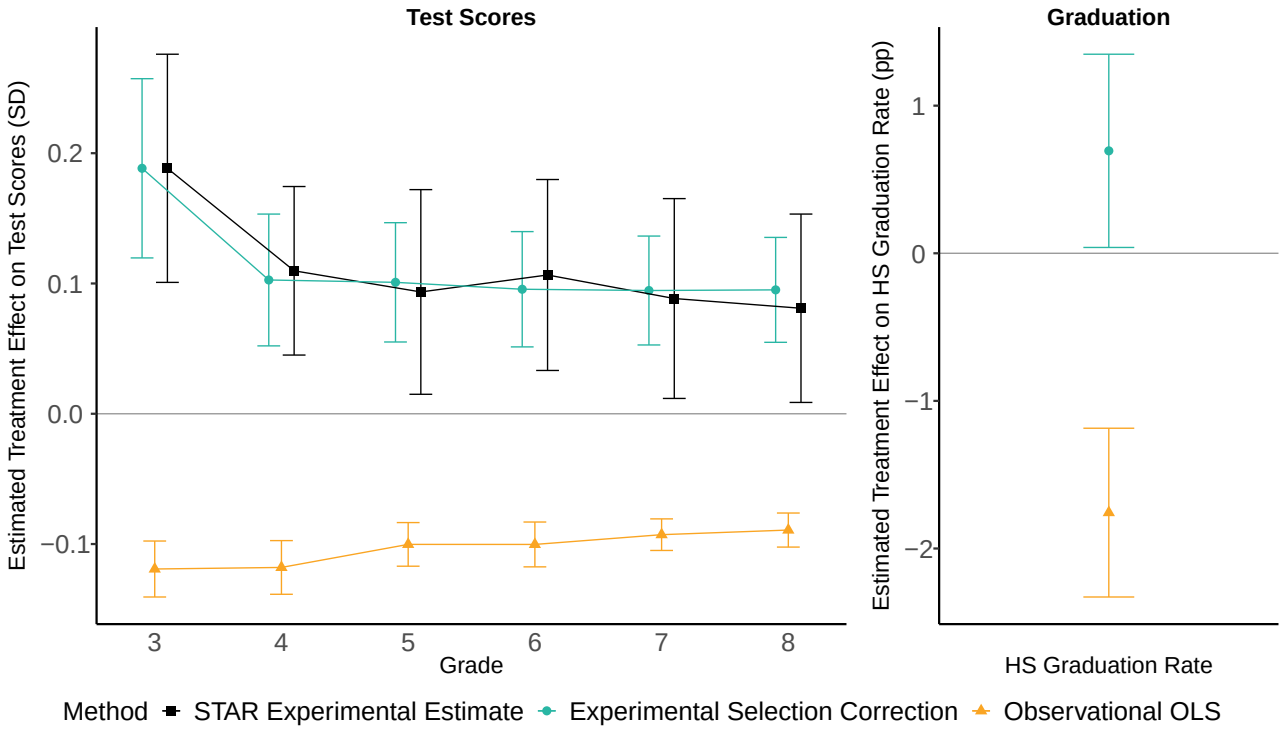
(ii) a standard external validity assumption; and (iii) a new assumption that we term *latent unconfoundedness*. External validity requires that (conditional on pretreatment observables), the treatment effect in the experimental sample is the same as the treatment effect in the population represented by the observational sample [Shadish et al., 2002, Hotz et al., 2005]. Latent unconfoundedness requires that the unobserved confounders that affect the primary outcome (graduation rates) are the same as those that affect the secondary outcome (test scores). Under this assumption, the difference between the actual secondary outcome in the observational data and the predicted secondary outcome based on the experimental estimate (α_i^{S}) fully captures any selection bias that affects the primary outcome. Thus, controlling for α_i^{S} is sufficient to identify the causal effect of W_i on Y_i^{P}. Intuitively, α_i^{S} functions as a selection correction, similar to parametric selection correction approaches dating to Heckman [1979] and control function methods (Heckman and Robb [1985], Imbens and Newey [2009], Wooldridge [2015]).

The main theoretical result of this paper is that the treatment effect of W_i on Y_i^{P} is point-identified under latent unconfoundedness, external validity, and random assignment in the experimental sample (without any functional form or distributional assumptions). We also present a control function approach to estimation for the general, nonlinear case. A corollary of our main result is that if an observational estimate of the treatment effect on the secondary outcome (Y_i^{S}) is the same as the experimental estimate, then under linearity, latent unconfoundedness and external validity together imply that the observational estimator is unconfounded for the primary outcome. Many empirical studies show that an observational and experimental estimator yield similar estimates for secondary outcomes and then use this as a heuristic justification for estimating impacts on a broader range of primary outcomes using the observational estimator (*e.g.*, Chetty et al. [2014], Bleemer [2022]). Our analysis makes precise the conditions—most importantly, latent unconfoundedness across outcomes—under which this heuristic is justified.

We also propose falsification tests for the underlying identifying assumptions that make use of additional "holdout" post-treatment outcomes (Y_i^H) observed in both datasets. One could use these measures as additional secondary outcomes. An alternative is to use them to validate the estimator instead of using them to implement the selection correction itself. Tests of whether our ESC estimator that uses 3rd grade scores for selection correction matches experimental estimates for 4th and 5th grade test scores (holdout outcomes) serve as tests for whether the underlying latent unconfoundedness and external validity assumptions jointly hold.

We apply the experimental selection correction estimator to estimate the causal effects of 3rd grade class size reduction on high school graduation rates in the New York City data, using end-of-3rd-grade test scores as the secondary outcome for the selection correction. We use the Tennessee STAR sample as the experimental sample in which we estimate the treatment effect of class size reduction on 3rd grade test scores.[3]

Figure 2: Estimated Treatment Effects of Assignment to Small Class in 3rd Grade



Notes: This figure plots estimates of the effect of being assigned to a small class in 3rd grade on two outcomes: standardized test scores (left panel) and high school (HS) graduation rates (right panel). In the left panel, effect sizes are reported in standard deviations of test scores; in the right panel, they are reported in percentage points of HS graduation. The "Experimental Estimate" series presents estimates from OLS regressions with school fixed effects in the STAR data. The "Observational OLS" estimate is obtained from OLS regressions with school and cohort fixed effects in the NYC observational data. The "Experimental Selection Correction" (ESC) estimate adjusts the observational OLS estimate using a selection correction term estimated using the experimental sample as described in the text. Vertical bars denote 95% confidence intervals.

[3]Naturally, one may have concerns about the external validity of the Tennessee sample for the New York City data. Both samples reflect a relatively low-income population with fairly similar demographic characteristics. Furthermore, we show that adjusting for remaining differences in observable demographics does not affect our conclusions meaningfully.

As discussed above, standard OLS regression estimates of 3rd grade test scores on an indicator for small class size in the NYC data yield a treatment effect estimate of -0.12 SD. The ESC estimator (shown in Column 3 of Table 1) yields an estimate of 0.19 SD, coinciding with the STAR experimental estimate by construction. The ESC treatment effect estimates for test scores in grades 4-8 (holdout outcomes) are nearly identical to the STAR experimental estimates (Figure 2). Notably, they capture the well-known "fadeout" pattern on test score impacts documented in prior work (Deming [2009], Chetty et al. [2011], Cascio and Staiger [2012]). These results support the identification assumptions underlying our method and more broadly serve to validate the ESC approach.

Finally, the ESC estimator implies that assignment to a small class in 3rd grade (which has 25% fewer students on average) increases the probability of graduating from a New York City public high school by 0.69 percentage points (pp), relative to a sample mean of 51.4%. This estimate is one of the first estimates of the causal effect of class size reduction on high-school graduation rates in the U.S.

In contrast, the standard OLS estimator yields significant negative estimates on test scores in later grades and on high school graduation rates. When we control for observable characteristics, the OLS estimates remain negative, while the ESC estimates remain similar to the experimental estimates (Figure 3). These findings demonstrate how our proposed experimental selection correction can detect and adjust for selection biases that are difficult to address with conventional methods in observational data without relying on strong surrogacy assumptions.

In addition to the literature on statistical surrogates, our analysis relates to other studies that have examined similar observation schemes, including Rosenman et al. [2018, 2020], Kallus and Mao [2020], Mealli and Pacini [2013] and Imbens et al. [2025]. Rosenman et al. [2018] focuses on the problem where assignment is unconfounded in both samples and combining the samples increases precision. Rosenman et al. [2020] allow for unobserved confounders in the observational sample and consider shrinkage estimators to decrease bias. Kallus and Mao [2020] analyze the case where assignment in the combined experimental and observational sample is unconfounded, but not in each sample separately. Kallus et al. [2018] focus on a case where the same variables (including the primary outcome) are observed in the two samples, but where unconfoundedness is violated in the observational sample and the experimental sample is used to estimate bias. Mealli and Pacini [2013] focuses on an instrumental variables setting where the presence of multiple

outcomes improves estimates. Our approach also relates to the Changes-in-Changes estimator 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. Under our maintained assumptions here, differences in the estimated effect of the treatment between the experimental and observational sample are attributed to violations of unconfoundedness in the observational data.

The paper is organized as follows. Section 2 analyzes a linear model that captures the intuition underlying our approach. Section 3 presents the identification result for the general case. Section 4 discusses estimation. Section 5 presents the application. Section 6 concludes.

2 Linear Models

In this section, we introduce our key identifying assumption and a control function estimator in the context of linear models. The linear case simplifies exposition and captures the key ideas that apply in more general models.

2.1 Setup

Using the potential outcome set up for observational studies introduced by Rubin [1974] (see Imbens and Rubin [2015] for a textbook discussion), let the pair of potential outcomes for the primary outcome for unit i be denoted by $Y_i^{\mathrm{P}}(0)$ and $Y_i^{\mathrm{P}}(1)$, where the superscript "P" stands for "Primary". In many applications, Y_i^{P} is a long-term outcome; in our application, it is a binary indicator for high school graduation. The treatment received by unit i is $W_i \in \{0, 1\}$. In our application, W_i is an indicator for small class size in third grade, with $W_i = 1$ indicating a small class size and $W_i = 0$ indicating a regular class size. There is also a secondary outcome, with the pair of potential outcomes for unit i denoted by $Y_i^{\mathrm{S}}(0)$ and $Y_i^{\mathrm{S}}(1)$, where the superscript "S" stands for "Secondary". In our application, Y_i^{S} is a student's end-of-third-grade test score.

We focus in this section on the case where both the primary and secondary outcomes are scalars, but both may be vector-valued (*e.g.,* test scores in multiple grades could be used as secondary outcomes).

The realized values for the primary and secondary outcomes are $Y_i^{\mathrm{P}} \equiv Y_i^{\mathrm{P}}(W_i)$ and $Y_i^{\mathrm{S}} \equiv Y_i^{\mathrm{S}}(W_i)$. We may also observe pretreatment variables, denoted by X_i, that are known not to be

affected by the treatment.

We focus on identifying the average treatment effect on the primary outcome,

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

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

We have two samples to draw on to estimate τ^{P}, as in the literature on combining datasets, *e.g.,* Hotz et al. [2005], Ridder and Moffitt [2007], Pearl et al. [2014]. The first is an observational study that 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 random assignment of treatment W_i. For all units in this experimental sample, we observe the triple $(W_i, Y_i^{\mathrm{S}}, X_i)$, but not the primary outcome Y_i^{P}.

Let $G_i \in \{\mathrm{E}, \mathrm{O}\}$, be an indicator for the sample a unit is drawn from. Then we can conceptualize the combined sample as a random sample of size N from an artificial super-population for which we observe the quintuple $(W_i, G_i, Y_i^{\mathrm{S}}, Y_i^{\mathrm{P}}\mathbf{1}_{G_i=\mathrm{O}}, X_i)$, 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 Control Function Estimator

Suppose we have a linear model for the secondary potential outcomes in combination with a constant treatment effect τ^{S}:

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

This model holds in both the experimental and observational samples. However, the properties of the unobserved component α_i^{S} differ between the two samples. In the experimental sample, randomization guarantees the following conditional independence condition:[4]

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

[4]In fact the randomization implies an even stronger condition, $W_i \perp\!\!\!\perp \alpha_i^{\mathrm{S}}, X_i | G_i = \mathrm{E}$, but we do not need that condition here.

In the observational study, the same conditional independence does not generally hold:

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

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 unobserved component might be correlated with the treatment in the observational sample:

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

that is, W_i is again endogenous.

To identify the treatment effect on the primary outcome τ^{P} in the observational sample, we make the following assumption that links the endogeneity problems for the primary and secondary outcomes:

$$\alpha_i^{\mathrm{P}} = \delta \alpha_i^{\mathrm{S}} + \varepsilon_i^{\mathrm{P}}, \qquad \text{with } W_i \perp\!\!\!\perp \varepsilon_i^{\mathrm{P}} \;\Big|\; X_i, \alpha_i^{\mathrm{S}}, G_i = \mathrm{O}. \tag{2.2}$$

This assumption requires that the component of the residual in the primary outcome that is not explained by the residual in the secondary outcome, $\varepsilon_i^{\mathrm{P}} \equiv \alpha_i^{\mathrm{P}} - \mathbb{E}[\alpha_i^{\mathrm{P}} | \alpha_i^{\mathrm{S}}]$, is orthogonal to treatment. The key substantive restriction captured by this condition is that the unobserved confounders that affect the secondary outcome are the same as those that affect the primary outcome. This assumption, which we term *latent unconfoundedness*, is the key to identifying τ^{P} in the general case below as well.

We now show how this latent unconfoundedness assumption allows us to identify τ^{P} using a simple control function approach in the linear case. First, we exploit randomization in the experimental sample to estimate τ^{S} and γ^{S} using ordinary least squares regression. Denote these least squares estimates by $\hat{\tau}^{\mathrm{S}}$ and $\hat{\gamma}^{\mathrm{S}}$.

Next, we estimate the residual α_i^{S} for the units in the observational sample as

$$\hat{\alpha}_i^{\mathrm{S}} = Y_i^{\mathrm{S}} - W_i \hat{\tau}^{\mathrm{S}} - X_i^\top \hat{\gamma}^{\mathrm{S}}. \tag{2.3}$$

If the assignment to treatment in the observational sample were random (and assuming the linear model is correct), the population value of these residuals α_i^{S} would be uncorrelated with the treatment indicator in the observational sample.

When treatment assignment is non-random, we can use the association between the secondary outcome residuals α_i^{S} and the treatment to correct for selection bias in the estimating equation for the primary outcome. We do so by including α_i^{S} as a control variable in an ordinary least squares regression of the primary outcome on treatment. To see why this yields a consistent estimate of τ^{P}, observe that we can use the linear representation in (2.2) to write the primary outcome as:

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

Because the error term $\varepsilon_i^{\mathrm{P}}$ is orthogonal to treatment in this specification, estimating this equation using OLS yields a consistent estimator for τ^{P} under our assumptions.

3 The General Case

In this section, we generalize the linear example above to accommodate (i) non-linear models and (ii) multiple secondary outcomes.

We are interested in causal estimands defined for the population of interest. 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]). For expositional simplicity, we focus here on average treatment effects. Define

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

to be 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 primary 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.*

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 $w = 0, 1$,*

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

Although internal validity of the experimental sample is guaranteed by design, external validity of the experimental study does not follow. 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*

$$G_i \perp\!\!\!\perp \left(Y_i^{\rm P}(0), Y_i^{\rm P}(1), Y_i^{\rm S}(0), Y_i^{\rm S}(1)\right) \,\Big|\, X_i. \tag{3.4}$$

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 imply non-trivial 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* $\tau^{\rm P}$.

The proof for this result is given in the appendix.

Next, let us briefly mention a common assumption that we do *not* wish to make in this context. Specifically, we consider the assumption that assignment in the observational study is unconfounded. For $w = 0, 1$,

$$W_i \perp\!\!\!\perp \left(Y_i^{\rm S}(w), Y_i^{\rm P}(w)\right) \,\Big|\, X_i, G_i = {\rm O}, \tag{3.5}$$

This assumption is made, for example, in Rosenman et al. [2018]. This assumption is sufficient for identification of τ, 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.

3.2 Latent Unconfoundedness

Suppose that we reject the combination of Assumptions 2-3 and unconfoundedness (3.5). If we maintain unconfoundedness in the experimental sample (Assumption 2), it must be that either conditional external validity in the experimental study (Assumption 3), or unconfoundedness in the observational study (3.5) must be violated. In many cases we may wish to maintain conditional external validity and interpret the finding that the combination does not hold as evidence that unconfoundedness in (3.5) does not hold for the observational study.

The fundamental idea behind our approach, although not the implementation, can be seen as related 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 between adjusted comparisons for the primary and secondary outcomes, is the following.

Assumption 4. (Latent Unconfoundedness)
For $w \in \{0,1\}$,

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

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

Remark 1. *Compared to a regular unconfoundedness assumption, we add the variable $Y_i^{\rm S}(w)$ 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 an unconfoundedness assumption to create subpopulations defined by the conditioning variables, and then compare treated and control units within those subpopulations. To be specific, suppose we wish to estimate $\mathbb{E}[Y_i^{\rm P}(1)|G_i = {\rm O}]$. We would first estimate the conditional expectation $\mu(y^{\rm S}, x) = \mathbb{E}[Y_i^{\rm P}(1)|Y_i^{\rm S}(1) = y^{\rm S}, W_i = 1, X_i = x, G_i = {\rm O}]$. Then, in the second step, we would average this over the marginal distribution of $(Y_i^{\rm S}(1), X_i)$ in the observational sample. However, in the observational sample we only see draws from the conditional distribution of $(Y_i^{\rm S}(1), X_i)$ 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.*

Remark 2. *The precise version of the unconfoundedness assumption here is slightly different from than the (stronger) unconfoundedness assumption in, say, Rosenbaum and Rubin [1983] where it is assumed that W_i is independent of the full set of $Y_i^{\rm P}(0), Y_i^{\rm P}(1)$). It is what is referred to in Imbens [2000] as "weak unconfoundedness."*

To highlight the link to the control function literature (Heckman [1979], Heckman and Robb [1985], Imbens and Newey [2009], Wooldridge [2010], 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 potential outcomes as

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

with the function $h^{\rm S}(w, \eta, x)$ strictly monotone in η. In the context of this model we can write the latent unconfoundedness assumption as

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

Although it is not generally true that $W_i \perp\!\!\!\perp \nu_i | X_i, G_i = \text{O}$ (without conditioning on η_i), adding η_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 nonparametric instrumental variables setting (*e.g.,* Imbens and Newey [2009]), where the two models are

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

with $r(z, \eta, x)$ strictly monotone in η. 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, $Y_i^{\text{P}}(w) = h^{\text{P}}(w, \nu_i, X_i)$. In both cases we address the endogeneity by conditioning on an additional variable, the control variable η_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 Imbens and Newey [2009]. In the Imbens-Newey nonparametric 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 comparison shows one of the limitations of the approach: the unobserved confounder η_i cannot have a dimension higher than that of the secondary outcome.

Formally, adding Assumption 4 (latent unconfoundedness) to Assumptions 1-3 allows us to point-identify the average effect of interest. The following theorem states our main identification result.

Theorem 1. *Suppose that Assumptions 1-4 hold, so that the experimental study is unconfounded and has conditional external validity, and the observational study has latent unconfoundedness. Then the average effect of the treatment on the primary outcome in the observational study is point-identified.*

3.3 Missing At Random

There is an interesting connection between Assumptions 1-4 and the Missing At Random (MAR) assumption in the missing data literature (Rubin [1976], Little and Rubin [2019], Rubin [1987]).

Lemma 2. *Suppose that Assumptions 1-4 hold. Then:*

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

Because $G_i = {\rm E}$ is equivalent to an indicator that $Y_i^{\rm P}$ missing, and because W_i, X_i, and $Y_i^{\rm S}$ are observed for all individuals in the sample, the conditional independence in (3.7) 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 1-3 and 4 does imply some inequality restrictions on the joint distribution of the observed variables. Kallus and Mao [2020] starts with a MAR 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 discuss estimation and inference. There are multiple approaches here, some of which we discussed in the examples in Section 2. These strategies include imputation, weighting, control function methods, and influence-function based methods. Because the model is just-identified, all four of these methods are first-order equivalent, although they will have different finite sample properties, see Newey [1994], Chen and Santos [2018] for a general discussion.

We focus here on the control function approach that is special to this setting with observational and experimental data. In the appendix, we discuss the imputation, weighting, and influence function approaches, which closely resemble their equivalents in standard unconfoundedness settings.

In the control function approach, we directly estimate the unobserved confounder. We then estimate the average treatment effect in the observational sample adjusting for both the observed covariates and the estimated confounder. The procedure consists of three steps.

In the first step, we estimate the conditional cumulative distribution function of the secondary outcome, conditional on the treatment and pre-treatment variables, in the experimental sample:

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

Note that if the secondary outcome is a vector, this is a vector of conditional cumulative distribution functions, demonstrating that multiple secondary outcomes can weaken the identifying assumptions.

In the second step, we calculate for all units in the observational sample the control variable as

$$\eta_i = F_{Y^{\rm S}|W,X}(Y_i^{\rm S}|W_i, X_i).$$

In the third step, we estimate the adjusted difference

$$\mathbb{E}\left[\mathbb{E}\left[Y_i^{\rm P}\middle|\, W_i = 1, X_i, G_i = {\rm O}\right] - \mathbb{E}\left[Y_i^{\rm P}\middle|\, W_i = 0, X_i, G_i = {\rm O}\right]\middle|\, G_i = {\rm O}\right],$$

which by the assumptions in Theorem 1 is equal to the average causal effect τ. Here we can use any of the conventional methods for estimating average treatment effects under unconfoundedness for the observational data (including matching, regression, inverse propensity score weighting, augmented inverse propensity score weighting, or doubly robust methods), where we use the combination of the pretreatment variables X_i and the estimated control function η_i as the variables to be adjusted for. For example, using a imputation/regression approach, one would estimate the conditional mean of the primary outcome in the observational sample given treatment status, control variable, and pre-treatment variables:

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

These estimated conditional means would then be use to estimate the average treatment effect τ as

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

where $N_w^g = \sum_{i=1}^N 1_{W_i=w, G_i=g}$. Under standard conditions (*e.g.,* Newey [1994], Chen and Santos [2018]) this estimator will be semiparametrically efficient and asymptotically linear and normally distributed. From Newey [1994] it follows that the control function estimator is first order equivalent to the imputation, weighting, and influence function estimators, and so one can use the semiparametric efficiency bound for variance estimation. Alternatively, one can use the regular bootstrap.

5 Application: Effects of Class Size on Graduation

In this section, we evaluate the performance of our approach by estimating the long-term impacts of reducing class sizes in elementary school. Many experimental and quasi-experimental studies have analyzed the impacts of educational inputs – such as class size, teacher quality, and resources – on test scores [Krueger, 1999, Kane and Staiger, 2008, Biasi et al., 2025]. Meanwhile, observational data with information on educational inputs as well as long-term outcomes such as high school graduation rates from school districts' administrative records have become widely available. We combine these two sets of information to estimate the effects of class size on high school graduation rates.[5] We first describe the data we use and then present results.

5.1 Data

We combine information from two datasets: experimental data from Tennessee STAR and observational data from the New York City public school district.

Experimental Sample: Tennessee STAR. The STAR experiment was conducted at 79 low-income public schools in Tennessee between 1985-89. In the 1985-86 school year, 6,323 kindergarten students in participating schools were randomly assigned to a small (target size 13-17 students) or regular-sized (20-25 students) class within their schools. An additional 5,248 children joined the 1985-86 entry cohort at the participating schools after kindergarten in grades 1-3. These new entrants were also randomly assigned to small vs. large classrooms within school upon entry. Students were intended to remain in the same class type (small vs. large) through 3rd grade, at which point all students returned to regular class sizes.

In each year from grades 3-8, STAR students were administered standardized tests that measure performance in math and reading. We standardize the average of math and reading scores to have mean 0 and standard deviation 1 within each grade among students in the STAR sample. We also observe information on students' race and ethnicity, sex, and eligibility for free or reduced-price lunch (an indicator for having low-income parents). For further information on

[5] A few studies have linked experimental data to administrative data from tax records and other sources to measure impacts on outcomes such as college attendance rates and earnings (see e.g., Chetty et al. 2011 STAR, Dynarski et al STAR, Fredrikkson et al. QJE on class size effects). However, such linkages remain challenging and relatively rare. Our objective here is to show how one can make progress in identifying treatment effects of interest even when direct measurement of primary outcomes in the experimental sample is infeasible.

the STAR experiment, see Word et al. [1990] , Krueger [1999], and Chetty et al. [2011].

Observational Sample: New York City. We obtain observational information from the administrative records of the New York City public school district for 1.76 million children in grades 3-8 between the 1991-2009 school years. Starting from the raw data, we impose the same sample restrictions as in Chetty et al. [2014] – such as excluding special education classrooms and classrooms with less than 10 or more than 50 students – and additionally limit the sample to students for whom we observe test scores throughout grades 3-8. For comparability to the STAR treatment of dichotomous assignment to small vs. large classes, we define a "small" class in NYC as one with 26 (the sample median) or fewer students in third grade.

We observe math and reading test scores at the end of grades 3-8, which we standardize to have mean 0 and standard deviation within each grade in the NYC sample.[6] Critically, unlike in the STAR sample, we also observe an indicator for graduating from a NYC public high school by 2016, which we view as the primary outcome of interest.[7] We also observe information on students' race and ethnicity, sex, and (after the 1999 school year) eligibility for free or reduced-price lunch. For further information on the New York City data, see Chetty et al. [2014] and Mariano et al. [2024].

Summary Statistics. Appendix Table 1 presents summary statistics for the two samples. Although they are from different time periods and geographic settings, the two samples overlap on key student characteristics. Both districts serve primarily low-income students, with 61% of students in the STAR sample and 81% of the students in the NYC sample eligible for free or reduced-price lunches. Approximately one-third of the students are Black in both datasets, while the New York City sample has a significantly larger share of Hispanic students than the STAR sample. On average, there are 7.0 fewer students in third grade classrooms defined as "small" in the New York City data and 6.7 fewer students in the classrooms of students assigned to small classes in the STAR sample. 51% of students in New York City public schools graduated from

[6]Chetty et al. (2014) show that the within-grade variation in achievement in the NYC school district is comparable to the within-grade variation in other urban school districts nationwide, and hence is likely comparable to that in the STAR population, which exhibits broadly similar socioeconomic characteristics.

[7]We can only observe whether students graduated from a high school in the New York City public school district. 27% of students in our sample leave the NYC school district before the end of high school; we include these students in our analysis and code them as not graduating from an NYC high school. The estimated effects of class size reduction on graduation rates are larger in schools where fewer students leave the district, suggesting that this missing data issue leads us to understate the overall impact of class size reduction on graduating from any high school.

high school, consistent with official statistics [New York State Education Department, 2025].

5.2 Results

We begin with OLS regressions of 3rd grade test scores on an indicator for being assigned to a small class in the STAR and NYC datasets.[8] We include school fixed effects in all regressions run in the STAR sample because randomization was conducted within schools among children who entered in a given birth cohort. We analogously include school and birth cohort fixed effects in the NYC sample to isolate within-school and cohort variation in class size.

Table 1 (in the introduction) reports estimates from these regressions. In the STAR sample (Column 1), small class assignment in third grade increases end-of-third-grade test scores by 0.19 SD (se= 0.04).[9] In the NYC sample, the corresponding OLS estimate is -0.12 SD (se = 0.01). The difference between these estimates implies that the observational estimates are confounded under our maintained external validity assumption (Assumption 1).

Next, we implement the ESC estimator by estimating equation (2.4). We first calculate the difference between actual 3rd grade test scores and predicted test scores based on the student's class size ($\alpha_i^{\mathrm{S}} = Y_i^{\mathrm{S}} - \tau^{\mathrm{S}} W_i$) in the NYC sample. We then replicate the OLS specification in Column 2, additionally controlling for the residuals α_i^{S}. The ESC correction yields an estimated treatment effect of class size on third grade test scores in the NYC sample of 0.19 SD (se = 0.04), which coincides with the experimental estimate in the STAR sample by construction (Column 3 of Table 1).[10]

[8] Throughout our analysis of the STAR data, we use initial assignment to small class (rather than actual realized class size) as the independent variable, thereby reporting intent-to-treat estimates. Compliance with treatment assignment was imperfect because principals had to re-balance classes on other dimensions, such as gender composition. The ITT estimates provide the appropriate scaling for comparison to the observational NYC sample because the difference in average class size between those initially assigned to small vs. large classes in STAR of 6.7 students is comparable to that in the New York City data.

[9] Students who entered STAR schools before 3rd grade and were assigned to small classes in 3rd grade were assigned to small classes in earlier grades as well. We find that assignment to a small class has similar effects of end-of-3rd-grade test scores for those who entered STAR schools in 3rd grade (and thus were treated for only one year) as for those who entered in earlier grades. This is a consequence of the rapid fade-out of treatment effects on subsequent test scores documented in Figure 2. We therefore interpret the treatment effect on test scores as the causal effect of being assigned to a small class in third grade when we construct an analogous observational estimate in the NYC sample.

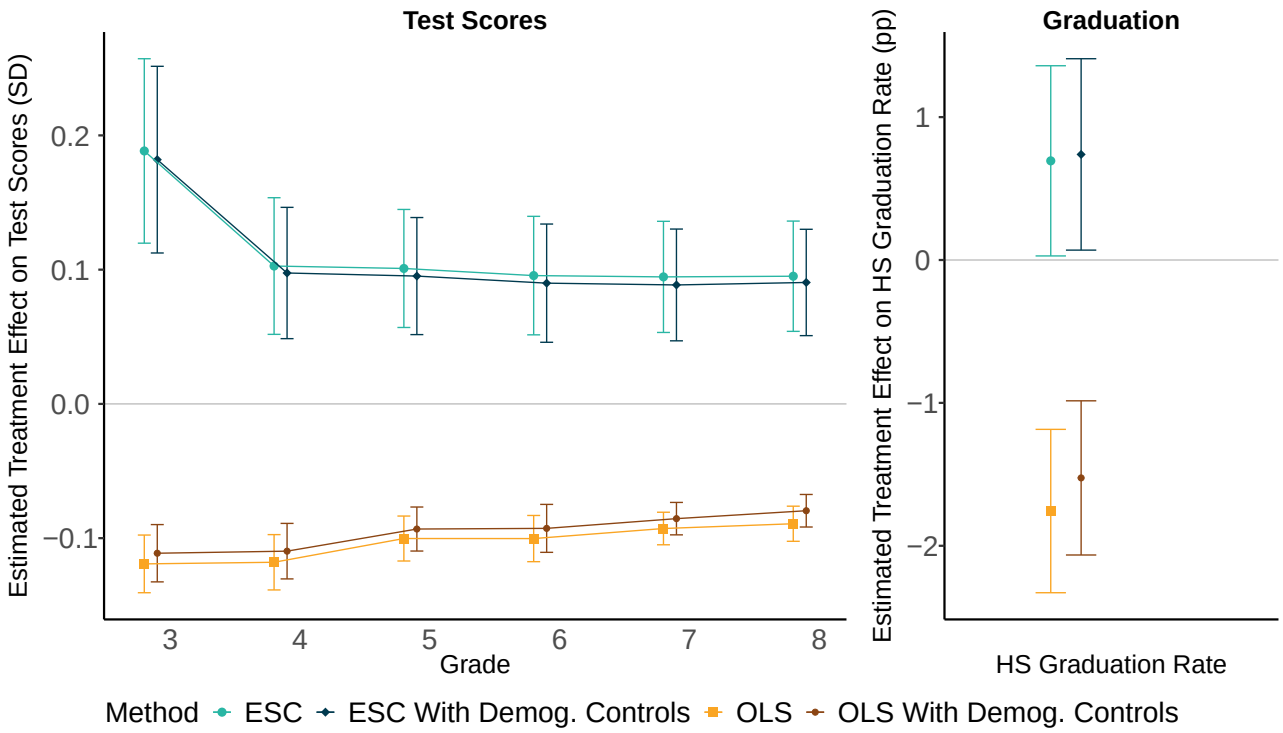
[10] We estimate standard errors for this and all other ESC estimates reported below using a bootstrap procedure, where we resample the student-level observations with replacement B = 1,000 times, re-estimate both the first-stage class-size effect $\hat{\tau}^{\mathrm{S}}$ (to form residuals α_i^{S}) and the second-stage ESC OLS controlling for α_i^{S} in each replication, and compute standard errors as the empirical standard deviation of the resulting bootstrap estimates.

We next use the ESC estimator to estimate treatment effects on subsequent outcomes. Figure 2 plots treatment effects on test scores in grades 3-8. The STAR experimental estimates (shown in black squares) are positive in all grades, while the OLS estimates in the NYC data (orange triangles) are all negative. The ESC estimates in grades 4-8 are all positive and very similar in magnitude and temporal pattern to the STAR estimates. Notably, the ESC estimates capture the well-known "fadeout" pattern in the STAR estimates – where the effects of interventions in early grades on test scores diminish in later grades. The close correspondence between the ESC estimates and the experimental estimates for the holdout outcomes (Y_i^H) of test scores in grades 4-8 supports the latent unconfoundedness assumption and demonstrates the ability of our approach to adjust for selection.

Finally, in the right panel of Figure 2, we turn to the primary outcome of interest – high school graduation – which we observe in the observational but not experimental sample. The ESC estimator implies that assignment to a small third grade class increases the probability of graduating from a NYC high school by 0.69 percentage points (Column 3 of Table 1). Small classes have 7 fewer students on average relative to a sample mean of 28 students; hence, a 25% reduction in class size in third grade increases high school graduation rates by 0.69 pp.

In contrast, the OLS estimator yields a negative association between small class assignment and high school graduation rates in the observational sample. Importantly, adjusting for selection on observables by controlling flexibly for key demographic covariates in the observational sample – the interaction of indicators for gender, race and ethnicity, eligibility for free or reduced-price lunch, and cohort – has little impact on the OLS estimates on test scores and graduation rates (Figure 3, Appendix Table 2). This result demonstrates that the experimental selection correction can adjust for selection on dimensions that are typically unobserved in standard administrative datasets.

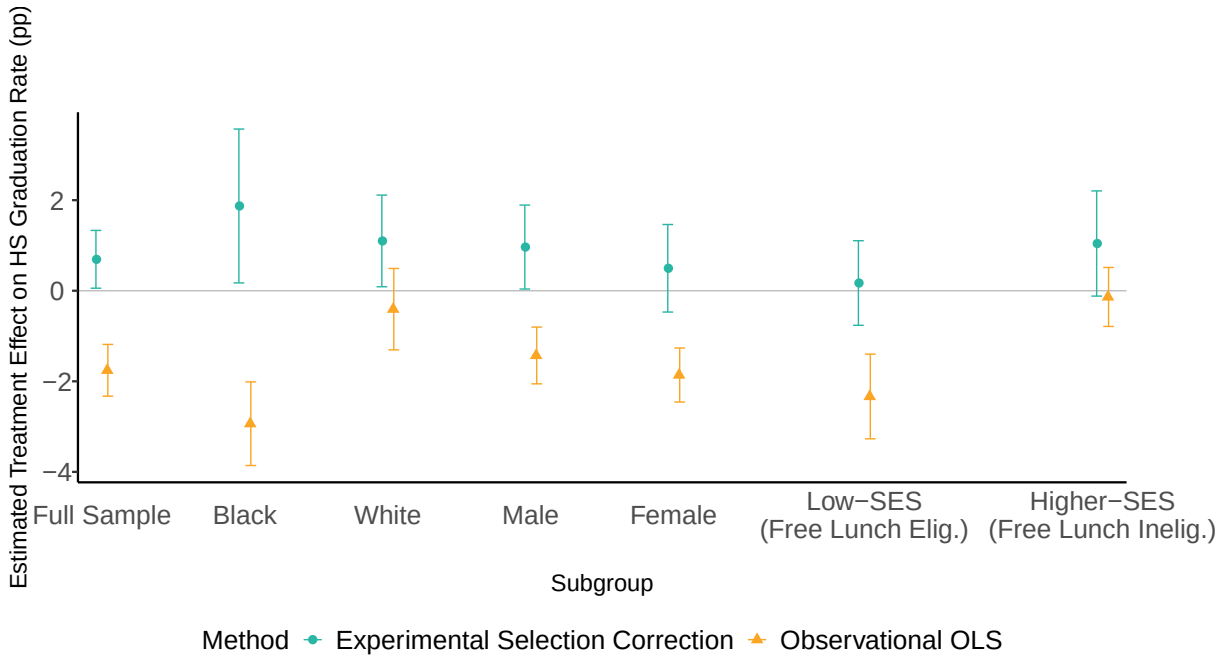
Figure 3: Effect of Controlling for Observables on Treatment Effect Estimates



Notes: This figure shows how controlling for demographic observables influences the estimated effect of assignment to a small 3rd-grade class on test scores (left panel) and high school graduation rates (right panel). "ESC" refers to the Experimental Selection Correction estimate, and "OLS" denotes the Observational OLS estimate in the NYC sample, constructed as described in the notes to Figure 2. All specifications include school and cohort fixed effects. Estimates labeled "With Demog. Controls" additionally control for the interaction of indicators for gender, race and ethnicity, eligibility for free or reduced-price lunch, and cohort, as well as an indicator for missing the free lunch variable (in the NYC data). Vertical bars represent 95% confidence intervals.

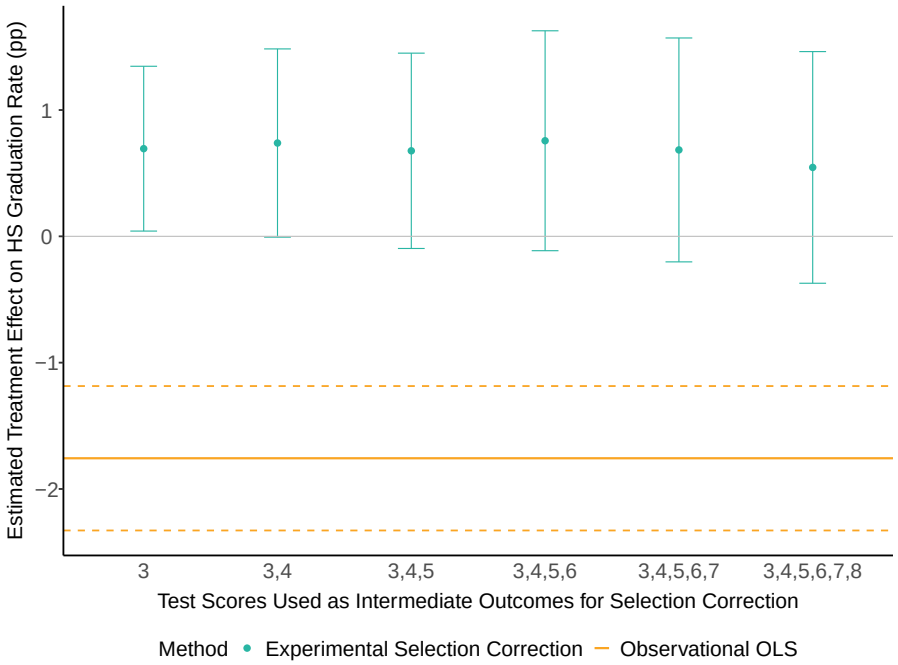
Robustness. We find very similar estimated impacts on high school graduation rates when focusing on specific demographic groups (*e.g.*, by race or sex), as shown in Figure 4. These findings allay the concern that differences in the demographic distribution between the STAR and NYC samples may lead to violations of the external validity assumption. We also find that the estimated impacts on high school graduation remain similar when we correct for selection using all test scores from grades 3-8 instead of just 3rd grade scores (Figure 5). These findings are consistent with the finding that treatment effects on test scores in grades 4-8 are very similar in the NYC and STAR data once we adjust for selection using 3rd grade test scores (Figure 2).

Figure 4: Heterogeneity of Treatment Effect Estimates on HS Graduation Across Subgroups



Notes: This figure presents subgroup-specific estimates of the effect of being assigned to a small 3rd-grade class on high school (HS) graduation rates. Each point corresponds to a point estimate obtained using either the observational OLS or Experimental Selection Correction (ESC) estimator in the NYC data, constructed as described in the notes to Figure 2. All specifications include school and cohort fixed effects. Estimates are expressed as percentage-point changes in HS graduation rates. Vertical lines represent 95% confidence intervals.

Figure 5: Robustness of ESC Estimates to Choice of Intermediate Outcomes

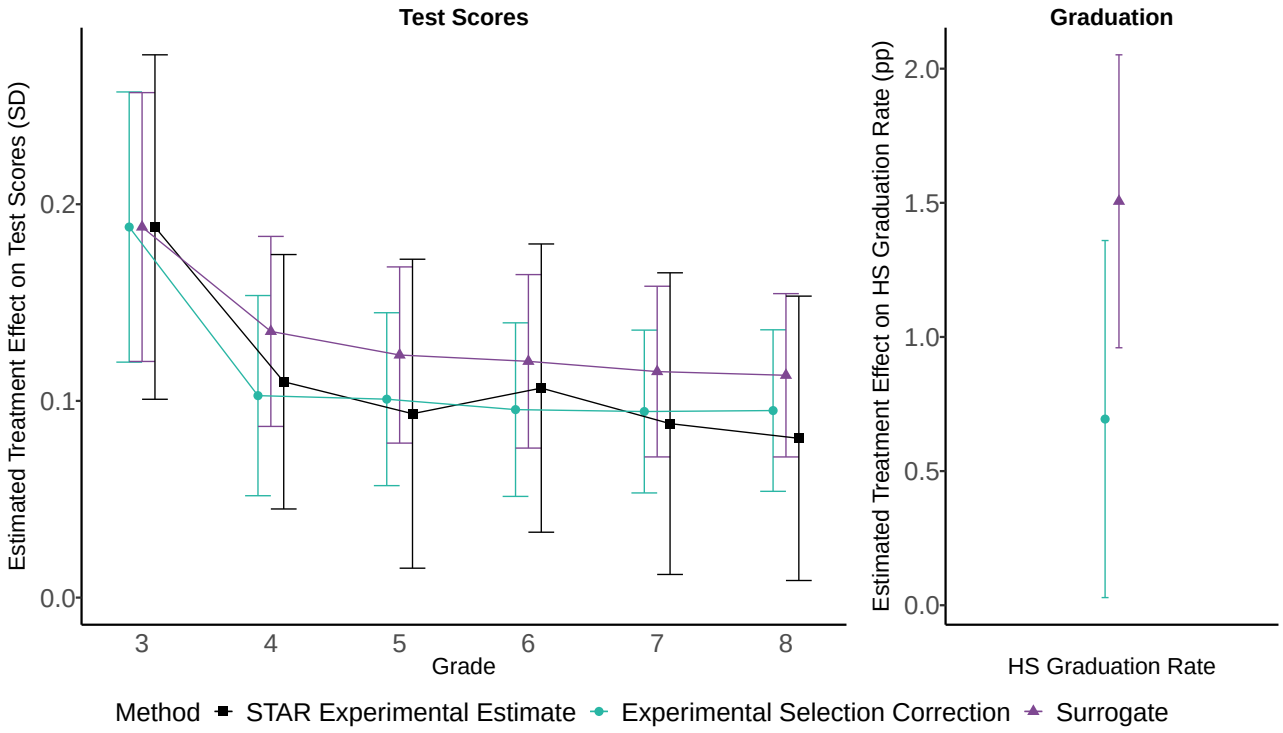


Notes: This figure illustrates how the estimated effect of assignment to a small 3rd-grade class on high school graduation rates varies with the intermediate outcomes used in the experimental selection correction procedure. Green points with vertical bars plot ESC estimates accompanied by 95% confidence intervals. The first estimate replicates the ESC estimate reported in Column 3 of Table 1, using only 3rd grade test scores for selection correction. The remaining estimates use additional test scores from grades 4-8 as intermediate outcomes in the selection correction procedure. The orange line shows the observational OLS estimate in the NYC sample from Column 2 of Table 1 (which does not use intermediate outcomes and hence is constant across the figure). The dashed orange line shows a 95% confidence interval for the OLS estimate.

Comparison to Surrogate Estimates. In Figure 6, we compare the ESC estimates to estimates from a surrogate index approach [Athey et al., 2019]. To construct the surrogate-based estimates, we multiply treatment effects on third grade test scores in the STAR sample by coefficients from OLS regressions of the outcome of interest on third grade scores in the NYC sample (including school and cohort fixed effects). The surrogate estimates on test scores in grades 4-8 are all higher than the experimental estimates (though not significantly so), while the ESC estimates match the experimental estimates more closely. This pattern is consistent with the education literature on fadeout, which finds that treatment effects of early interventions on test scores persist less than one would expect based on the serial correlation of test scores across grades, violating the assumption that test scores in earlier grades provide surrogates for later outcomes.

Accordingly, the ESC estimate yields an estimated treatment effect on high school graduation that is about half as large as the surrogate-based estimate.

Figure 6: Comparison of Surrogate and ESC Estimates



Notes: This figure compares the ESC estimates reported in Figure 2 to estimates that use third grade test scores as a surrogate for subsequent outcomes. The surrogate estimates are constructed by multiplying treatment effects on third grade test scores in the STAR sample by coefficients from OLS regressions of the outcome of interest (test scores in later grades or HS graduation) on third grade scores in the NYC sample. The ESC and experimental estimates reproduce the series plotted in Figure 2. See notes to Figure 2 for further details.

6 Conclusion

This paper has proposed a new method of combining experimental and observational data to improve causal inference about a primary outcome of interest. We leverage the internal validity of the experimental data to implement a selection correction based on secondary outcomes when estimating treatment effects on the primary outcome in the observational dataset. Our estimator relies on a key new assumption that we term latent unconfoundedness, which requires that the unobserved confounders that affect the primary and secondary outcomes are the same.

Our approach strictly weakens the assumptions underlying the popular approach of using intermediate outcomes as surrogates, yielding more credible estimates of causal effects when both the treatment and primary outcome are observed in the observational dataset.

We apply our Experimental Selection Correction estimator to estimate the effects of 3rd grade class size on test scores in later grades and high school graduation rates. In observational data, we find wrong-signed estimates that are likely biased by unobserved selection. The ESC estimator yields estimates that match holdout experimental estimates for test scores in later grades and provides one of the first estimates of the causal effects of class sizes on high school graduation rates in the U.S. – showing that a 25% reduction in class size in third grade increases high school graduation rates by 0.7 pp.

As observational data become more widely available, it would be valuable to build on the ideas proposed here by developing approaches to using experiments to correct for selection in observational data. Recent econometric advances have extended the framework we propose here in both identification and estimation—e.g., Meza and Singh [2024], Park and Sasaki [2024a,b], Imbens et al. [2025]—but there remains substantial scope for further development. In particular, future work could seek to characterize and weaken the latent unconfoundedness condition, potentially by leveraging multiple intermediate variables or partial identification strategies.[11] Empirically, it would be useful to characterize settings where latent unconfoundedness is a good approximation using validation studies in order to guide future applications.

[11]For example, the latent unconfoundedness assumption we use requires that all variation in students' test scores arises from unobservables that affect graduation rates as well, ruling out shocks that may affect test performance but not later outcomes such as illness or noise on the day of the test. In practice, test-retest reliability tends to be very high (exceeding 0.8), so such noise is likely minimal in our application. But in other settings, noise in the intermediate outcome may be more substantial; in such cases, it may be possible to use instrumental variables approaches to adjust for such noise.

Appendix

A. Proofs

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 Theorem 1:[12] 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

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

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

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

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

By Latent Unconfoundedness (Assumption 4) 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^{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. □

B. Alternative Approaches to Estimation

In this appendix, we present three approaches to estimation that are alternatives to the control function approach described in the main text.

Imputation. In the imputation approach, we impute the missing primary outcomes in the experimental sample and then difference the average imputed outcome by treatment status in the experimental sample, adjusted for pretreatment variables.
In the first step, we estimate the conditional mean of the primary outcome given the secondary outcome, treatment and pre-treatment variables in the observational sample:

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

In the second step we impute, for all units in the experimental sample, the primary outcome as $\hat{Y}_i^{\rm P}=\hat{\kappa}(W_i,X_i,Y_i^{\rm S})$. In the third step we use the standard program evaluation methods under unconfoundedness on the experimental sample with the imputed primary outcomes adjusted for differences between treated and control units in X_i. This last step can be based on matching, regression adjustment, (augmented) inverse propensity score weighting, and doubly robust methods, see for a general discussion Imbens and Wooldridge [2009].
If in the experimental sample the treatment is completely random, we could in this step estimate the average treatment effect in the experimental sample as the simple difference in average outcomes,

$$\hat{\tau}^{\rm imp,E}=\frac{1}{N_1^{\rm E}}\sum_{i:P_i={\rm E}} W_i\hat{\kappa}(1,X_i,Y_i^{\rm S})-\frac{1}{N_0^{\rm E}}\sum_{i:P_i={\rm E}}(1-W_i)\hat{\kappa}(0,X_i,Y_i^{\rm S}),$$

although this would not be efficient in the presence of covariates, the same way the difference in means estimator is not efficient in a randomized experiment with covariates.

Weighting. Another alternative is to estimate the average effect by differencing weighted averages of outcome in the treated and control subsamples of the observational sample. The difference of unweighted averages is not consistent for the average treatment effect because of the violation of unconfoundedness in the observational sample. The weighting is used to correct for that. First estimate the conditional distribution of $(Y_i^{\rm S},W_i)$ in the observational and experimental sample as

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

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

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

These weights adjust for the differences between the observational and experimental sample.
Assuming we have completely random assignment in the experimental sample, we can in the third step estimate the average treatment effect as

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

For efficiency, 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

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

This leads to

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

Influence Function. A third approach is to directly estimate an influence function and use that as the basis for an efficient estimator. There are some theoretical advantages to influence-function based methods in terms of robustness to misspecification of some of the nonparametric components. See Chernozhukov et al. [2022] for general discussion and Chen and Ritzwoller [2023] for a discussion in this setting. Chen and Ritzwoller [2023]. Here we re-write their estimator in the notation of the current paper.

To characterize the influence function estimator we need to define a number of additional functions:

$$\kappa(w,x,y^{\text{S}}) \equiv \mathbb{E}_P[Y_i^P | W_i = w, Y_i^S = y^s, X_i = x, G_i = O],$$

$$\overline{\kappa}(w,x) \equiv \mathbb{E}_P[\kappa(W_i, Y_i^s, X_i) | W_i = w, X_i = x, G_i = 0],$$

$$\rho(w,x,y^{\text{S}}) \equiv \text{pr}(W_i = w | Y^{\text{S}}(w) = y^{\text{S}}\, X_i = x, G_i = \text{E}),$$

$$\pi \equiv \text{pr}(G_i = \text{O}),$$

$$r(x) \equiv P(G = 1 | X = x),$$

$$e(x,g) \equiv \text{pr}(W_i = 1 | X = x, G_i = g),$$

$$\nu(x,y^{\text{S}}) = \mathbb{E}[Y_i | X_i = x, Y_i^{\text{S}} = y^{\text{S}}, G_i = \text{E}],$$

and

$$\eta(w,x) = \mathbb{E}[\nu(Y_i^{\text{S}}, X_i) | W_i = w, X_i = x, G_i = \text{E}].$$

Then the influence function is

$$\psi(y^p, y^s, w, x, g) = \frac{1_{g=\text{O}}}{\pi} \left(\frac{w(y^p - \kappa(1, y^s, x))}{\rho(1, y^s, x)} - \frac{(1-w)(y^p - \kappa(0, y^s, x))}{\rho(0, y^s, x)} + (\overline{\kappa}(1,x) - \overline{\kappa}(0,x) - \tau \right)$$

$$+ \frac{1_{g=\text{E}}}{\pi} \left(\frac{r(x)}{1-r(x)} \left(\frac{w(\nu(x,y^s) - \eta(1,x))}{e(x,\text{O})} - \frac{(1-w)(\nu(x,y^s) - \eta(0,x))}{1-e(x,\text{O})} \right) \right),$$

and the influence function based estimator is based on averaging an estimated version of this:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1_{g=\text{O}}}{\hat{\pi}} \left(\frac{w(y^p - \hat{\kappa}(1, y^s, x))}{\hat{\rho}(1, y^s, x)} - \frac{(1-w)(y^p - \hat{\kappa}(0, y^s, x))}{\hat{\rho}(0, y^s, x)} + (\hat{\overline{\kappa}}(1,x) - \hat{\overline{\kappa}}(0,x) \right) \right.$$

$$\left. + \frac{1_{g=\text{E}}}{\hat{\pi}} \left(\frac{\hat{r}(x)}{1-\hat{r}(x)} \left(\frac{w(\hat{\nu}(x,y^s) - \hat{\eta}(1,x))}{\hat{e}(x,\text{O})} - \frac{(1-w)(\hat{\nu}(x,y^s) - \hat{\eta}(0,x))}{1-\hat{e}(x,\text{O})} \right) \right) \right).$$

C. Stata Code for Implementing ESC Estimator

```
*Data Structure: stacked dataset with observations for experimental
sample (exp=1) and observational sample (obs=1)

*Variables: treatment (treatment indicator),score (secondary outcome),
 graduation (primary outcome)

*** Implementing Experimental Selection Correction Estimator

*Step 1: Estimate Treatment Effect on Secondary Outcome in
Experimental Sample

reg score treatment if exp==1

*Step 2: Estimate Selection Correction Term in Observational Sample

predict score_pred
gen selection = score - score_pred if obs==1

*Step 3: Estimate Treatment Effect on Primary Outcome in Observational
 Sample

reg graduation treatment selection if obs==1

*Note: conventional standard errors are invalid; bootstrap is needed

*** Surrogate Estimator (for comparison)

*Step 1: Predict Primary Outcome Based on Secondary Outcome in
Observational Sample

reg graduation score if obs==1
predict graduation_pred

*Step 2: Estimate Treatment Effect on Predicted Primary Outcome in
Experimental Sample

reg graduation_pred treatment if exp==1
```

See GitHub Repository for an R version of this code[13]

[13] https://github.com/OpportunityInsights/Experimental-Selection-Correction-Replication-Code.git.

Appendix Table 1: Summary Statistics for STAR and New York Data

	STAR		New York	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
A. Student Background Variables				
Female (%)	47.1	49.9	50.0	50.0
Eligible for Free or Reduced-Price Lunch (%)	60.6	48.9	80.8	39.4
Missing Free Lunch Indicator (%)	1.5	12.2	43.2	49.5
Graduated from NYC Public High School (%)	–	–	51.4	50.0
Race/Ethnicity (%)				
White	62.8	48.3	15.7	36.3
Black	36.4	48.1	33.9	47.3
Asian	0.28	5.3	11.9	32.4
Hispanic	0.18	4.3	38.1	48.6
Native American	0.1	3.5	0.3	5.8
Other	0.17	4.2	–	–
B. Classroom Characteristics				
Class Size in Grade 3	21.3	4.4	25.1	4.5
In Small Class at Grade 3 (%)	18.7	39.0	58.9	49.2
Size Given Large Class	23.9	2.4	29.2	2.2
Size Given Small Class	15.7	1.7	22.2	3.2
Initial Assignment to Small Class Size (%)	26.1	43.9	–	–
Size Given Large Class Assignment	23.0	3.3	–	–
Size Given Small Class Assignment	16.3	2.9	–	–
Number of Observations	11,599		1,758,838	

Notes: This table presents summary statistics for the analysis samples we use from two datasets: Project STAR (Experimental) and New York City school district (Observational). Panel A presents student background variables, including the percentage of female students, eligibility for free or reduced-price lunch, high school graduation rates, and race and ethnicity. High school graduation is omitted for the STAR dataset (denoted by "–") because it is not observed. Panel B reports classroom characteristics, with all statistics calculated as student-weighted means. The variable "In Small Class at Grade 3" indicates whether a student was actually placed in a small class, while "Initial Assignment to Small Class Size" captures the student's original randomized assignment in the STAR experiment. The table also reports average class size conditional on actual classroom assignment (small or large) and conditional on initial assignment status. For the NYC dataset, the reported sample size corresponds to the union of the variables examined—that is, the number of students with at least one non-missing variable (e.g., Grade 3–8 test scores, graduation outcomes, or demographic covariates).

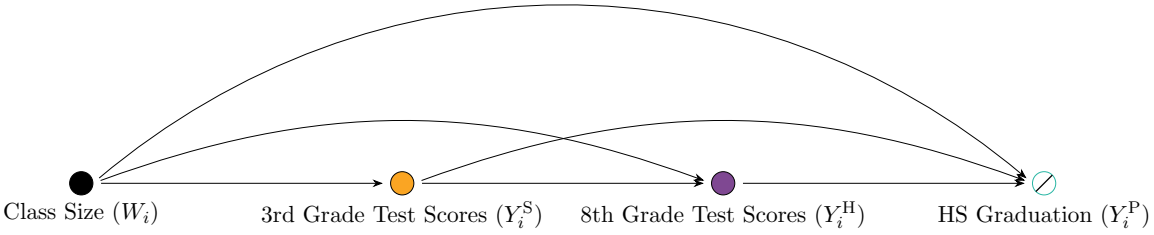
Appendix Table 2: Treatment Effect Estimates on HS Graduation: Comparison of Estimators

	OLS	OLS w/ Controls	ESC	ESC w/ Controls
Assigned to Small Class in 3rd Grade (W_i)	−1.76 (0.29)	−1.53 (0.28)	0.69 (0.34)	0.74 (0.34)
N	368,339	368,339	368,339	368,339

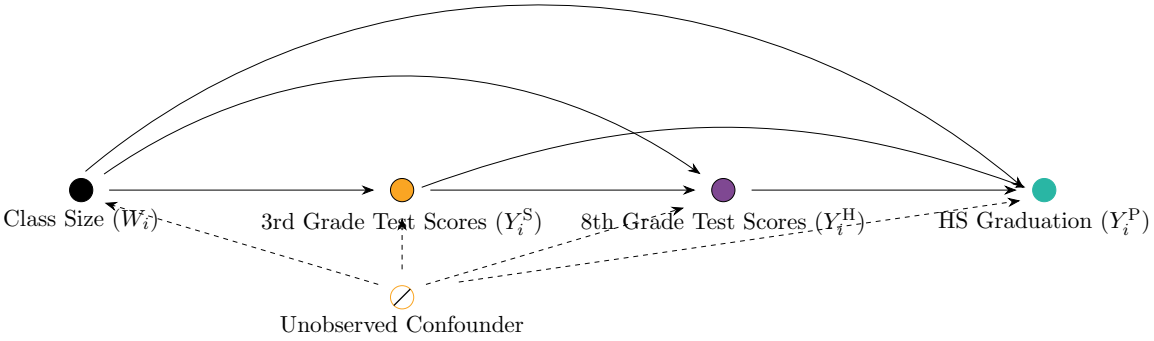
*Notes:*In this table, "OLS" refers to the ordinary least squares estimator based exclusively on the observational data, while "ESC" denotes the Experimental Selection Correction method implemented by combining experimental and observational data. Columns 1 and 3 control for school and cohort fixed effects. Columns 2 and 4 additionally control for the interaction of indicators for gender, race and ethnicity, eligibility for free or reduced-price lunch, and cohort, as well as an indicator for missing the free lunch variable (in the NYC data). Bootstrapped standard errors are reported in parentheses.

Appendix Figure 1: Experimental Selection Correction Model with Holdout Outcome

A. **Experimental Data**



B. **Observational Data**



Notes: This figure extends the framework presented in Figure 1 by introducing a holdout outcome, Y_i^H, representing 8th-grade test scores. Panel A illustrates the experimental data, where both the holdout outcome Y_i^H and the short-term outcome Y_i^S are observed under randomized assignment to class size W_i. Panels B illustrates the observational data, where Y_i^H is also observed. Dashed arrows represent potential unobserved confounding in the observational setting.

References

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.

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

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.

Barbara Biasi, Julien Lafortune, and David Schönholzer. What works and for whom? effectiveness and efficiency of school capital investments across the u.s. The Quarterly Journal of Economics, page qjaf013, 2025.

Zachary Bleemer. Affirmative action, mismatch, and economic mobility after california's proposition 209. The Quarterly Journal of Economics, 137(1):115–160, 2022.

David Card. The impact of the mariel boatlift on the miami labor market. Industrial and Labor Relations Review, 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.

Elizabeth U Cascio and Douglas O Staiger. Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research, 2012.

Jiafeng Chen and David M Ritzwoller. Semiparametric estimation of long-term treatment effects. Journal of Econometrics, 237(2):105545, 2023.

Xiaohong Chen and Andres Santos. Overidentification in regular models. Econometrica, 86(5):1771–1817, 2018.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. Econometrica, 90(3):967–1027, 2022.

Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. The Quarterly Journal of Economics, 126(4):1593–1660, 2011.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. American Economic Review, 104(9):2593–2632, 2014.

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.

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

David Deming. Early childhood intervention and life-cycle skill development: Evidence from head start. American Economic Journal: Applied Economics, 1(3):111–134, 2009.

Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. Top challenges from the first practical online controlled experiments summit. ACM SIGKDD Explorations Newsletter, 21(1):20–35, 2019.

James Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. Journal of Econometrics, 30(1-2):239–267, 1985.

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

James J Heckman, Jora Stixrud, and Sergio Urzua. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics, 24(3):411–482, 2006.

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, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. Journal of the Royal Statistical Society Series B: Statistical Methodology, 87(2):362–388, 2025.

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.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. Journal of Economic Literature, 47(1):5–86, 2009.

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. Advances in Neural Information Processing Systems, 31, 2018.

Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research, 2008.

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

Alan B Krueger. Experimental estimates of education production functions. The Quarterly Journal of Economics, 114(2):497–532, 1999.

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.

Louis T Mariano, Paco Martorell, and Tiffany Berglund. The effects of grade retention on high school outcomes: Evidence from new york city schools. Journal of Research on Educational Effectiveness, pages 1–31, 2024.

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.

Isaac Meza and Rahul Singh. Nested nonparametric instrumental variable regression: Long term, mediated, and time varying treatment effects, 2024. URL `https://arxiv.org/abs/2112.14249`.

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.

New York State Education Department. NYSED Data Site: Downloads. `https://data.nysed.gov/downloads.php`, 2025. Accessed: 2025-04-25.

Whitney K Newey. The asymptotic variance of semiparametric estimators. Econometrica, pages 1349–1382, 1994.

Yechan Park and Yuya Sasaki. A bracketing relationship for long-term policy evaluation with combined experimental and observational data. arXiv preprint arXiv:2401.12050, 2024a.

Yechan Park and Yuya Sasaki. The informativeness of combined experimental and observational data under dynamic selection. arXiv preprint arXiv:2403.16177, 2024b.

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

Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in Medicine, 8(4):431–440, 1989.

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. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1987. ISBN 047108705X, 9780471087052.

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. MIT Press, 2010. ISBN 9780262232586.

Elizabeth Word, John Johnston, Helen P Bain, B DeWayne Fulton, Jayne B Zaharias, Charles M Achilles, Martha N Lintz, John Folger, and Carolyn Breda. The state of tennessee's student/teacher achievement ratio (star) project. Tennessee Board of Education, 1990.

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