A Unified Framework for Transparent Quasi-Experimental Research*

Bryan A. Stuart and Evan J. Taylor

August 31, 2026

Abstract

This paper develops a framework that allows researchers to describe a wide class of quasi-experimental treatment effect estimates using direct analogs of the features that make a randomized controlled trial transparent. The outcome weight measures how much each observation's outcome contributes to an estimate, and its sign defines effective treatment and control groups. We show that estimates take a Wald form, equal to the weighted outcome difference between these groups divided by an effective first stage that quantifies the identifying treatment contrast. Our framework also allows researchers to assess covariate balance, examine how differences throughout the outcome distribution contribute to mean impacts, and identify influential observations. The treatment effect weight—the outcome weight times treatment—instead measures how much each observation's unobserved treatment effect contributes to an estimate, and exactly decomposes an estimate into subgroup-specific components. When the implied weighting of an estimate is undesirable, we show how to construct alternative estimates that satisfy researcher-specified criteria. We illustrate the framework with published analyses of democracy and growth, Chinese import competition, and returns to schooling.

JEL Classification Codes: C13, C81, C82

Keywords: Treatment effect estimation, transparency

***Stuart:** Research Department, Federal Reserve Bank of Philadelphia, and IZA, Email: bryan.stuart@phil.frb.org; **Taylor:** Department of Economics, University of Arizona, Email: evantaylor@arizona.edu.
For helpful comments and discussions, we thank Jeff Smith and numerous seminar and conference participants. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

1 Introduction

One key benefit of a randomized controlled trial (RCT) is that the resulting treatment effect estimates are very transparent. To be concrete, consider a setting where observations are randomly assigned to a treatment or control group, with treatment status for observation i indicated by the variable $D_i \in \{0, 1\}$. The resulting average intent-to-treat (ITT) effect on outcome Y is:

$$\hat{\theta}_{RCT} = \frac{1}{n_1} \sum_i D_i Y_i - \frac{1}{n_0} \sum_i (1 - D_i) Y_i, \tag{1}$$

where the number of observations assigned to the treatment and control groups is n_1 and n_0. The estimate $\hat{\theta}_{RCT}$ is transparent in five ways. First, it is straightforward to identify whether an observation is in the treatment or control group (*group membership*). Second, the exact contribution of each observation's outcome to the estimated treatment effect is clear (*observation contribution*). Third, researchers can easily examine whether there are differences in covariates—evaluated at the mean or quantiles of the distribution—between observations in the treatment and control groups (*covariate balance*). Fourth, researchers can plot the distribution of the outcome for the treatment and control groups to understand whether the mean effect $\hat{\theta}_{RCT}$ is driven by differences at particular quantiles (*distributional comparison*). Fifth, it is straightforward to decompose the overall effect estimate into subgroup-specific estimates by partitioning the sample (*subgroup decomposition*).

When motivating quasi-experimental research designs, researchers regularly describe the ideal experiment that their design is meant to approximate. However, the resulting quasi-experimental estimates are typically much less transparent than estimates from RCTs along each of the five dimensions named above. For example, consider the linear regression model: $Y_i = \alpha + D_i\theta + \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i$, where $D_i \in \{0, 1\}$ and $\mathbf{X}_i$ is a vector of control variables that are necessary to ensure $E[D_i\epsilon_i|\mathbf{X}_i] = 0$. While it is straightforward to compare individuals with $D_i = 1$ and $D_i = 0$, this simple comparison is not the variation that is used to estimate θ via ordinary least squares (OLS) (Frisch and Waugh, 1933; Lovell, 1963). As a result, this simple comparison might provide poor guidance about how the OLS estimate is produced. Relatedly, the contribution of each observation's outcome to the OLS estimate of θ is considerably more complicated than in the analysis of an RCT.

In this paper we develop a framework that enhances the transparency of quasi-experimental treatment effect estimators. Our results apply to settings where the treatment variable is binary or continuous for a wide class of econometric models—including those based on OLS, two-stage least squares (2SLS), matching, inverse propensity weighting (IPW), marginal treatment effects (MTE), and kernel regression—in which an estimate of interest can be written as a linear function of observed outcomes:

$$\hat{\theta} = \sum_i \omega_i Y_i, \tag{2}$$

where ω_i is the *outcome weight* (OW). For the ITT estimator in equation (1), the OW is $1/n_1$ for treatment group observations and $-1/n_0$ for control group observations.

We show how to calculate the OW for a wide class of econometric models and then describe how the OW delivers direct analogs of the first four transparency-generating features above. Group membership follows from defining the *effective treatment and control groups* as observations with $\omega_i > 0$ and $\omega_i < 0$; these are the groups that contribute positively and negatively to the estimated treatment effect, mirroring the contribution of treatment and control groups in an RCT. Observation contributions are given by the OW itself, as is immediately evident from equation (2). Covariate balance can be assessed by comparing OW-weighted means of covariates across the two groups, and distributional comparisons by plotting OW-weighted distributions for each group. The fifth feature, subgroup decomposition, requires the treatment effect weight introduced below.

Our framework also delivers two objects that are not visible in most quasi-experimental designs. First, we show that estimates in this class take a Wald form equal to the OW-weighted difference in outcomes between the effective treatment and control groups, divided by their OW-weighted difference in treatment. The denominator is the *effective first stage*, and it quantifies the size of the treatment contrast that produces an estimate—an object that is typically reported only for designs with a single instrument. Second, *outcome-influence* identifies observations whose outcomes contribute a disproportionate amount to an estimate, a concern that does not arise for the estimator in equation (1) because every observation within a group receives the same weight.

Our framework also describes how a treatment effect estimate relates to the heterogeneous treatment effects that researchers frequently care about. Letting τ_i denote the treatment effect for observation i, we define the *treatment effect weight* as $TW_i = \omega_i D_i$, the product of the OW and the amount of treatment received. For the estimator in equation (1), the TW is $1/n_1$ for treatment group observations and zero for control group observations, so that the estimate reflects a simple average of treatment effects among those assigned to treatment. More generally, we show that any estimate in our class can be written as:

$$\hat{\theta} = \Delta + \sum_i TW_i \tau_i, \tag{3}$$

where Δ is the OW-weighted difference in untreated outcomes between the effective treatment and control groups. The first term is fundamentally unobservable and reflects both bias and finite-sample noise; standard identifying assumptions imply that it is zero in expectation or vanishes as the sample grows. The second term shows that the TW governs how individual treatment effects are aggregated into an estimate, just as the OW governs how outcomes are aggregated. TWs generally sum to one, making this term a weighted average of treatment effects, but the weights can be negative. Finally, summing TWs within subgroups delivers the fifth feature above, *subgroup decomposition*: it exactly decomposes an estimate into subgroup-specific contributions. These contributions are informative because they generally differ from a subgroup's share of the sample or of the treatment group, and because a subgroup's contribution is increasing in its own effective first stage.

The OW is determined directly by the data and estimator. In some cases, the OW of an estimate will not have desirable features. For example, a group's contribution to a treatment effect estimate—which is determined by the sum of TWs for observations in that group—might be quite different from that group's share of the entire sample or treated observations. To address this, we show how to construct alternative OWs that target researcher-specified contributions for each subgroup. As a different example, some observations with high treatment levels can have negative OWs; these observations are treated, but their outcomes contribute negatively to an estimate, and a researcher might not want to use this variation because it does not align with an idealized experiment. We show how to construct alternative OWs that ensure convex weights.

We apply our framework to three important and careful studies. The first is the analysis of how democracy affects GDP growth by Acemoglu et al. (2019). Their baseline approach is a linear dynamic panel model in which log GDP per capita of a country depends on an indicator for democratic status, several lags of log GDP per capita, and fixed effects for country and year. Our framework yields several results that enhance the transparency of the resulting estimates. First, 28 percent of actual treatment group observations have negative OWs—putting them in the effective control group—while the same share of actual control group observations have positive OWs, putting them in the effective treatment group. Second, a simple comparison of the actual treatment and control groups provides an incomplete and potentially misleading characterization of the effective treatment and control groups. For example, Western Europe and other developed countries account for 33 percent of the actual treatment group, 1 percent of the actual control group, and 12 percent of both effective groups. Third, the effective first stage is 0.71, which is sizable but notably lower than the first stage of 1 that is implied by a comparison of actual treatment and control groups. Fourth, the positive effect of democracy on average GDP growth is largely driven by fewer periods of GDP decline rather than more periods of high growth.

Fifth, a substantial portion of the treatment effect is driven by Albania, Sierra Leone, Korea, Zimbabwe, and Georgia; if these countries had not outperformed the GDP growth predicted by covariates, the estimated effect of democracy would have fallen from 0.787 to 0.444. The presence of observations with large outcome-influence does not necessarily indicate a problem with a treatment effect estimate; instead, outcome-influence is a diagnostic tool that helps researchers identify which outcomes are most important to understand. Sixth, African countries account for 36 percent of total TWs, well above their 14 percent share of democracies. Finally, treatment effect estimates that eliminate nonconvex weights are larger in magnitude than the baseline estimate in Acemoglu et al. (2019).

The second study we revisit is Autor, Dorn and Hanson (2013), which estimates the effect of Chinese import competition on US manufacturing employment using a shift-share instrument. We show that five commuting-zone-year observations have a total outcome-influence that is equal to one-third of the estimated treatment effect. This means that, if these five observations instead saw average growth in manufacturing employment (based on their covariates), the main estimate would fall in magnitude from -0.596 to -0.405. One observation—San Jose for 2000–2007—has outcome-influence equal to 19 percent of the estimated effect.

Third is the estimated return to schooling using compulsory schooling laws as instruments from Stephens and Yang (2014). Motivated by research documenting regional variation in several domains, they show that adding fixed effects for individuals' region and year of birth leads to 2SLS estimates that are generally small and indistinguishable from zero, even though first stage power remains large. Our framework provides a clear and definitive explanation for why this happens. Without region-year fixed effects, effective treatment observations have a pupil-teacher ratio that is 3.9 percent lower than effective control observations. Moreover, the South has a total outcome-influence of 0.0784, which means that had the South experienced average wages (based on covariates), the 2SLS treatment effect estimate would fall from 0.105 to 0.027. When including region-birth-year fixed effects, the South no longer has outsized outcome-influence, and the effective treatment and control groups are well-balanced on school quality measures. Adding region-birth-year fixed effects also cuts the effective first stage in half, from 0.18 to 0.09 years of schooling. Given the small size of this effective first stage (which is distinct from a first-stage F-statistic), researchers may find the variation generated by these compulsory schooling laws to be relatively uninformative about the effects of education on earnings.

Previous work provides valuable insights on the relationship between a treatment effect estimate and *observed* moments in the data. For example, there has been notable progress for saturated OLS models (Angrist, 1998; Angrist and Pischke, 2009), OLS difference-in-differences estimators (Goodman-Bacon, 2021; Sun and Abraham, 2021), and shift-share instrumental variables (Borusyak, Hull and Jaravel, 2022; Goldsmith-Pinkham, Sorkin and Swift, 2020).[1] A distinct literature derives the weights that estimators place on observed outcomes—the object we call the OW—and uses them to construct diagnostics for treatment effect estimates (Chattopadhyay and Zubizarreta, 2023; Knaus, 2024).

Our contribution to this literature fundamentally arises from our partition of the sample by the sign of the OW into effective treatment and control groups. This partition yields several transparency-generating features described above. It allows researchers to describe the variation used for estimation by comparing two opposing groups, including for covariates that a model does not control for.[2] It also underlies the Wald representation and definition of the effective first stage. We further introduce outcome-influence, which provides an exact additive decomposition that sums across observations and subgroups, in contrast to existing measures that describe how an estimate would change if an observation were dropped (Broderick, Giordano and Meager, 2020; Chattopadhyay and Zubizarreta, 2023).

A related set of papers studies the relationship between a treatment effect estimate and *unobserved* heterogeneous treatment effects. This work covers OLS (e.g., Yitzhaki, 1996; Angrist and Krueger, 1999; Aronow and Samii, 2016; Słoczyński, 2022; Borusyak and Hull, 2024; Goldsmith-Pinkham, Hull and Kolesár, 2024), two-way fixed effects and event-study estimators (e.g., de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2023; Borusyak, Jaravel and Spiess, 2024), and 2SLS estimators (e.g., Imbens and Angrist, 1994; Angrist, Graddy and Imbens, 2000; Mogstad, Torgovitsky and Walters, 2021; Blandhol et al., 2022). Chen (2026) derives the weights that regressions place on potential outcomes and characterizes when a regression estimand can be interpreted as a weighted average of treatment effects.

Our contribution to this literature arises from the TW, which governs how individual treatment effects aggregate into an estimate. The decomposition in equation (3) is a finite sample gener-

[1] More broadly, Angrist and Pischke (2009) state that "regression can be motivated as a particular sort of weighted matching estimator" (p. 69).

[2] Aronow and Samii (2016) use the weight that multiple regression places on unobserved treatment effects to describe the estimation sample. That weight is distinct from the OW and does not separate the observations that contribute positively and negatively to an estimate. See Appendix C for details.

alization of decompositions derived for specific settings.[3] We also characterize the conditions under which TWs are nonnegative, so that an estimate is a convex combination of treatment effects, show how to decompose an estimate into subgroup-specific components, and show how to construct alternative estimates that satisfy researcher-specified criteria when the implied weighting is undesirable. Throughout, we aim to make this framework immediately useful for empirical researchers by emphasizing intuition and simple derivations.

2 Mapping Treatment Effect Estimates to Observation-Specific Outcomes

This section describes a unified framework that maps treatment effect estimates to observation-specific outcomes. We first introduce the core concept of the *outcome weight* (OW) and establish fundamental results. After describing how to calculate the OW, we show how it can be used to describe effective treatment and control groups and identify influential observations.

2.1 Outcome Weight: Definition and Basic Results

We consider a situation where a researcher observes several features for each observation $i = 1, \ldots, n$. The scalar outcome of interest is Y_i. The scalar treatment variable of interest is D_i, which can be binary or continuous. The $q \times 1$ vector $\mathbf{W}_i$ contains all data that the researcher uses for estimation besides the outcome variable. This vector includes the treatment variable and a constant term and may also include control variables and instruments. Stacking these objects across observations leads to the $n \times 1$ vectors $\mathbf{Y}$ and $\mathbf{D}$ and the $n \times q$ matrix $\mathbf{W}$. The scalar parameter of interest is θ.[4]

To motivate our definition of the OW, we begin by considering a particularly simple and transparent estimator. In a simple RCT where D_i is a randomly-assigned binary treatment variable, the

[3]For example, Proposition 3 in Blandhol et al. (2022) and equation (15) in Goodman-Bacon (2021).

[4]Focusing on a scalar treatment variable simplifies the exposition. This focus is not restrictive because it is straightforward to change which variable in $\mathbf{W}$ is designated as the treatment variable of interest. For the same reason, our set-up accommodates categorical treatment variables.

intent-to-treat estimator defined in equation (1) can be written:

$$\hat{\theta}_{RCT} = \sum_i \omega_i Y_i = \sum_{i:D_i=1} \omega_i Y_i - \sum_{i:D_i=0} |\omega_i| Y_i, \tag{4}$$

where the weight ω_i is $1/n_1$ for observations in the treatment group ($D_i = 1$) and $-1/n_0$ for observations in the control group ($D_i = 0$), with n_1 and n_0 denoting the number of observations in the treatment and control groups.

The RCT ITT estimator in equation (4) illustrates the two transparency-generating features that follow from the weights themselves. First, it is easy to describe the treatment and control groups. Treatment group observations contribute positively to the treatment effect estimate—in the sense that a higher value of Y_i for a treatment group observation increases $\hat{\theta}_{RCT}$—and control group observations contribute negatively. These groups can be identified easily using the treatment variable D_i. They also can be identified using ω_i, which is a simple function of D_i in this case. Second, it is straightforward to quantify how much each observation's outcome contributes to the treatment effect estimate; this is ω_i, which is either $1/n_1$ or $-1/n_0$. Consequently, the estimator does not assign outsize influence to any observation.

We now show that these two features exist for a wide class of treatment effect estimators. Section 2.3 describes how to apply the same weights to analyze covariate balance and make distributional comparisons. Sections 2.4 and 2.5 develop two objects that are not visible in most quasi-experimental designs: the effective first stage and outcome-influence. Section 3 then discusses subgroup decomposition, which combines the weights with treatment status.

We focus on the class of linear estimators, for which the estimated parameter of interest can be written as a linear function of the observed outcome variables: $\hat{\theta} = f_n(\mathbf{W})'\mathbf{Y}$, where $f_n(\cdot)$ is a function that maps the data matrix $\mathbf{W}$ into an $n \times 1$ vector and defines the estimator. Linear estimators include those based on matching, inverse propensity weighting (IPW), ordinary least squares (OLS), two-stage least squares (2SLS), marginal treatment effects (MTE), and kernel regression.[5] The function $f_n(\cdot)$ assigns a *weight* to each observation that depends on the data used

[5]Generalized method of moments estimators where the moments are linear in outcome variables are also included. Estimators where $E[Y_i \mid \mathbf{W}_i]$ is a nonlinear function, including logit, probit, and Poisson models, do not follow this

for estimation and the identifying assumptions. This leads to our first key definition.

Definition 1. The *outcome weight* (OW) for observation i is the ith element of $f_n(\mathbf{W})$. Denoting the OW by ω_i, the estimated parameter of interest can be written:

$$\hat{\theta} = \sum_{i=1}^{n} \omega_i Y_i. \tag{5}$$

Equation (5) immediately shows how each observation's outcome influences the treatment effect estimate, as $\omega_i = \partial\hat{\theta}/\partial Y_i$. Observations with $\omega_i > 0$ contribute positively to the treatment effect estimate, while observations with $\omega_i < 0$ contribute negatively. We label these the *effective* treatment and control groups.

Definition 2. The *effective treatment group*, denoted $\mathcal{T}_E$, consists of observations with $\omega_i > 0$. The *effective control group*, denoted $\mathcal{C}_E$, consists of observations with $\omega_i < 0$.

Observations with $\omega_i = 0$ are in neither group, as their outcomes do not contribute to the estimated treatment effect.[6] These definitions of effective treatment and control groups naturally extend the definitions of *actual* treatment and control groups in a simple RCT to a wide range of estimators and forms of treatment variables. In the case of the RCT ITT estimator, the actual treatment and control groups coincide with the effective treatment and control groups. However, this intuitive pattern does not hold in general. Instead, the effective treatment group can contain observations with small values of the treatment variable, and the effective control group can contain observations with large values of the treatment variable. We discuss this further in Section 2.2 and show how to calculate the OW for different estimators.

Proposition 1 establishes some basic results about the OW that are of independent interest and are used in subsequent analyses.

Proposition 1 (Basic OW Properties)**.** Assume an estimator satisfies the mean-invariance property:

structure; we leave the analysis of these estimators for future work.

[6]This does not imply that the estimated treatment effect would be identical if observations with $\omega_i = 0$ were removed before estimation, as these observations' $\mathbf{W}_i$ vector could affect the OW of observations in the effective treatment or control groups.

the estimate $\hat{\theta}$ does not change if some constant $\kappa \neq 0$ is added to every observation's outcome. Then:

(A) (Zero-Sum) The sum of OWs equals zero:

$$\sum_i \omega_i = 0.$$

(B) (Equal-Sum) The effective treatment and control groups have the same total OW in absolute value:

$$\sum_{i \in \mathcal{T}_E} \omega_i = \sum_{i \in \mathcal{C}_E} |\omega_i| \equiv \Omega.$$

Proof. See Appendix A. □

Mean-invariance is a basic property that is satisfied for the vast majority of treatment effect estimators that are used in practice. For example, this assumption is satisfied by all OLS and 2SLS estimators that include a constant term, as well as by matching, normalized IPW, and kernel regression estimators.[7] We assume that this property holds for all of our subsequent analysis. We also assume throughout that the estimator is nondegenerate, so that $\Omega > 0$; this rules out only the uninteresting case in which every observation has $\omega_i = 0$.

Both parts of Proposition 1 extend features of the RCT ITT estimator to a wide class of treatment effect estimators. Most notably, Proposition 1(B) implies that the outcomes of the *effective* treatment and control groups get equal total weight in producing the treatment effect estimate, echoing the analogous result for the *actual* treatment and control groups.[8]

Equation (5) implies that some observations generally matter more than others in producing a treatment effect estimate. This means that the number of observations in the effective treatment and control groups might not adequately describe the effective sample size. By analogy to an RCT,

[7] Normalized IPW estimators (also sometimes referred to as Hajek IPW estimators) are normalized so that weights sum to 1.

[8] For the RCT ITT estimator, the sum of the ω_i terms for the actual treatment group is 1, and the sum for the actual control group is -1.

the *effective number* of observations in the effective treatment and control group is:

$$N^{\text{eff}}_{\mathcal{T}_E} = \left[\sum_{i\in\mathcal{T}_E}\left(\frac{\omega_i}{\Omega}\right)^2\right]^{-1}, \quad N^{\text{eff}}_{\mathcal{C}_E} = \left[\sum_{i\in\mathcal{C}_E}\left(\frac{|\omega_i|}{\Omega}\right)^2\right]^{-1}. \tag{6}$$

The effective sample size is inversely related to the dispersion of OWs.[9] For a simple RCT, the effective number of observations in the effective treatment group equals the actual number of observations in the actual treatment group, and similarly for the control groups.[10]

2.2 Calculating the Outcome Weight

Before describing how the OW can improve the transparency of quasi-experimental estimates, we summarize how to calculate the OW for OLS and 2SLS estimation. Appendix B provides details on how to construct the OW for weighted least squares, matching, normalized IPW, regression-adjustment, doubly-robust, kernel regression, and MTE estimators.

2.2.1 Ordinary Least Squares

Consider the equation:

$$Y_i = D_i\theta + \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i, \tag{7}$$

where D_i is a scalar variable of interest (which may be binary or continuous), θ is a scalar parameter of interest, $\mathbf{X}_i$ is a $(k-1)\times 1$ vector of control variables (including an intercept), and $\boldsymbol{\beta}$ is a $(k-1)\times 1$ vector of parameters. OLS estimation typically relies on the assumption $E[\epsilon_i|D_i,\mathbf{X}_i] = 0$. Our focus on a scalar parameter of interest simplifies the calculation of the OW and is without loss of generality, as a researcher can always relabel which variable plays the role of D_i. This set-up covers a large number of applications, including cross-sectional models that rely on conditional exogeneity for identification, difference-in-differences and event-study set-ups,

[9] To derive these expressions, note that the OW for each observation in the actual treatment group in the RCT ITT estimator is $\omega_i = 1/n_1$ and $\Omega = 1$. In this set-up, we have $\sum_{i\in\mathcal{T}_E}(\omega_i/\Omega)^2 = 1/n_1$, which implies that $n_1 = \left[\sum_{i\in\mathcal{T}_E}(\omega_i/\Omega)^2\right]^{-1}$. Similar steps apply for the control group.

[10] Chattopadhyay and Zubizarreta (2023) propose this measure for the actual treatment and control groups; we apply it to the effective groups.

and sharp regression discontinuity designs.

Proposition 2 (OLS OW)**.** The OW for observation i in an OLS estimate of $\hat{\theta}$ from equation (7) is:

$$\omega_i = \frac{\tilde{D}_i}{\sum_j \tilde{D}_j^2},$$

where $\tilde{D}_i = D_i - \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$ is the residual from regressing D_i on $\mathbf{X}_i$.

Proof. See Appendix A. □

The sign of the OW in Proposition 2 depends only on the sign of $\tilde{D}_i$. In turn, $\tilde{D}_i$ is positive for observations where D_i is larger than the predicted value from the linear projection of $\mathbf{D}$ on $\mathbf{X}$. This means that an observation with a large value of D_i could be in the effective control group ($\omega_i < 0$) if it has other characteristics $\mathbf{X}_i$ that are associated with an even larger value of D_i on average.

The derivation of the OLS OW relies on the widely-used Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963), and other researchers have noted its role in producing regression estimates (e.g., Chattopadhyay and Zubizarreta, 2023; Borusyak and Hull, 2024; Knaus, 2024). The OW differs from several existing measures used in regression diagnostics, as detailed in Appendix C. Leverage and Cook's distance are general diagnostics that summarize how unusual an observation is in the full covariate space (i.e., in terms of $(D_i, \mathbf{X}_i)$) or how much it shifts the vector of fitted values; both are nonnegative scalars that do not distinguish between observations that receive positive versus negative weight in forming the estimate. The change in a treatment effect estimate when leaving out an observation depends on the OW of an observation, as well as that observation's residual and leverage, and so conflates design-driven and outcome-driven components of influence. The multiple regression weight of Aronow and Samii (2016) is proportional to the squared residual $\tilde{D}_i^2$, and so does not distinguish between the effective treatment and control groups.

2.2.2 Two-Stage Least Squares

We next describe how to calculate the OW for models that are estimated using two-stage least squares (2SLS). This covers a large number of applications for cross-sectional and panel designs, as well as fuzzy regression discontinuities. We continue to use equation (7) as the starting point. However, in this set-up, the researcher is not willing to assume that $E[\epsilon_i|D_i, \mathbf{X}_i] = 0$. For concreteness, we treat D_i as endogenous and $\mathbf{X}_i$ as exogenous. There is a $p \times 1$ vector of instrumental variables, $\mathbf{Z}_i$, that satisfies $E[\epsilon_i|\mathbf{Z}_i, \mathbf{X}_i] = 0$. There is at least one instrument for D_i, and there might be more: $p \geq 1$.

Proposition 3 (2SLS OW)**.** The OW for observation i in a 2SLS estimate of $\hat{\theta}$ from equation (7) is:

$$\omega_i = \frac{\hat{D}_i - \breve{D}_i}{\sum_j (\hat{D}_j - \breve{D}_j)^2},$$

where $\hat{D}_i = \mathbf{Q}_i'(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{D}$ is the predicted first-stage value, $\breve{D}_i = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$ is the predicted first-stage value based on the exogenous controls but not the instruments, $\mathbf{Q}_i = (\mathbf{X}_i', \mathbf{Z}_i')'$ is the $(k - 1 + p) \times 1$ stacked vector of exogenous controls and instruments, and $\mathbf{Q}$ is stacked across observations. The OW is well-defined as long as the instruments are relevant conditional on the controls, so that $\sum_j (\hat{D}_j - \breve{D}_j)^2 > 0$.

Proof. See Appendix A. □

The sign of the OW in Proposition 3 depends only on the numerator, which measures the predicted value of D_i based on the instrument(s) and exogenous controls relative to the predicted value based only on the exogenous controls.[11] The OW is positive for observations where the instruments increase the predicted value of D_i relative to the prediction based only on exogenous controls. For example, if the instruments predict a higher value of D_i, then a positive OW does not simply require a high value of the instruments; it requires a high value of the instruments *conditional* on the controls. The 2SLS OW does not depend on the actual value of the endogenous treatment variable.

[11] In our set-up, there is always at least one exogenous control (the intercept).

2.3 Using the Outcome Weight to Describe Effective Treatment and Control Groups

Essentially every paper that analyzes an RCT reports averages of variables in the actual treatment and control groups. These averages are useful for understanding the context (e.g., the characteristics of observations used to produce a treatment effect estimate) and assessing balance between the treatment and control groups.[12]

The OW yields a direct analogy for a broad set of quasi-experimental estimators. In particular, the OW-weighted means of some variable x for the effective treatment and control groups are:

$$\bar{x}_{\mathcal{T}_E} = \sum_{i \in \mathcal{T}_E} \frac{\omega_i}{\Omega} x_i, \quad \bar{x}_{\mathcal{C}_E} = \sum_{i \in \mathcal{C}_E} \frac{|\omega_i|}{\Omega} x_i. \tag{8}$$

As in an RCT, the OW-weighted means in equation (8) are useful for two distinct reasons. First, they describe the sample used for estimation, assigning each observation a weight proportional to the one it receives in the treatment effect estimate, as shown in equation (5). The OW-weighted means can differ substantially from means that split the sample based on values of D_i but do not use the OW. Second, these means can be used to assess covariate balance between the effective treatment and control groups.[13]

The OW can also be used for distributional comparisons that describe the effective treatment and control groups beyond the mean. For example, plotting OW-weighted distributions of the outcome Y for the effective treatment and control groups sheds light on whether extreme values drive the treatment effect estimate. The OW ensures that the resulting distributions assign the same weight to each observation as the treatment effect estimate. In contrast, simply comparing densities of Y for observations with high versus low values of a treatment variable can produce misleading results outside of an RCT with full compliance.

[12] As shown in equation (4), a simple average reproduces the weighting used in the RCT ITT estimate. If the treatment probability differs across observations, it is straightforward to modify this equation to include randomization weights.

[13] By construction, control variables included in OLS or 2SLS regression are perfectly balanced. See Proposition A.1 in Appendix A for a formal statement and proof. However, the means in equation (8) can be used to assess balance on potentially important variables that are not included in a primary OLS or 2SLS specification.

2.4 Using the Outcome Weight to Quantify the Effective First Stage

The OW can also be used to show that treatment effects estimated using a wide range of estimators take on a generalized Wald (1940) form, where the difference in outcomes is scaled by the difference in treatment.

Proposition 4 (Wald Representation)**.** Assume an estimator satisfies the following weighted treatment sum condition:

$$\sum_i \omega_i D_i = 1.$$

This condition is satisfied by OLS, 2SLS, weighted least squares (WLS), matching, normalized IPW, regression-adjustment, and doubly-robust estimators. Then the estimated treatment effect can be written as the difference in the mean outcome among effective treatment and control groups divided by the difference in the mean treatment among effective treatment and control groups:

$$\hat{\theta} = \frac{\bar{Y}_{\mathcal{T}_E} - \bar{Y}_{\mathcal{C}_E}}{\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E}},$$

where OW-weighted means are constructed as in equation (8).

Proof. See Appendix A. □

Proposition 4 echoes the result from IV estimation with a binary instrument, Z_i, and no controls, where the estimated treatment effect is the ratio of the difference in mean outcomes and the difference in mean treatment status between observations with $Z_i = 1$ and $Z_i = 0$. Our result shows that a much broader class of estimators can be understood as making this type of comparison. For example, the Wald representation holds with controls, with continuous instruments, and for OLS estimation.

An important takeaway is that every estimator satisfying the weighted treatment sum condition has an *effective first stage*, equal to $\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E}$. This extends a commonly reported object for binary instrument designs and allows researchers to understand the size of the identifying treatment contrast across a wide range of settings.

2.5 Using the Outcome Weight to Identify Influential Outcomes

In the RCT ITT estimate, the outcome of each observation within the actual treatment and control groups gets the same weight. As a result, there is no reason to worry that the estimator assigns outsized influence to a small set of observations. In contrast, the same reassurance is not generally available for quasi-experimental estimators.[14] The OW can be used to identify observations with particularly influential outcomes.

We define the *outcome-influence* of observation i as:

$$\omega_i(Y_i - Y_i^b), \tag{9}$$

where Y_i^b is a researcher-specified baseline outcome. This definition follows from equation (5) and reflects the idea that an observation's contribution to a treatment effect estimate depends both on its OW, ω_i, and its realized outcome, Y_i. An observation has a large positive outcome-influence if it (i) has a large positive OW and has an outcome that is larger than the baseline value or (ii) has a large negative OW and has an outcome that is smaller than the baseline value.

The baseline value Y_i^b ensures that outcome-influence is well-defined no matter how the outcome is scaled and allows researchers to consider whether a realized outcome is larger or smaller than some natural reference point.[15] The simplest choice is to set $Y_i^b = \bar{Y}$. In practice, we suggest that researchers set Y_i^b equal to an estimate of $E[Y_i|\mathbf{X}_i]$, such as the predicted value from a regression of Y_i on control variables $\mathbf{X}_i$. This allows the baseline value Y_i^b to depend on observed characteristics of an observation besides its treatment value.[16]

The treatment effect estimate that would arise if all observations took on their baseline outcome is $\hat{\theta}^b \equiv \sum_i \omega_i Y_i^b$. As a result, the difference between the estimated treatment effect and the baseline

[14] Young (2022) shows that published 2SLS estimates are frequently driven by a small number of high-leverage observations, and Broderick, Giordano and Meager (2020) show that the sign of many published estimates can be reversed by dropping a small fraction of the sample.

[15] By Proposition 1(A), adding a constant to every Y_i and every Y_i^b leaves both $\hat{\theta}$ and $\omega_i(Y_i - Y_i^b)$ unchanged.

[16] An alternative approach is to set Y_i^b equal to an estimate of $E[Y_i|\mathbf{X}_i, D_i = 0]$. We view this approach as less desirable in most situations because it typically requires more assumptions to estimate and does not account for the fact that average treatment levels may vary with $\mathbf{X}_i$.

treatment effect is:

$$\hat{\theta} - \hat{\theta}^b = \sum_i \omega_i (Y_i - Y_i^b). \tag{10}$$

Observations with a large positive outcome-influence drive up the treatment effect estimate relative to the specified baseline. Observations with negative outcome-influence pull down the estimate. A natural benchmark is that the treatment effect estimate when all observations take on their baseline outcome is zero (i.e., $\hat{\theta}^b = 0$); in this case, the right-hand side of equation (10) provides an exact decomposition of the estimated treatment effect.[17] It is also straightforward to aggregate outcome-influence into different subgroups. The result is a precise understanding of how a treatment effect estimate depends on OWs and outcomes of specific observations or groups of observations. If some observations have especially high outcome-influence, a researcher may want to provide reassurance that the contribution of these observations is desirable (e.g., not likely caused by confounding variation).

Outcome-influence is related to, but distinct from, existing influence diagnostics. The leave-one-out change in a coefficient and the influence scores underlying the robustness metric of Broderick, Giordano and Meager (2020) both describe how an estimate would change if observations were removed.[18] Outcome-influence instead shows how outcomes are aggregated and provides an exact additive decomposition of $\hat{\theta} - \hat{\theta}^b$ that sums across observations and subgroups. Appendix C provides further comparisons.

3 Mapping Treatment Effect Estimates to Observation-Specific Treatment Effects

We next introduce the *treatment effect weight* (TW), which maps treatment effect estimates to unobservable, heterogeneous treatment effects. After providing some basic results, we show how the TW delivers subgroup decomposition, the fifth transparency-generating feature of an RCT.

[17] This would be true in an OLS or 2SLS model if Y_i^b is any linear function of exogenous controls.

[18] The leave-one-out change in a coefficient is equivalent, up to scaling, to the sample influence curve that Chattopadhyay and Zubizarreta (2023) derive for regression estimators.

3.1 Treatment Effect Weight: Definition and Basic Results

We use the potential outcomes framework and define $Y_i(d)$ as the potential outcome for observation i when the treatment status is d, with observation-specific treatment effect per unit of treatment τ_i. This implies that the observed outcome for observation i can be written:

$$Y_i(D_i) = Y_i(0) + D_i\tau_i.$$

The potential outcomes framework allows us to decompose a treatment effect estimate into these unobservable treatment effects. This decomposition relies on the treatment effect weight, which combines the OW with the amount of treatment.

Definition 3. The *treatment effect weight* (TW) for observation i is $TW_i \equiv \omega_i D_i$.

Proposition 5 (Treatment Effect Decomposition)**.** The estimated treatment effect can be written as Ω times the OW-weighted mean difference in untreated outcomes between the effective treatment and control groups, plus the sum of the TW-weighted treatment effects:

$$\hat{\theta} = \underbrace{\Omega \left(\sum_{i \in \mathcal{T}_E} \frac{\omega_i}{\Omega} Y_i(0) - \sum_{i \in \mathcal{C}_E} \frac{|\omega_i|}{\Omega} Y_i(0) \right)}_{\text{Mean difference in untreated outcomes}} + \underbrace{\sum_i \omega_i D_i \tau_i}_{\substack{\text{Weighted sum of} \\ \text{treatment effects}}} \equiv \Delta + \sum_i TW_i \tau_i$$

Proof. See Appendix A. □

Proposition 5 provides a familiar expression of a treatment effect estimate as reflecting a combination of bias and causal effects, and is a finite sample generalization of many other decompositions applied to specific settings. The first term, which we label Δ, represents the OW-weighted mean difference in untreated outcomes between the effective treatment and control groups, multiplied by Ω. This term is fundamentally unobservable, and it reflects both bias (e.g., omitted variable bias or selection into treatment) and finite sample noise (e.g., groups might be balanced in expectation, but not in realization). The second term reflects TW-weighted sums of observation-specific treatment effects.

Standard identifying assumptions imply that $\Delta \rightarrow_p 0$ as $n \rightarrow \infty$, and in estimators that are unbiased, $E[\Delta] = 0$. For example, in the OLS setting the assumption that $E[Y_i(0)|D_i, \mathbf{X}_i] = \mathbf{X}_i'\boldsymbol{\beta}$ implies $E[\Delta] = 0$.[19] A similar argument applies in 2SLS, but only for consistency. For difference-in-differences and event study designs, standard parallel trends and other assumptions imply that $E[\Delta] = 0$ and $\operatorname{plim} \Delta = 0$.

The second term gives a weighted sum of treatment effects. For estimators that satisfy the weighted treatment sum condition ($\sum_i TW_i = 1$)—which includes all of the estimators listed in Proposition 4—this can be interpreted as a weighted average. However, there is no guarantee that $TW_i \geq 0$. For a binary treatment variable, OLS TWs are nonnegative if and only if the predicted value from the linear projection of $\mathbf{D}$ on $\mathbf{X}$ is less than or equal to 1 for all treated observations.[20]

An important takeaway from Proposition 5 is that the TW is central to the aggregation of treatment effects into a treatment effect estimate. This complements equation (5), which shows that the OW is central to the aggregation of outcomes. Intuitively, the TW is key for aggregating treatment effects because it combines the OW with the amount of treatment.

3.2 Decomposing Estimates into Group-Specific Treatment Effects

Using TWs, we can show how a treatment effect estimate depends on group-specific treatment effects.

Corollary 1 (Subgroup Decomposition)**.** Let $g_i \in \{1, \ldots, G\}$ denote the mutually-exclusive, collectively-exhaustive group that contains observation i. Then the estimated treatment effect can be written:

$$\hat{\theta} = \Delta + \sum_{g=1}^{G} s_g \bar{\tau}_g,$$

[19]This relies on the result that control variables included in OLS regression are perfectly balanced (cf., footnote 13). The same assumption on $E[Y_i(0)|D_i, \mathbf{X}_i]$ implies that $\operatorname{plim} \Delta = 0$. The alternative assumption that $E[D_i|Y_i(0), \tau_i, \mathbf{X}_i] = \mathbf{X}_i'\boldsymbol{\lambda}$, as in the design-based approach discussed in Borusyak and Hull (2024), implies that $\operatorname{plim} \Delta = 0$.

[20]This is guaranteed in fully saturated models. It does not hold in many difference-in-differences estimators (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). Mogstad, Torgovitsky and Walters (2021) document an analogous failure for 2SLS with multiple instruments.

where the sum of TWs for group g is:

$$s_g \equiv \sum_{i:g_i=g} TW_i,$$

and the TW-weighted average treatment effect for group g is:

$$\bar{\tau}_g = \sum_{i:g_i=g} \frac{TW_i \tau_i}{\sum\limits_{i':g_{i'}=g} TW_{i'}}.$$

Proof. See Appendix A. □

It is straightforward to calculate the TW for each observation, which immediately leads to s_g. While the share terms add up to 1 for estimators satisfying the weighted treatment sum condition (i.e., $\sum_g s_g = 1$ if $\sum_i TW_i = 1$), there is no general guarantee that each s_g term is positive. As a result, a subgroup's average treatment effect, $\bar{\tau}_g$, might contribute negatively to the estimated treatment effect. Moreover, the share of group g in the treatment effect, s_g, generally differs from the share of group g in the overall sample and the actual treatment group. Because of this, an estimated treatment effect might not reflect an intuitive combination of group-specific treatment effects. In any case, Corollary 1 provides a useful way of understanding where a treatment effect estimate comes from.

In many cases, researchers want to decompose treatment effect heterogeneity in terms of a variable that is included as a control. For example, a researcher might include year fixed effects in a regression model and want to understand which years' treatment effects contribute the most. For OLS and 2SLS estimation in which the model includes a full set of fixed effects for the groups $g = 1, \ldots, G$, the sum of TWs contributed by a subgroup can be written:

$$s_g = \Omega_g(\bar{D}_{g,\mathcal{T}_E} - \bar{D}_{g,\mathcal{C}_E}), \tag{11}$$

where $\Omega_g = \sum_{i:g_i=g, i\in\mathcal{T}_E} \omega_i$ is the sum of effective treatment OWs for a subgroup, $\bar{D}_{g,\mathcal{T}_E}$ is the OW-weighted mean of the treatment variable for subgroup observations in the effective treatment group, and $\bar{D}_{g,\mathcal{C}_E}$ is defined similarly for the effective control group.[21] Equation (11) shows that the

[21]OLS and 2SLS estimation ensure that $\sum_{i:g_i=g} \omega_i = 0$ when the regression model includes fixed effects for each

sum of TWs contributed by a subgroup depends on (i) the overall importance of that subgroup as measured by the sum of OWs and (ii) the subgroup-specific effective first stage (cf., Proposition 4). An important implication is that a subgroup's contribution to the treatment effect decomposition is increasing in the effective first stage.

4 Modifying the Weights Used in Treatment Effect Estimation

So far, we have described how commonly-used estimators weight outcomes and treatment effects across observations to produce a treatment effect estimate. The weighting scheme imposed by these estimators might not be desirable in a particular application. Here we show how to modify the weights assigned to different observations and generate alternative treatment effect estimates. This analysis focuses on the OLS estimator, which is widely used and sometimes implements undesirable weights.

4.1 Constrained Optimization Representation of OLS

We first show that the OLS OW can be constructed by solving a constrained minimization problem. This result is a building block for our approach to constructing treatment effects using alternative weights. It also connects our framework to the literature on optimization-based balancing weights (e.g., Hainmueller, 2012; Zubizarreta, 2015), which chooses minimum-dispersion weights subject to covariate balance constraints. That literature constructs weights to target a chosen estimand, and typically restricts weights to be nonnegative; we instead begin from the weights that a researcher's existing specification already implies, which need not be positive.

Proposition 6 (Constrained Optimization Representation of OLS)**.** The OLS OW in Proposition 2 can be obtained by solving the following problem:

$$\min_{\{\omega_i\}} \sum_i \omega_i^2 \quad s.t. \sum_i \omega_i D_i = 1 \, , \; \sum_i \omega_i \mathbf{X}_i = \mathbf{0}.$$

Proof. See Appendix A. □

group. This result implies that $\sum_{i:g_i=g, i\in\mathcal{T}_E} \omega_i = \sum_{i:g_i=g, i\in\mathcal{C}_E} |\omega_i|$, which is used to derive equation (11).

Proposition 6 provides an alternative way to describe how OLS estimates are produced. In particular, OLS chooses OWs with the smallest variance subject to satisfying the weighted treatment sum condition and the covariate balance condition. This echoes the Gauss-Markov theorem, where the constraints in Proposition 6 are linked to OLS being unbiased and the minimization is linked to OLS having minimal variance.[22] The weighted treatment sum and covariate balance conditions are the restrictions on the weights needed for $\sum_i \omega_i Y_i$ to be unbiased for θ in equation (7), given the conditional exogeneity assumption that underlies OLS. There is a unique minimum to the problem in Proposition 6 if $(\mathbf{D}, \mathbf{X})$ has full rank. Proposition A.2 in Appendix A provides an analogous constrained optimization representation of the 2SLS OW.[23] Together, Propositions 6 and A.2 extend Proposition 3 of Chattopadhyay and Zubizarreta (2023), which provides an analogous result for OLS with a binary treatment variable.

4.2 Specifying Treatment Effect Weight Totals by Subgroup

As noted in the discussion of Corollary 1, the total weight placed on a subgroup's treatment effects in a treatment effect estimate might not be desirable. For example, if women and men are equally represented in the analysis sample, then a researcher might want to estimate a treatment effect that is an equally weighted average of the treatment effects on men and women. Corollary 2 shows how to produce a treatment effect estimate that addresses this issue.

Corollary 2 (Alternative Group Weights)**.** Let $(a_1, \ldots, a_G)$ be a vector of specified constants that describes a researcher's desired contribution of each group to the treatment effect estimate, satis-

[22] In particular, for a sample of independent observations with homoskedastic error terms ($\mathrm{Var}(\epsilon_i) = \sigma^2$), where explanatory variables are taken as fixed, we have:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\sum_i \omega_i Y_i\right) = \sum_i \mathrm{Var}(\omega_i Y_i) = \sum_i \omega_i^2 \mathrm{Var}(Y_i) = \sigma^2 \sum_i \omega_i^2 = n\sigma^2 \mathrm{Var}(\omega_i).$$

This illustrates the close correspondence between minimizing the variance of the treatment effect estimate, $\hat{\theta}$, and minimizing the variance of OWs.

[23] The 2SLS problem is similar to the OLS representation, with the added restriction that the weights must be linear functions of the exogenous controls and excluded instruments: $\omega_i = \boldsymbol{\gamma}_X' \mathbf{X}_i + \boldsymbol{\gamma}_Z' \mathbf{Z}_i$. The intuition behind this restriction is that it ensures that an observation's weight does not depend on its own value of the endogenous treatment variable, echoing the fact that the 2SLS OW in Proposition 3 is a function of $(\mathbf{X}_i, \mathbf{Z}_i)$. It is straightforward to include interactions between instruments and exogenous covariates.

fying $\sum_{g=1}^{G} a_g = 1$. OWs that implement this desired weighting can be obtained by solving the following problem:

$$\min_{\{\omega_i\}} \sum_i \omega_i^2 \quad s.t. \; \sum_i \omega_i D_i = 1 \; , \; \sum_i \omega_i \mathbf{X}_i = \mathbf{0} \; , \; \sum_{i:g_i=g} \omega_i D_i = a_g \; \forall g.$$

Denoting the resulting OW by ω_i^{alt}, the resulting treatment effect estimator is

$$\hat{\theta}^{\text{alt}} = \sum_i \omega_i^{\text{alt}} Y_i.$$

Proof. This follows immediately from adding a constraint to Proposition 6. The last constraint ensures that a subgroup's total contribution to the treatment effect, $s_g \equiv \sum_{i:g_i=g} \omega_i D_i$, equals the contribution desired by a researcher. □

Implementing the alternative estimator in Corollary 2 is straightforward. Appendix A proves that the solution to this minimization problem is identical to the estimate obtained by augmenting the OLS regression of interest by including interactions between the treatment variable, D_i, and indicators for the mutually exclusive groups of interest, and then estimating a linear combination of these treatment-group interaction terms evaluated at the desired shares $(a_1, \ldots, a_G)$.[24] This approach also makes clear that there is a unique solution to the minimization problem as long as the augmented regression has full rank.

Researchers should be aware that, if errors are homoskedastic and observations are independent, this estimator has a weakly higher variance than the one without specified group weights, and a strictly higher variance whenever $a_g \neq s_g$ for some group. This estimator may put higher weights on groups with little variation in treatment status (as we show later in an empirical application).

4.3 Eliminating Nonconvex Weights

A distinct issue is that the effective treatment and control groups might not correspond to the actual treatment and control groups. A researcher might want to avoid using actually-treated observations

[24]By the same logic, evaluating the treatment-group interaction terms at the shares $(s_1, \ldots, s_G)$ yields an identical estimate as running an OLS regression without interaction terms.

in the effective control group because these observations do not correspond to the ideal experiment that they have in mind. Corollary 3 shows how to address this issue.

Corollary 3 (Convex Weights for Actual Treatment Group)**.** Assume the treatment variable is binary: $D_i \in \{0, 1\}$. OWs that are constrained to be nonnegative for all observations in the actual treatment group can be obtained by solving the following problem:

$$\min_{\{\omega_i\}} \sum_i \omega_i^2 \quad s.t. \ \sum_i \omega_i D_i = 1 \ , \ \sum_i \omega_i \mathbf{X}_i = \mathbf{0} \ , \ \omega_i D_i \geq 0 \ \forall i.$$

Denoting the resulting OWs by ω_i^{convex}, the resulting treatment effect estimator is

$$\hat{\theta}^{\text{convex}} = \sum_i \omega_i^{\text{convex}} Y_i.$$

Proof. This follows immediately from adding a constraint to Proposition 6. The last constraint ensures that actually-treated observations have a nonnegative OW. □

Corollary 3 provides a straightforward algorithm for modifying the OLS estimator with a binary treatment variable to have convex weights, such that observations in the actual treatment group ($D_i = 1$) are in the effective treatment group ($\omega_i > 0$) or have no weight. The corollary extends to a continuous treatment variable.[25] It is not generally possible to provide an analytical expression for the OW in this setting, but it is straightforward to solve this minimization problem on a computer.[26] The additional constraint generally changes the OWs, and in some cases observations will go from having a nonzero OW without the convexity constraint to a zero OW with it.[27] The minimization problem in Corollary 3 has a unique solution as long as the feasible set that satisfies the constraints is nonempty.[28]

[25]This can be done by modifying the last constraint to be $\omega_i(D_i - c) \geq 0$ for a researcher-specified constant c that divides high and low treatment values.

[26]We solve the minimization problem by iteratively dropping observations with $\omega_i D_i < 0$ and re-estimating OWs until the number of observations in the estimation sample converges. Then we estimate OWs on the final sample.

[27]The solution to the minimization problem in Corollary 3 differs from an approach where OWs are set to be zero for observations with $D_i\omega_i < 0$. This is because simply replacing ω_i to be 0 for some observations leads the other constraints to be slack.

[28]The objective function is strictly convex and coercive. The set satisfying the constraints is convex and closed. Thus, there is at most one minimum, and the minimum exists if the feasible set is nonempty.

It is possible to modify or augment the constraints in Corollary 3. For example, adding the constraint $\omega_i(1 - D_i) \leq 0$ leads to a treatment effect estimate where all actual treatment observations have a nonnegative OW *and* all actual control observations have a nonpositive OW.[29]

5 Applications

We illustrate how our framework can bring greater transparency to treatment effect estimation by revisiting three important empirical papers. We provide a comprehensive analysis of the effects of democracy on GDP growth by Acemoglu et al. (2019) and more targeted analyses of the effects of Chinese import competition on manufacturing employment from Autor, Dorn and Hanson (2013) and the effects of education on earnings from Stephens and Yang (2014). Across these applications, we show how the OW and TW deliver substantive new insights into the estimated effects.

5.1 The Effect of Democracy on GDP Growth

Acemoglu et al. (2019) use a variety of estimators and empirical strategies and find consistent evidence of a positive effect of democracy on GDP growth. We focus on their baseline results, which come from a linear panel model:

$$y_{ct} = \beta D_{ct} + \sum_{j=1}^{p} \gamma_j y_{ct-j} + \alpha_c + \delta_t + \epsilon_{ct}, \tag{12}$$

where y_{ct} is the log of real GDP per capita in country c and year t (multiplied by 100 for the purpose of interpretation), D_{ct} is an indicator for whether a country is a democracy in each year, α_c is a country fixed effect, and δ_t is a year fixed effect. Their preferred specification controls for four lags of log GDP per capita ($p = 4$). The parameter of interest is β, and the baseline estimate is 0.787 (standard error: 0.226) (Table 2, column 3). We use data from their replication package, which allows us to reproduce their results exactly. The data set includes 175 countries from 1960–2010.

[29] With a continuous treatment variable there is no additional constraint to impose: the single constraint $\omega_i(D_i - c) \geq 0$ already forces observations with $D_i > c$ to have a nonnegative OW and observations with $D_i < c$ to have a nonpositive OW.

Describing Effective Treatment and Control Groups. Table 1 compares the actual treatment and control groups, democracies and nondemocracies, with the effective treatment and control groups. We report unweighted means of key variables for the actual groups alongside OW-weighted means for the effective groups (cf., Section 2.3). The actual and effective groups differ because of the controls in equation (12). For example, country fixed effects absorb time-invariant differences across countries, while lags of log GDP control for time-varying changes in economic conditions. The OW facilitates a precise understanding of how these controls shape the outcomes used for estimation.

We begin by examining baseline covariates in the spirit of a balance test. Panel A displays variables that are controlled for in the main specification, while Panel B displays variables that are not in the main specification but could be. One takeaway is that the actual treatment and control groups differ substantially. For example, Western Europe and other developed countries account for 33 percent of the actual treatment group, but only 1 percent of the actual control group.[30] The first lag of log GDP per capita is 138.3 points (299 percent) higher in the actual treatment group than the actual control group. A second takeaway is that the actual treatment and control groups do a poor job of describing the effective treatment and control groups. Western Europe and other developed countries account for 12 percent of the effective groups—far from the shares among the actual groups—and the average of the first lag of log GDP in the effective groups is 98 points lower than the actual treatment group and 41 points higher than the actual control.[31]

There is no difference in the Panel A variables between the effective treatment and control groups by construction (cf., footnote 13). The same is not true of the Panel B variables, which are not in the main specification. The controls in the model eliminate most of the very large differences in further-lagged GDP, as intended. However, the effective treatment group has GDP growth over the previous 5–8 years that is about 4 points (4 percent) faster than the effective control group.[32]

[30]Equation (12) includes country fixed effects, which absorb geographic region fixed effects. We focus on regions here because they provide a more concise summary.

[31]It is straightforward to calculate standard errors on the weighted averages and differences in this table.

[32]Acemoglu et al. (2019) control for these variables in a robustness test and find that the effect of democracy increases somewhat when controlling for these variables.

We focus here on variables that are in the Acemoglu et al. (2019) replication data. More broadly, examining a wide range of variables can help researchers easily assess a range of potential threats to identification.

In Panel C, we report similar means for the treatment and outcome variables. The first stage implied by a simple comparison of the actual treatment and control groups is 1, while the effective first stage, which accounts for controls, is 0.71. This is a substantial treatment contrast; as we show below, some empirical settings lack this. Because the model controls for lagged GDP, it is easier to interpret the outcome as growth: relative to its first lag, log GDP per capita grows by 1.68 points in the effective treatment group ($= 719.51 - 717.83$) and 1.12 points in the effective control group. This representation provides the empirically relevant comparison of means for the two groups, which is standard for an RCT but generally not shown for quasi-experimental work. The Wald representation of the treatment effect (Proposition 4) is $(1.68 - 1.12)/(0.87 - 0.16) \approx 0.787$, which is the estimated effect from equation (12).

Panel D reports summary statistics on OWs. Notably, 28 percent of observations in the actual treatment group have a negative OW—which puts them in the effective control group—while the same share of the actual control observations have a positive OW—putting them in the effective treatment group.[33] Proposition 2 explains why this happens. In particular, democracies where the predicted democracy variable—based on the included controls—exceeds 1 have a negative OW, putting them in the effective control group. Conversely, nondemocracies where the predicted democracy variable is below 0 have a positive OW, putting them in the effective treatment group. These out-of-bounds predictions occur because equation (12) includes continuous control variables and does not include a full set of interactions between controls.

Comparing Distributions in Actual and Effective Treatment and Control Groups. As noted in Section 2.3, it is possible to compare the unweighted distribution of key variables for the actual treatment and control groups to the OW-weighted distribution for the effective treatment and

[33] In general, these numbers need not be equal. Appendix Figure D.1 displays the density of OWs for the actual treatment and control groups.

Table 1: Means of Key Variables, Actual and Effective Treatment and Control Groups, Acemoglu et al. (2019)

	Actual Control (1)	Actual Treatment (2)	Effective Control (3)	Effective Treatment (4)
Panel A: Controls in Main Model				
Africa	0.44	0.14	0.32	0.32
East Asia and the Pacific	0.13	0.08	0.10	0.10
Eastern Europe and Central Asia	0.09	0.11	0.10	0.10
Western Europe and other developed countries	0.01	0.33	0.12	0.12
Latin America and the Caribbean	0.13	0.29	0.25	0.25
Middle East and North Africa	0.16	0.01	0.05	0.05
South Asia	0.04	0.05	0.05	0.05
Lag 1, Log GDP per capita ($\times$ 100)	677.28	815.58	717.83	717.83
Lag 2, Log GDP per capita ($\times$ 100)	675.73	813.50	716.46	716.46
Lag 3, Log GDP per capita ($\times$ 100)	674.00	811.41	714.97	714.97
Lag 4, Log GDP per capita ($\times$ 100)	672.28	809.40	713.48	713.48
Panel B: Variables Not in Main Model				
Lag 5, Log GDP per capita ($\times$ 100)	671.10	807.55	712.72	712.05
Lag 6, Log GDP per capita ($\times$ 100)	669.90	805.70	712.17	710.71
Lag 7, Log GDP per capita ($\times$ 100)	668.83	803.88	711.90	709.05
Lag 8, Log GDP per capita ($\times$ 100)	667.84	802.07	711.66	707.12
Panel C: Treatment and Outcome Variables				
Democracy [treatment]	0.00	1.00	0.16	0.87
Log GDP per capita ($\times$ 100) [outcome]	678.75	817.77	718.95	719.51
Panel D: Outcome Weights				
Share of OWs that are positive	0.28	0.72	0.00	1.00
Share of OWs that are negative	0.72	0.28	1.00	0.00
Sum of OWs	-1.00	1.00	-1.40	1.40
Smallest OW ($\times$ 1000)	-2.51	-0.51	-2.51	0.00
Largest OW ($\times$ 1000)	0.63	2.24	-0.00	2.24
Number of country-year observations	2991	3345	3095	3241

Notes: Columns 1 and 2 report unweighted means of the actual control and treatment groups used in the baseline analysis of Acemoglu et al. (2019). Columns 3 and 4 report OW-weighted means for the effective control and treatment groups in their baseline analysis. OW-weighted means for the effective control group use the absolute value of the OW as a weight. Panel A displays variables that are controlled for in the baseline specification, and Panel B includes variables that are not in the baseline specification but could be. Panel C shows the treatment and outcome variables. Panel D reports summary statistics on OWs. The OW-weighted means of control variables in Panel A are equal by construction. The averages in Panel B are calculated among observations with nonmissing variables.

Source: Authors' analysis of data from Acemoglu et al. (2019).

control groups. We focus this analysis on the outcome variable, though it can also be done for explanatory variables. Because it is easier to interpret, we focus on the one-year change in log

GDP per capita.[34]

Panel A of Figure 1 displays the empirical CDF of the one-year change in log GDP per capita for the actual treatment and control groups. This is an immediate extension of reporting the mean outcome for these groups, as is commonly done. The actual treatment group has higher GDP growth up to about the 60th percentile of the distribution, while the actual control group has higher GDP growth beyond that point. One interpretation of this pattern is that nondemocracies are more likely to generate very high and very low levels of GDP growth relative to democracies. However, this comparison does not use any controls, and it turns out to be misleading.

Panel B displays the empirical CDF for the effective treatment and control groups. This is the distribution of the outcome that is used to produce the mean estimate of β in equation (12). The distribution of the change in log GDP per capita for the effective treatment group is shifted rightward up to the 75th percentile of the distribution, beyond which the two distributions largely overlap. This figure shows that the positive average effect of democracy on GDP growth comes from a reduction in periods of real GDP *declines*. In contrast, there is little evidence that the positive effect of democracy comes from increases in periods with large positive GDP growth.[35] It is also possible to construct analogous distributional figures for covariates.[36]

Visualizing Weights for Selected Countries. A final way to describe the effective treatment and control groups in this setting is to visualize OWs for selected countries. Figure 2 does this for four illustrative cases. The Gambia transitions to a nondemocracy in 1994. As one would hope, the Gambia is in the effective treatment group in years when it is a democracy and the effective control group in years when it is a nondemocracy. South Korea transitions to a democracy in 1988, and there is a similar correspondence across years in actual and effective treatment status. The USA is

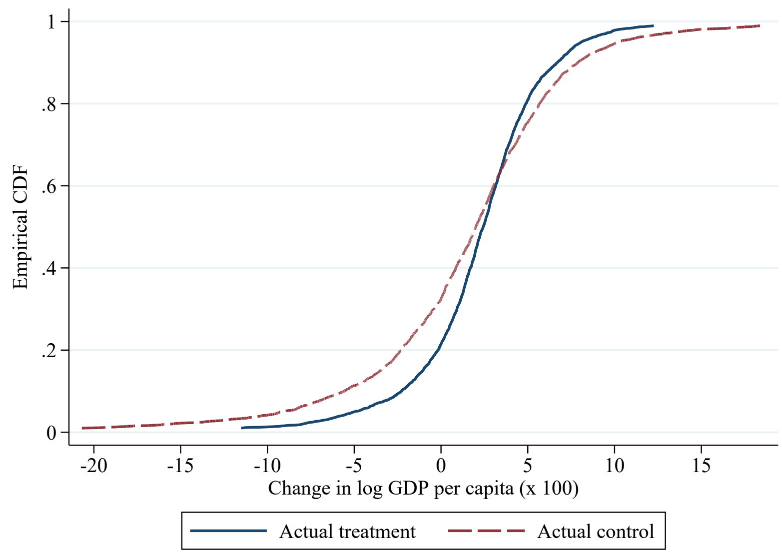
[34]Changing the dependent variable in equation (12) to the one-year change in log GDP per capita yields an identical estimate of β because lagged GDP is a control.

[35]These distributions are a helpful diagnostic tool for understanding mean effects and can motivate analyses of causal effects throughout the distribution, which require additional assumptions (e.g., Heckman, Smith and Clements, 1997; Firpo, Fortin and Lemieux, 2009).

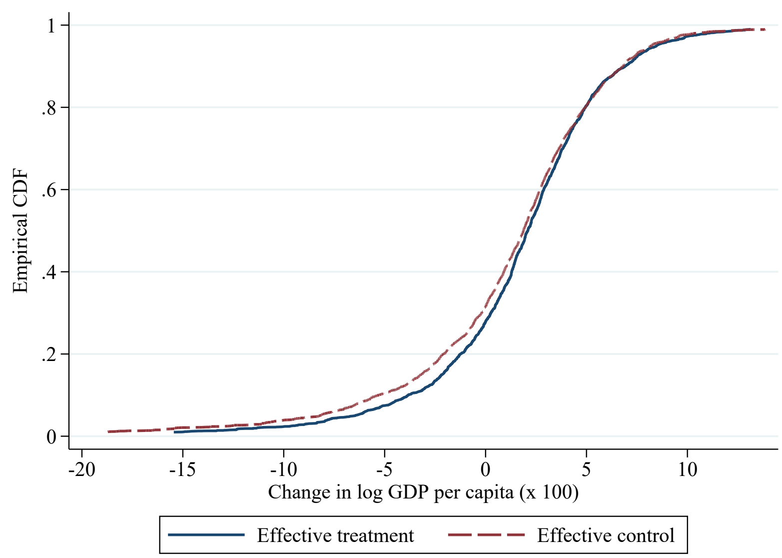
[36]For variables included in equation (12), means in the effective treatment and control groups are equal by construction. However, this equality need not apply to the entire distribution. As an example, Appendix Figure D.2 shows that the OW-weighted distribution of the one-year lag of log GDP per capita is quite similar for the effective treatment and control groups, though there are some differences at the highest levels of GDP per capita.

Figure 1: Distribution of Outcomes, Acemoglu et al. (2019)

(a) Empirical CDFs of Outcome, Actual Treatment and Control



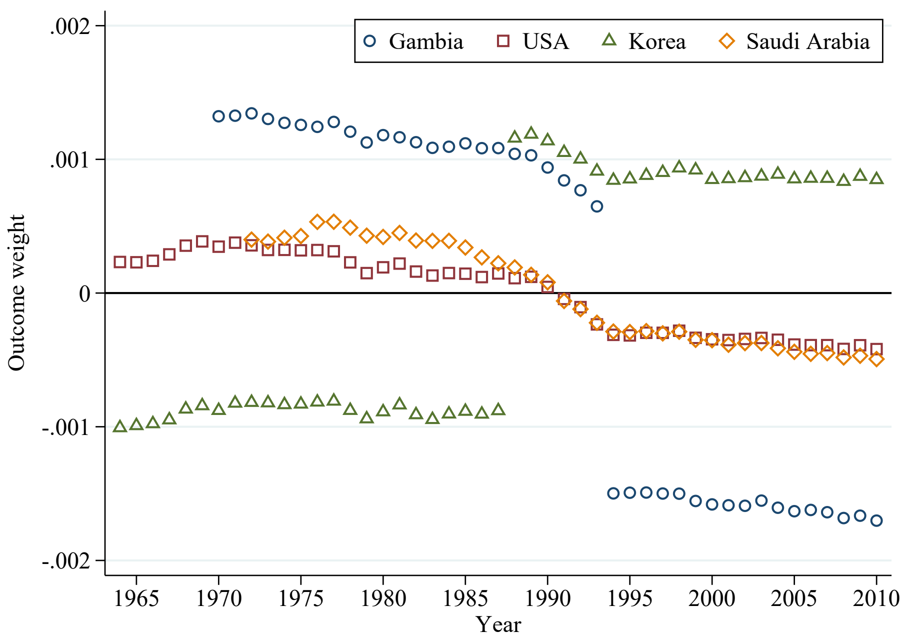
(b) OW-Weighted Empirical CDFs of Outcome, Effective Treatment and Control



Notes: Panel A displays the empirical CDFs of the one-year change in log GDP per capita (times 100) for the actual treatment and control groups. Panel B displays empirical CDFs for the effective treatment and control groups weighted by the OW. We use the one-year change in log GDP per capita because it is easier to interpret and produces identical results when estimating equation (12). We only show results between percentiles 1 and 99 to make the difference between the CDFs in the middle of the distribution more visible. We use the absolute value of the OW as a weight for the effective control group.
Source: Authors' analysis of data from Acemoglu et al. (2019).

a democracy throughout the sample period, and Saudi Arabia is a nondemocracy. However, both of these countries shift from the effective treatment group to the effective control group in 1991. This is driven by two forces. First, because the regression includes country fixed effects, the sum of OWs for each country is zero; this is a consequence of OLS balancing country fixed effects (cf., footnote 13). This implies that, unless a country has an OW of zero for all years, it will appear in *both* the effective treatment and control groups irrespective of its actual treatment status. Second, the year fixed effects in the regression capture the rapid rise in the early 1990s in the share of countries that are democracies; this lowers the OW as the predicted value of D_i conditional on $\mathbf{X}_i$ rises over time (Proposition 2).

Figure 2: Outcome Weights by Year for Selected Countries, Acemoglu et al. (2019)



Notes: Figure displays OWs for four countries by year. The Gambia transitions from a democracy to a nondemocracy in 1994. South Korea transitions from a nondemocracy to a democracy in 1988. USA is a democracy for the whole period. Saudi Arabia is a nondemocracy for the whole period.
Source: Authors' analysis of data from Acemoglu et al. (2019).

Identifying Observations with Influential Outcomes. Following the approach described in Section 2.5, Table 2 lists observations with the largest positive and negative outcome-influence.

We continue to use the one-year change in log GDP per capita because it is easier to interpret and does not change the estimate of β. We set the baseline value Y_i^b for each observation as the predicted value from a modified version of equation (12) that excludes the treatment variable D_{ct}, minus a one-year lag of log GDP per capita.

There are two main findings. First, no single observation plays an especially large role in the estimate. The observation with the largest positive outcome-influence is 2003 Liberia, which is in the effective (and actual) control group. Log GDP per capita in Liberia declined by 38.9 points (32 percent) in 2003, far below the baseline value of 2.41 log points. This observation drives up the estimated effect of democracy because it is in the effective control group and has a below-baseline outcome. If Liberia had instead experienced this baseline GDP growth, the overall estimate of the effect of democracy would change from 0.787 to 0.748 ($= 0.787 - 0.039$). Second, large outcome-influence observations have roughly the same impact on both the positive and negative side. Liberia in 1997, which is a nondemocracy and in the effective control group, saw growth in log GDP per capita of 65.5 points (93 percent). If this outcome were replaced by the baseline value (5.9 log points), the estimated effect of democracy would increase to 0.857. Replacing all of the top 5 outcome-influence observations on both the positive and negative side would leave the overall estimate virtually unchanged.

There are 6,336 country-year observations in the analysis sample, so the country-year analysis in Table 2 could miss certain patterns. For this reason, Table 3 zooms out and aggregates outcome-influence across countries. The countries with the five largest positive outcome-influence are Albania, Sierra Leone, Korea, Zimbabwe, and Georgia. Taken together, the magnitude of their outcome-influence implies that the estimated effect of democracy would fall from 0.787 to 0.444 if these countries experienced baseline GDP growth in every year. Put differently, a significant amount of the estimated treatment effect arises from the fact that these five countries outperformed the baseline level of GDP growth. The large outcome-influence for these countries is driven by (i) moderately large OWs, (ii) average GDP growth that exceeds the baseline for effective treat-

Table 2: Observations with Largest Positive and Negative Outcome-Influence, Acemoglu et al. (2019)

Country	Year	Democracy (1)	Outcome Weight (2)	GDP Growth (3)	Baseline Growth (4)	Outcome-Influence (5)
Panel A: Largest Positive Outcome-Influence Observations						
Liberia	2003	0	-0.0009	-38.91	2.41	0.0392
Albania	1991	0	-0.0012	-35.15	-3.08	0.0371
Guinea-Bissau	1998	0	-0.0010	-34.93	0.64	0.0342
Georgia	1992	0	-0.0007	-60.38	-10.62	0.0342
Latvia	1992	0	-0.0009	-37.36	-5.22	0.0281
Panel B: Largest Negative Outcome-Influence Observations						
Belarus	1994	1	0.0020	-12.33	-0.46	-0.0243
Bangladesh	1972	1	0.0015	-15.95	-0.01	-0.0246
Nigeria	1981	1	0.0015	-16.76	-0.28	-0.0250
Kiribati	1980	1	0.0006	-60.82	-9.17	-0.0334
Liberia	1997	0	-0.0012	65.54	5.94	-0.0696

Notes: Table reports the five country-year observations with the largest positive and negative outcome-influence following the approach in Section 2.5. Column 1 reports the value of the democracy indicator for the observation. Column 2 reports the OW. Column 3 reports the one-year change in log GDP per capita. Column 4 reports the baseline value, constructed as the predicted value from a modified version of equation (12) that excludes the democracy indicator, minus a one-year lag of log GDP per capita. Column 5 reports the outcome-influence, calculated as in equation (9).
Source: Authors' analysis of data from Acemoglu et al. (2019).

ment years, and (iii) average GDP growth that is below the baseline for effective control years.[37] The countries with the largest negative outcome-influence are Hungary, Kiribati, Paraguay, Belarus, and Tajikistan. The outcomes in these countries pull down the estimated treatment effect. The total top-five outcome-influence is almost three times as large (in absolute value) for positive outcome-influence countries as negative countries.[38]

Table 4 reports a similar analysis for years grouped into 5-year time periods (with the exception of the initial 1964–1965 window). We find that outcomes from 1991–1995 are especially

[37]None of these countries have OWs that dominate the sample. The sum of OWs in the effective treatment group is 1.4. This means that Korea, which has the largest effective treatment OW sum, accounts for only 1.5 percent of the effective treatment group. The magnitudes are identical for the effective control group by construction.

[38]This table shows that former Soviet and Communist countries played large roles in both directions. In a robustness check, Acemoglu et al. (2019) include a specification where they add interactions between a dummy for Soviet and Soviet satellite countries and dummies for the years 1989, 1990, 1991, and post-1992. This highlights how analysis like that in Table 3 might guide researchers in their robustness checks.

Table 3: Countries with Largest Positive and Negative Outcome-Influence, Acemoglu et al. (2019)

Country	Total Obs. (1)	Demo. Obs. (2)	Sum of OWs (3)	Effective Treatment Avg D_i (4)	Effective Treatment Avg Y_i (5)	Effective Treatment Avg Y_i^b (6)	Effective Control Avg D_i (7)	Effective Control Avg Y_i (8)	Effective Control Avg Y_i^b (9)	Outcome-Influence (10)
Panel A: Largest Positive Outcome-Influence Countries										
Albania	27	18	0.0096	1.00	4.73	1.97	0.00	-3.95	1.40	0.0782
Sierra Leone	47	14	0.0154	1.00	4.69	2.49	0.00	-3.65	-1.00	0.0748
Korea	47	23	0.0212	1.00	4.83	3.15	0.00	6.14	7.76	0.0701
Zimbabwe	46	9	0.0171	0.99	0.37	-1.43	0.00	-2.22	-0.28	0.0640
Georgia	42	16	0.0118	1.00	6.21	4.48	0.00	-6.51	-3.52	0.0558
Panel B: Largest Negative Outcome-Influence Countries										
Hungary	42	21	0.0159	1.00	0.59	1.37	0.00	3.30	2.78	-0.0207
Kiribati	32	32	0.0048	1.00	-10.86	-7.11	1.00	0.57	-0.01	-0.0208
Paraguay	47	18	0.0163	1.00	0.66	1.40	0.00	2.33	1.79	-0.0209
Belarus	17	1	0.0020	1.00	-12.33	-0.51	0.00	7.26	6.06	-0.0266
Tajikistan	22	0	0.0028	0.00	-14.60	-8.44	0.00	6.55	2.03	-0.0299

Notes: Table reports the five countries with the largest positive and negative outcome-influence following the approach in Section 2.5. Column 1 reports the total number of years a country appears in the data, and column 2 reports the number of years a country is a democracy. Column 3 shows the sum of OWs in the effective treatment group (which equals the sum of the absolute value of OWs in the effective control group by construction). Column 4 reports the OW-weighted share of effective treatment year observations for which a country is a democracy. Column 5 shows the OW-weighted average of the one-year change in log GDP per capita for years when the country is in the effective treatment group. Column 6 shows the OW-weighted baseline for the country when it is in the effective treatment group. The baseline value is constructed as the predicted value from a modified version of equation (12) that excludes the democracy indicator, minus a one-year lag of log GDP per capita. Columns 7–9 show analogous variables for years when a country is in the effective control group. Column 10 is the sum of the outcome-influence of all observations for the country. This can be constructed from other columns as $(3) \times [(5) - (6)] - (3) \times [(8) - (9)]$.
Source: Authors' analysis of data from Acemoglu et al. (2019).

influential, with a total outcome-influence of 0.295 (relative to an estimated effect of 0.787). The second-most influential window is 2006–2010, which has outcome-influence of 0.183. These results show that most of the estimated effect of democracy comes from GDP growth during these two periods.

Decomposing the Estimate into Subgroup Contributions. Following the approach in Section 3.2, we can calculate the degree to which different subgroups' unobserved treatment effects contribute to the overall estimated treatment effect by calculating the sum of TWs. Table 5 reports the share of different subgroups in the actual treatment group as a benchmark, along with the sum of TWs, denoted as s_g in Corollary 1.

Panel A reports heterogeneity by region. Equation (11) shows that a subgroup's total TW

Table 4: Outcome-Influence Across Years, Acemoglu et al. (2019)

Period	Total Obs. (1)	Demo. Obs. (2)	Sum of OWs (3)	Effective Treatment Avg D_i (4)	Avg Y_i (5)	Avg Y_i^b (6)	Effective Control Avg D_i (7)	Avg Y_i (8)	Avg Y_i^b (9)	Outcome-Influence (10)
1964–1965	163	64	0.0393	0.78	2.96	2.60	0.00	3.24	3.84	0.0379
1966–1970	446	158	0.1039	0.70	2.76	2.72	0.00	3.38	3.59	0.0251
1971–1975	499	172	0.1208	0.68	2.10	2.23	0.00	2.78	2.84	-0.0086
1976–1980	541	210	0.1267	0.73	1.23	1.52	0.00	1.91	1.60	-0.0759
1981–1985	618	280	0.1388	0.80	-0.03	0.21	0.00	-0.85	-0.08	0.0718
1986–1990	703	324	0.1550	0.80	1.67	1.36	0.00	0.33	0.80	0.1199
1991–1995	784	462	0.1650	0.95	-0.14	-0.62	0.20	-2.18	-0.87	0.2947
1996–2000	851	541	0.1776	0.99	1.78	1.85	0.29	2.23	2.39	0.0141
2001–2005	870	562	0.1804	0.99	2.60	2.58	0.35	1.48	2.15	0.1247
2006–2010	861	572	0.1967	0.99	2.73	2.44	0.37	1.43	2.08	0.1830

Notes: Table reports the outcome-influence by time period following the approach in Section 2.5. Column 1 reports the total number of country-year observations for each time period, and column 2 reports the number of observations that are a democracy. Column 3 shows the sum of OWs in the effective treatment group (which equals the sum of the absolute value of OWs in the effective control group by construction). Column 4 shows the OW-weighted average of the democracy indicator for observations in the effective treatment group. Column 5 shows the OW-weighted average of the one-year change in log GDP per capita for observations in the effective treatment group. Column 6 shows the OW-weighted baseline log GDP per capita for observations in the effective treatment group. The baseline value is constructed as the predicted value from a modified version of equation (12) that excludes the democracy indicator, minus a one-year lag of log GDP per capita. Columns 7–9 show analogous variables for observations in the effective control group. Column 10 is the sum of the outcome-influence of all observations for the time period. This can be constructed from other columns as $(3) \times [(5) - (6)] - (3) \times [(8) - (9)]$.
Source: Authors' analysis of data from Acemoglu et al. (2019).

depends on its sum of OWs and effective first stage. African countries account for 36 percent of total TWs, well above their 14 percent share of democracies, because both factors are large: they account for 32 percent of OWs, and their effective first stage of 0.78 exceeds the overall value of 0.71. Western Europe and other developed countries show the opposite pattern, accounting for 34 percent of democracies but only 4 percent of TWs, because both factors are small: they account for just 12 percent of OWs, and their effective first stage is 0.25.[39]

Panel B explores heterogeneity by year. These results show that each 5-year period from 1981 to 2010 accounts for a similar share of TWs. This similarity stands in contrast to the finding that years 1991–1995 and 2006–2010 have outsized outcome-influence (Table 4). These results

[39] These effective first stages reflect differences in treatment status within each region. For Western Europe and other developed countries, 100 percent of effective treatment observations are democracies compared to 75 percent of effective control observations. For African countries, the corresponding shares are 80 and 2 percent. (These numbers are OW-weighted averages.)

highlight the difference between the two measures. Outcome-influence equals the product of the OW and the difference of an outcome from a baseline value, while the TW equals the product of the OW and the treatment level. Years 1991–1995 have large outcome-influence because high-OW observations also have unusually high outcomes in the same direction (i.e., observations with large positive OWs have especially high outcomes and observations with large negative OWs have low outcomes). In contrast, the sum of TWs is not unusually large for 1991–1995 because the effective first stage is similar to surrounding time periods.

Panel C reports results when grouping observations into quintiles of one-year lagged log GDP per capita. The share of actual treatment observations is monotonically increasing in lagged GDP, reflecting a positive association between GDP and democracy. In contrast, the sum of TWs is generally decreasing in lagged GDP. Observations in the bottom quintile account for 11 percent of democracies, but 30 percent of TWs. Observations in the top quintile represent 34 percent of democracies but only 2 percent of TWs. These patterns are driven by differences in the effective first stage between these subgroups.

The effect of democracy on GDP might vary with the amount of time that has passed since the democratic transition. Equation (12) does not model this heterogeneity, but it nonetheless might exist, and our framework can be used to understand it. Treated observations fall into three categories. First, some countries are democracies for the entire sample period. Because country fixed effects are included in the regression model, the total TW of these observations is zero, and these observations' treatment effects do not contribute to the estimated effect.[40] Second, some countries start as a democracy and transition to a nondemocracy. Third, some countries start as a nondemocracy and transition to a democracy. We break this third group down further based on the number of years since the democratic transition.[41]

Panel D shows that most of the treatment effect contribution comes from the early years after a

[40]For example, the US is a democracy for the whole sample period. Some of the US observations receive a positive OW and some receive a negative OW, but the sum of these OWs is zero, which implies that s_g for the US is zero.

[41]As an example, Bangladesh has data from 1972–2010. It is categorized as a democracy from 1972–1973, 1991–2006, and 2009–2010. Only the years 1972 and 1973 are in the category of "Democracy at start." Both 1991 and 2009 are considered "0 years after democratic transition."

Table 5: Decomposition into Treatment Effects by Subgroup, Acemoglu et al. (2019)

	Share of Demo. Obs. (1)	Share of TWs (2)	Subgroup Effect (3)	S.E. (4)
Panel A: Heterogeneity by Region				
Africa	0.141	0.356	0.075	0.358
East Asia and the Pacific	0.080	0.102	2.300	0.675
Eastern Europe and Central Asia	0.106	0.133	1.594	0.761
Western Europe and Other Developed	0.330	0.043	0.627	1.015
Latin America and the Caribbean	0.292	0.290	0.722	0.390
Middle East and North Africa	0.005	0.014	2.312	1.838
South Asia	0.045	0.060	0.696	0.866
Panel B: Heterogeneity by Year				
1964–1965	0.019	0.030	-0.100	0.833
1966–1970	0.047	0.073	-0.439	0.545
1971–1975	0.051	0.082	-0.810	0.519
1976–1980	0.063	0.093	0.093	0.570
1981–1985	0.084	0.111	0.236	0.453
1986–1990	0.097	0.124	2.096	0.428
1991–1995	0.138	0.124	3.132	0.640
1996–2000	0.162	0.124	1.150	0.525
2001–2005	0.168	0.116	0.370	0.531
2006–2010	0.171	0.122	0.149	0.445
Panel C: Heterogeneity by Lag GDP Per Capita				
First quintile	0.107	0.300	0.131	0.364
Second quintile	0.128	0.235	1.454	0.422
Third quintile	0.180	0.244	0.835	0.406
Fourth quintile	0.248	0.203	0.964	0.382
Fifth quintile	0.338	0.018	0.356	0.611
Panel D: Heterogeneity by Time Since Democratic Transition				
Democracy at start	0.558	0.086	0.259	0.695
Nondemocracy at start				
0–4 years after democratic transition	0.146	0.426	0.800	0.301
5–9 years after democratic transition	0.111	0.242	0.654	0.331
10–14 years after democratic transition	0.084	0.140	1.454	0.381
15–19 years after democratic transition	0.056	0.072	0.401	0.456
20+ years after democratic transition	0.044	0.034	0.961	0.540

Notes: Table reports treatment effect decomposition following the approach in Section 3.2. Column 1 reports the share of the actual treatment group in each subgroup. Column 2 reports the share of TWs for each group as in Corollary 1. Columns 1 and 2 add up to 1 within each panel. Columns 3 and 4 report the coefficient and standard error from a regression that modifies equation (12) by interacting the treatment variable, D_{ct}, with subgroup indicators. Using column 2 as weights, the weighted average of the estimates in column 3 is identical to the main treatment effect estimate from equation (12). The standard errors in column 4 are the maximum of the standard error clustered by country and the standard error that is heteroskedasticity robust but not clustered. Source: Authors' analysis of data from Acemoglu et al. (2019).

country transitions to a democracy. The first five years after the transition account for 43 percent of TWs, but only 15 percent of treated observations. In contrast, countries that are already democracies at the start of the sample account for 56 percent of democratic observations but only 9 percent of TWs.

Columns 3 and 4 of Table 5 report point estimates and standard errors from replacing the single democracy indicator in equation (12) with a full set of interactions between the democracy indicator and subgroup indicators. As discussed following Corollary 2, the baseline treatment effect estimate is equal to a weighted average of these subgroup-specific estimates, where the weights are the TW sums shown in column 2. This result does not rely on the assumption that the interacted OLS model recovers estimates of subgroup-specific treatment effects, τ_g. However, under that additional assumption, this decomposition corresponds to that shown in Corollary 1.[42] While the subgroup-specific treatment effect estimates are imprecise, the results are still interesting. For example, even though (sub-Saharan) Africa contributes the most TWs, it has the smallest democracy coefficient. The coefficient is so small as to suggest that the estimated treatment effect depends more on the Middle East and the North of Africa—with a total TW of 0.014 and a subgroup-specific coefficient of 2.3—than on Africa—with a total TW of 0.356 and a coefficient of 0.075.

Alternative Treatment Effect Estimates. The results above show that the total TW of subgroups generally differs from their representation in the sample. Our framework not only allows researchers to understand how subgroups contribute to a treatment effect estimate, but also to construct alternative treatment effect estimates in which subgroups contribute a researcher-specified amount. Following Corollary 2, we produce treatment effect estimates for two alternative sets of weights. First, we ensure that the sum of TWs of each region matches the representation in the actual treatment group. Second, we ensure that the sum of TWs of each quintile of lagged log GDP per capita matches the actual treatment group. Columns 2 and 3 of Table 6 report point estimates that are very similar to the baseline point estimate shown in column 1. Even though there are meaningful differences between subgroups' representation in the actual treatment group and their contribution to the sum of TWs, adjusting for these differences has little net effect on the resulting

[42] A notable feature of this corollary is that we can calculate the sum of TWs, s_g, without estimating subgroup-specific treatment effects, τ_g.

Table 6: Baseline and Alternative Treatment Effect Estimates, Acemoglu et al. (2019)

	Baseline (1)	Subgroup: Region (2)	Subgroup: Lag GDP (3)	Convex: Treated Only (4)	Convex: Treated and Control (5)
Democracy	0.787	0.826	0.781	0.890	0.993
	(0.226)	(0.379)	(0.400)	(0.255)	(0.284)
Country-year observations	6,336	6,336	6,336	4,392	2,226
Countries	175	175	175	126	81
Effective Observations: Effective Control	1,325	373	812	1,106	713
Effective Observations: Effective Treatment	1,340	854	1,021	1,168	702

Notes: Table reports baseline and alternative treatment effect estimates following the approach in Section 4. Column 1 reports the baseline treatment effect estimate from Acemoglu et al. (2019). Column 2 reports an alternative estimate where the sum of TWs for each region match the representation in the treatment group; column 3 repeats this approach for quintiles of lagged log GDP per capita. Column 4 reports an alternative estimate where OWs are constrained to be nonnegative for observations in the actual treatment group. Column 5 reports an estimate where OWs are constrained to be nonnegative in the actual treatment group and nonpositive in the actual control group. The effective number of observations are calculated as in equation (6).
Source: Authors' analysis of data from Acemoglu et al. (2019).

point estimate in this setting.[43] This is a valuable robustness check. As previewed in Section 4.2, the alternative treatment effect estimates have meaningfully higher variance.

A distinct issue is that the estimate of β from equation (12) features negative OWs for 28 percent of the actual treatment group and positive OWs for 28 percent of the actual control group. A researcher might want to avoid relying on these comparisons (e.g., because they do not reflect the intuitive treatment-control logic that a researcher has in mind or because they put negative weight on the treatment effect for some observations). Corollary 3 provides one way of doing this. We report the resulting alternative estimates of β in columns 4 and 5 of Table 6. In column 4, we restrict all democracies to have nonnegative OWs. This increases the estimated effect from 0.787 (0.226) to 0.890 (0.255). The sample size falls by about 30 percent, while the standard error increases by 13 percent.[44] In column 5, we require democracies to have nonnegative weights

[43] Table 5 helps explain these results. The baseline estimate is given by the weighted sum of columns 2 and 3, while the alternative estimate is given by the weighted sum of columns 1 and 3 (for the relevant panel). For example, the alternative weight estimator puts less weight on the small estimate for sub-Saharan Africa and more weight on the moderate (and imprecise) estimate for Western Europe. The result is a slightly higher treatment effect estimate.

[44] The additional constraint assigns an OW of zero to any country that is a democracy for the entire sample period. It also removes some observations from countries that transition. For example, after the constraint is imposed, years 1993–2010 for Greece are removed (when it is a democracy but in the effective control group under the baseline approach).

and nondemocracies to have nonpositive weights.[45] This further reduces the sample size and the number of countries in the regression. Now, the regression only includes countries that switch at some point. The resulting treatment effect estimate is 0.993 (0.284).

The bottom rows of Table 6 report the effective number of observations in the effective treatment and control groups following equation (6). The results clearly show the difference between the number of observations and the effective sample size. For example, when modifying subgroup weights (columns 2 and 3), the number of observations used for estimation does not change, but the effective sample size falls considerably as the distribution of OWs becomes more dispersed. When eliminating observations with nonconvex OWs in columns 4 and 5, the number of observations falls, but the change in this count is considerably larger than the change in the effective sample size. The effective sample size is more closely linked to the precision of the estimates.

5.2 The Effect of Chinese Import Competition on US Local Labor Markets

Autor, Dorn and Hanson (2013) estimate the effects of Chinese import competition on local labor markets in the US. Their main regression model is:

$$\Delta L_{it}^{m} = \beta_1 \Delta IPW_{uit} + \mathbf{X}_{it}'\boldsymbol{\beta}_2 + \gamma_t + e_{it}, \tag{13}$$

where ΔL_{it}^{m} is the decadal change in the manufacturing employment share of the working-age population in commuting zone i; the key independent variable of interest, ΔIPW_{uit}, is the change in Chinese import exposure per worker; $\mathbf{X}_{it}$ is a vector of start-of-decade controls; and γ_t is a year fixed effect. They weight the regression using start-of-period population and instrument for ΔIPW_{uit} using a shift-share instrumental variable based on previous period local employment shares and changes in Chinese import exposure in eight other developed countries. Equation (13) stacks differences from 1990 to 2000 and 2000 to 2007.[46] We use data from their replication package, which allows us to reproduce their results exactly. The data set contains 722 commuting

[45] Appendix Figure D.3 shows the OW-weighted distribution of outcomes for the effective treatment and control groups using these OWs. The pattern is similar to the main regression, with the results being driven by less negative growth in the effective treatment group.

[46] They scale up changes from 2000 to 2007 into a ten-year equivalent change to match the 1990 to 2000 change.

zones and two time periods. We focus on their preferred specification, shown in column 6 of Table 3 of their paper. The estimate of β_1 is -0.596 (0.099).

Instead of repeating the full set of possible analyses that are done in Section 5.1, we focus this analysis more narrowly on using our framework to enhance the transparency of two-stage least squares estimates. We describe the effective treatment and control groups, assess balance on variables outside the model, report the effective first stage, characterize the industry variation behind the shift-share instrument, and identify influential observations.

Describing Effective Treatment and Control Groups. Table 7 compares the overall estimation sample with the effective treatment and control groups.[47] There are two notable takeaways. First, the effective first stage of 2.34 implies that the effective treatment group sees $2,340 higher growth in import exposure per worker than the effective control group, which has average growth in import exposure per worker of $1,660.[48] Second, some observations have relatively large OWs. The largest OW in the effective treatment group is 8.5 percent of the sum of effective treatment OWs, and the analogous magnitude for the effective control group is 5.3 percent. These results suggest that some observations could have large outcome-influence, which we examine directly below.

Describing Shift-Share Variation. Previous work studying shift-share instrumental variables like the one used by Autor, Dorn and Hanson (2013) shows how identification can rely on assumptions about the conditional exogeneity of either shares or shifts (Goldsmith-Pinkham, Sorkin and Swift, 2020; Borusyak, Hull and Jaravel, 2022). The shares-based approach relies on the assumption that commuting zones with different initial industry mixes would have evolved similarly in the absence of Chinese import competition. The shifts-based approach instead requires that industries facing larger import shocks would have evolved similarly absent import competition. Irrespective of a researcher's preferred identifying assumption, differences between the effective treatment and

[47] We focus on the overall estimation sample here because the instrument is continuous.

[48] Following Proposition 4, the main estimate of interest is equal to the OW-weighted mean difference in the outcome (-1.4) divided by the effective first stage. Differences in both the treatment and outcome variables are apparent throughout the distribution (Appendix Figure D.4).

Table 7: Means of Key Variables, Full Sample and Effective Treatment and Control Groups, Autor, Dorn and Hanson (2013)

	Full Sample (1)	Effective Control (2)	Effective Treatment (3)
Panel A: Controls in Main Model			
Starting percentage of employment in manufacturing	18.46	20.36	20.36
Starting percentage of college-educated population	50.74	50.33	50.33
Starting percentage of foreign-born population	12.41	11.29	11.29
Starting percentage of employment among women	63.98	63.93	63.93
Starting percentage of employment in routine occupations	32.05	31.15	31.15
Starting average offshorability index of occupations	0.05	-0.07	-0.07
Division: New England	0.05	0.04	0.04
Division: Middle Atlantic	0.15	0.13	0.13
Division: East North Central	0.17	0.16	0.16
Division: West North Central	0.07	0.09	0.09
Division: South Atlantic	0.18	0.20	0.20
Division: East South Central	0.06	0.09	0.09
Division: West South Central	0.11	0.09	0.09
Division: Mountain	0.06	0.06	0.06
Division: Pacific	0.15	0.13	0.13
Panel B: Variables Not in Main Model			
Starting manufacturing emp / working-age pop	11.60	11.95	12.61
Starting average log weekly wage manufacturing	6.60	6.60	6.61
Starting average log weekly wage non-manufacturing	6.44	6.42	6.42
Starting population (1000s)	3,109	2,210	1,645
Starting share unemployed	4.54	4.53	4.28
Starting share employed	70.05	69.77	69.74
Panel C: Treatment and Outcome Variables			
Change in predicted import exposure per worker [instrumental variable]	1.79	1.31	5.02
Change in import exposure per worker [endogenous variable]	1.88	1.66	4.00
Decadal change in mfg emp / pop [outcome variable]	-2.40	-2.05	-3.45
Panel D: Outcome Weights			
Share of OWs that are positive	0.39	0.00	1.00
Share of OWs that are negative	0.61	1.00	0.00
Sum of OWs	0.00	-0.43	0.43
Smallest OW ($\times$ 1000)	-22.94	-22.94	0.00
Largest OW ($\times$ 1000)	36.63	-0.00	36.63
Number of CZ-time-period observations	1,444	884	560

Notes: Column 1 reports starting-period-population-weighted means for the estimation sample of Autor, Dorn and Hanson (2013). Columns 2 and 3 report OW-weighted means for the effective control and treatment groups. OW-weighted means for the effective control group use the absolute value of the OW as a weight. Panel A displays variables that are controlled for in the main specification, and Panel B includes variables that are not in the main specification but could be. Panel C shows the instrumental, treatment, and outcome variables. Panel D reports summary statistics on OWs. The OW-weighted means of control variables in Panel A are equal by construction.

Source: Authors' analysis of data from Autor, Dorn and Hanson (2013).

Table 8: Differences in Industry Composition, Effective Treatment and Control Groups, Autor, Dorn and Hanson (2013)

Industry (SIC code)	Effective Control (1)	Effective Treatment (2)	Difference (3)
Panel A: Largest Positive Differences			
Semiconductors and Related Devices (3674)	0.0009	0.0081	0.0072
Electronic Computers (3571)	0.0007	0.0064	0.0058
Household Audio and Visual Equipment (3651)	0.0002	0.0033	0.0030
Electronic Components, NEC (3679)	0.0010	0.0040	0.0030
Guided Missiles and Space Vehicles (3761)	0.0008	0.0035	0.0026
Panel B: Largest Negative Differences			
Motor Vehicles and Passenger Car Bodies (3711)	0.0041	0.0021	-0.0020
Sawmills and Planing Mills, General (2421)	0.0045	0.0023	-0.0022
Manufacturing Industries, NEC (3999)	0.0139	0.0111	-0.0028
Paper Mills (2621)	0.0047	0.0018	-0.0029
Aircraft (3721)	0.0050	0.0010	-0.0040

Notes: Columns 1 and 2 report OW-weighted means of lagged employment shares in each manufacturing industry. Column 3 reports column 2 minus column 1. The instrument in Autor, Dorn and Hanson (2013) is constructed using 397 four-digit industries. Panel A reports the 5 industries with the largest positive difference according to column 3, and Panel B reports the 5 industries with the largest negative difference.
Source: Authors' analysis of data from Autor, Dorn and Hanson (2013).

control groups help describe the variation actually used for estimation.

Table 8 reports the four-digit industries (the level used to construct the instrument) with the largest differences in employment shares between the effective treatment and control groups. The effective treatment group is specialized in electronics manufacturing—semiconductors, computers, consumer audio and visual equipment, and electronic components—while the effective control group is specialized in autos, lumber, paper, and aircraft. These differences describe the variation used for estimation under either identifying assumption: they are the industry mixes that distinguish commuting zones in the two groups, and they point to the industries whose shocks generate the contrast in the instrument. Consequently, when assessing the plausibility of the chosen identifying assumption, these industries deserve heightened attention.[49]

[49]The top four industries in Panel A are all in the top seven highest Rotemberg weights (Goldsmith-Pinkham, Sorkin and Swift, 2020). Guided Missiles and Space Vehicles is a notable outlier, with a Rotemberg weight of zero—Chinese exports were zero in this industry—but a large positive difference between the effective treatment and control groups. Even though this industry was not exposed to trade, it is still a potential source of bias if lagged industry

Table 9: Observations with Largest Positive and Negative Outcome-Influence, Autor, Dorn and Hanson (2013)

Commuting Zone	Period	Outcome Weight (1)	Observed Outcome, Y_i (2)	Baseline Outcome, Y_i^b (3)	Outcome-Influence (4)
Panel A: Largest Positive Outcome-Influence Observations					
Detroit, MI	2000–2007	-0.009	-5.19	-2.73	0.021
Kokomo, IN	1990–2000	0.003	0.65	-4.18	0.015
Cedar Rapids, IA	2000–2007	0.003	-0.94	-2.74	0.006
Hutchinson, MN	2000–2007	0.002	-0.71	-3.95	0.005
Greensboro, NC	1990–2000	-0.003	-7.26	-5.71	0.005
Panel B: Largest Negative Outcome-Influence Observations					
New York City, NY	2000–2007	-0.023	-1.58	-2.28	-0.016
Austin, TX	2000–2007	0.007	-5.01	-2.58	-0.018
Dutchess County, NY	2000–2007	0.018	-3.33	-2.07	-0.023
Washington, DC	2000–2007	-0.019	-1.27	-2.50	-0.023
San Jose, CA	2000–2007	0.037	-5.96	-2.93	-0.111

Notes: Table reports the five commuting-zone-time-period observations with the largest positive and negative outcome-influence following the approach in Section 2.5. Column 1 reports the OW. Column 2 reports the observed change in the manufacturing employment share of the working-age population. Column 3 reports the baseline value, constructed as the predicted value from an OLS regression of a modified version of equation (13) that excludes the treatment variable. Column 4 reports the outcome-influence, calculated as in equation (9).
Source: Authors' analysis of data from Autor, Dorn and Hanson (2013).

Identifying Observations with Influential Outcomes. Table 9 identifies commuting-zone-time-period observations with the largest positive and negative outcome-influence. San Jose, CA from 2000–2007 has the largest negative outcome-influence, at -0.111. This observation is in the effective treatment group and experienced a 5.96 percentage point decrease in the manufacturing employment share, which is considerably larger than the baseline value (based on the controls in the regression) of a 2.93 p.p. decrease. As a result, this observation contributes to the negative estimated treatment effect. If San Jose had experienced its baseline outcome, the main estimate would change from -0.596 to -0.485. The total outcome-influence of the observations with the five largest negative outcome-influence is -0.191, which is almost one-third of the total treatment effect. By comparison, observations with large positive outcome-influence are less important.

shares are correlated with residual employment trends. Similarly, the industries in Panel B all had Rotemberg weights extremely close to zero, but are also potential sources of bias.

As noted above, observations with large outcome-influence do not necessarily indicate a problem with a treatment effect estimate. Instead, we view this type of outcome-influence exercise as providing a useful diagnostic tool that helps researchers identify which observations' outcomes are most important to understand.

5.3 The Effect of Education on Wages

Stephens and Yang (2014) estimate the impact of education on wages using an instrumental variable approach based on compulsory schooling laws in the US. Their main regression model is:

$$Y_{st,i} = \alpha Educ_{st,i} + \chi_s + \delta_t + \mathbf{X}'_{st,i}\boldsymbol{\beta} + \epsilon_{st,i}, \tag{14}$$

where the main outcome of interest $Y_{st,i}$ is log weekly wages for individual i born in state s in year t, $Educ_{st,i}$ is the years of schooling attained by the individual, χ_s is a birth state fixed effect, δ_t is a birth year fixed effect, and $\mathbf{X}_{st,i}$ is a vector of control variables that includes a quartic in age, indicators for the census year in which the person is observed, and an indicator for gender. The primary instruments used by Stephens and Yang (2014), denoted $\mathbf{CSL}_{st}$, are indicators for whether an individual born in state s in year t is required to attend 7, 8, or 9 or more years of schooling (the omitted category is 6 years of schooling). We use data from their replication package, which allows us to reproduce their results exactly. We focus on the sample of White men and women ages 25–54, as this is the largest sample they analyze.

Stephens and Yang (2014) note that prior research suggests the possibility that unobserved determinants of earnings vary over time across states (Card and Krueger, 1992*a*,*b*; Bleakley, 2007; Aaronson and Mazumder, 2011). As a result, in some specifications they include in $\mathbf{X}_{st,i}$ interactions between birth year and four census birth region indicators or measures of school quality from Card and Krueger (1992*a*). Their 2SLS estimate of α in equation (14) is 0.105 (95 percent confidence interval: 0.083 to 0.123) when not including region-birth-year fixed effects and −0.003 (confidence interval: −0.058 to 0.016) when including these controls. The corresponding first-stage F-statistics are 91.7 and 40.6.

We use this example to show how our framework describes what changes when controls are added. Across specifications that exclude and include region-birth-year fixed effects, we compare the effective treatment and control groups, assess covariate balance, report the effective first stage, and identify the regions with the largest outcome-influence.

Describing How Effective Treatment and Control Groups Vary with Controls. Table 10 compares the effective treatment and control groups from a specification that excludes region-birth-year fixed effects and a specification that includes them. There is little difference between the two specifications among the included control variables in Panel A. However, Panel B provides evidence of imbalance among key measures of school quality when not including the region-birth-year fixed effects. In particular, the effective treatment group has a mean pupil-teacher ratio of 27.93, which is lower (i.e., more advantaged) than the effective control group value of 29.06. Similar differences also exist for the term length and relative teacher wage. These differences are exactly what motivate Stephens and Yang (2014) to include region-birth-year fixed effects. Our approach uses the effective treatment and control group contrast to provide direct evidence that this imbalance exists when not including region-birth-year fixed effects and largely vanishes when including these fixed effects.

Describing the Effective First Stage. Panel C of Table 10 provides a useful way of describing the effective first stage (cf., Proposition 4). When region-birth-year fixed effects are not included, the difference in average years of education between the effective treatment and control groups is 0.18 years. Including region-birth-year fixed effects reduces the effective first stage by half, to 0.09 years. These fixed effects also reduce the difference in mean log weekly wages from 0.02 to 0.00, corresponding to the lack of a treatment effect in this specification.

Figure 3 provides a complementary way of visualizing the effective first stage. We plot the OW-weighted histogram of education for the effective treatment and control groups with and without region-birth-year fixed effects. Without region-birth-year fixed effects, the effective treatment and control groups differ primarily in the share of individuals with 12 years of education, with dif-

Table 10: Means of Key Variables, Effective Treatment and Control Groups, With and Without Region by Birth Year Fixed Effects, Stephens and Yang (2014)

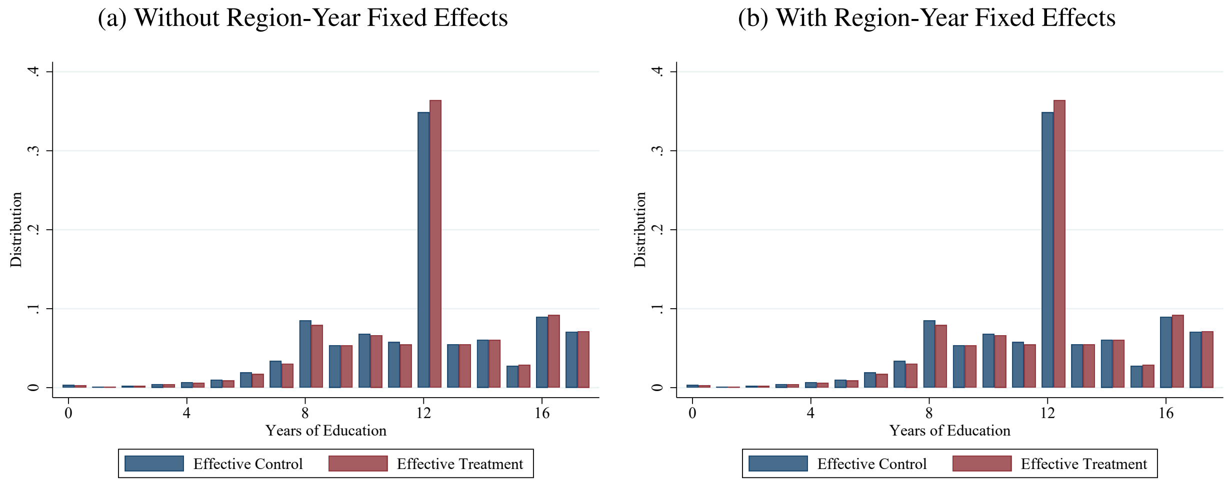
	Without Region-Year F.E.		With Region-Year F.E.	
	Effective Control (1)	Effective Treatment (2)	Effective Control (3)	Effective Treatment (4)
Panel A: Controls in Main Model				
Age	39.53	39.53	39.36	39.36
Male	0.61	0.61	0.61	0.61
Census year = 1960	0.32	0.32	0.32	0.32
Census year = 1970	0.31	0.31	0.30	0.30
Census year = 1980	0.37	0.37	0.37	0.37
Panel B: Variables Not in Main Model				
Pupil-teacher ratio	29.06	27.93	28.76	28.42
Term length, days	171.50	175.88	173.64	173.86
Relative teacher wage	0.99	1.07	1.02	1.01
Panel C: Treatment and Outcome Variables				
Compulsory years = 6 [instrument, omitted category]	0.22	0.01	0.20	0.02
Compulsory years = 7 [instrument]	0.39	0.11	0.39	0.10
Compulsory years = 8 [instrument]	0.30	0.27	0.32	0.30
Compulsory years = 9 [instrument]	0.09	0.61	0.09	0.58
Years of education [endogenous treatment]	11.72	11.90	11.75	11.84
Log weekly wage [outcome]	4.86	4.88	4.88	4.88
Panel D: Outcome Weights				
Share of OWs that are positive	0.00	1.00	0.00	1.00
Share of OWs that are negative	1.00	0.00	1.00	0.00
Sum of OWs	-5.47	5.47	-10.12	10.12
Smallest OW ($\times$ 1000)	-0.03	0.00	-0.11	0.00
Largest OW ($\times$ 1000)	-0.00	0.04	-0.00	0.10
Number of individual observations	1,840,397	1,839,826	1,917,705	1,762,425

Notes: Columns 1 and 2 report OW-weighted means for the effective control and treatment groups for equation (14) when region-birth-year fixed effects are not included in $\mathbf{X}_{st,i}$. This corresponds to column 5 of Table 1 of Stephens and Yang (2014). Columns 3 and 4 report OW-weighted means when region-birth-year fixed effects are included. This corresponds to column 6 of Table 1 of Stephens and Yang (2014). Panel A displays variables that are controlled for in the main specification, and Panel B includes variables that are not in the main specification but could be. Panel C shows the instrumental, treatment, and outcome variables. Panel D reports summary statistics on OWs. The OW-weighted means of control variables in Panel A are equal by construction. OW-weighted means for the effective control group use the absolute value of the OW as weights.
Source: Authors' analysis of data from Stephens and Yang (2014).

ferences also arising at 6 and 7 years. These differences narrow when including region-birth-year fixed effects, consistent with the overall reduction in the effective first stage.

The first-stage estimates in Stephens and Yang (2014) show an increasing, monotonic relationship between compulsory years of schooling and average educational attainment (without and with region-birth-year fixed effects), but do not reveal which requirements generate the identify-

Figure 3: OW-Weighted Histograms of Education, Effective Treatment and Control, Stephens and Yang (2014)



Notes: Figure displays OW-weighted histograms of completed years of education for the effective treatment and control groups. Panel A is based on a specification where $\mathbf{X}_{st,i}$ does not include region-birth-year fixed effects, and Panel B is based on a specification where these fixed effects are included. We use the absolute value of the OW as a weight for the effective control group.
Source: Authors' analysis of data from Stephens and Yang (2014).

ing variation. A monotonic first stage implies that observations facing the lowest requirements concentrate in the effective control group and those facing the highest requirements concentrate in the effective treatment group, with a crossover in between. Panel C of Table 10 shows that the effective treatment and control groups differ in terms of the share of observations subject to 6, 7, or 9 years of education, while there is little difference in the share subject to 8 years—the 8-year requirement is where the crossover occurs. The identifying contrast therefore comes almost entirely from comparing observations subject to a 9-year requirement with those subject to 6- or 7-year requirements.

Identifying Observations with Influential Outcomes. Table 11 describes the total outcome-influence by region. When not including region-birth-year fixed effects (Panel A), the South is a clear outlier in having very large positive outcome-influence; the value of 0.0784 implies that the estimated treatment effect would fall from 0.105 to 0.0266 if observations in the South experienced average wages predicted by covariates. Another anomalous result from this specification is that the

Table 11: Outcome-Influence by Region, Stephens and Yang (2014)

State/Region	Total Obs. (1)	Sum of OWs (2)	Effective Treatment Avg Educ. (3)	Effective Treatment Avg Y_i (4)	Effective Treatment Avg Y_i^b (5)	Effective Control Avg Educ. (6)	Effective Control Avg Y_i (7)	Effective Control Avg Y_i^b (8)	Outcome-Influence (9)
Panel A: Outcome-Influence by Region, Without Region-Year Fixed Effects									
Northeast	979,450	1.160	11.955	4.844	4.825	12.024	4.900	4.886	0.0062
South	1,114,096	2.325	11.849	5.004	5.008	10.751	4.656	4.694	0.0784
Midwest	1,178,214	1.459	11.813	4.750	4.742	12.513	5.051	5.049	0.0088
West	408,463	0.524	12.262	4.818	4.799	13.126	5.195	5.199	0.0121
Panel B: Outcome-Influence by Region, With Region-Year Fixed Effects									
Northeast	979,450	2.053	11.962	4.827	4.824	11.761	4.826	4.826	0.0064
South	1,114,096	4.477	11.350	4.821	4.824	11.288	4.821	4.825	0.0058
Midwest	1,178,214	3.029	12.301	4.948	4.945	12.186	4.949	4.944	-0.0056
West	408,463	0.565	12.889	5.114	5.131	12.954	5.121	5.122	-0.0092

Notes: Panels A and B report outcome-influence by birth region following the approach in Section 2.5. Column 1 reports the total number of individual observations in each region. Column 2 shows the sum of OWs in the effective treatment group (which equals the sum of the absolute value of OWs in the effective control group by construction). Column 3 reports the OW-weighted average years of education for effective treatment observations. Column 4 shows the OW-weighted average log weekly wage for effective treatment observations. Column 5 shows the OW-weighted average baseline value, constructed as the predicted value from an OLS regression of a modified version of equation (14) that excludes the treatment variable. Columns 6–8 show analogous variables for observations in the effective control group. Column 9 is the sum of the outcome-influence of all observations for each region. This can be constructed from other columns as $(2) \times [(4) - (5)] - (2) \times [(7) - (8)]$.
Source: Authors' analysis of data from Stephens and Yang (2014).

effective first stage is negative for all regions besides the South. This implies that the total TW for each of these regions is negative. The sum of TWs for the South is 2.55, while the sum of TWs for all other regions is -1.55. When including region-birth-year fixed effects, the South no longer has outsized outcome-influence, and the effective first stage is positive for three out of four regions.

6 Conclusion

This paper develops a framework that enhances the transparency of a large class of quasi-experimental treatment effect estimators. We show how to calculate the outcome weight (OW)—the weight that an estimator places on each observation's outcome—for models estimated by OLS, 2SLS, matching, inverse propensity weighting, regression-adjustment, doubly-robust methods, kernel regression, and marginal treatment effects. Partitioning the sample by the sign of the OW yields effective treatment and control groups, which support direct analogs of the analyses that accompany a randomized controlled trial: identifying which observations contribute positively and negatively, balance on variables that a model does not control for, and comparisons throughout the distribution of outcomes. The framework additionally delivers an effective first stage that quantifies the treatment contrast behind an estimate, and a measure of outcome-influence that identifies which observations' outcomes matter most. Combining the OW with the amount of treatment received yields the treatment effect weight (TW), which governs how unobserved, heterogeneous treatment effects aggregate into an estimate and exactly decomposes an estimate into subgroup-specific contributions. When the resulting weighting is undesirable, we show how to construct alternative estimates that satisfy researcher-specified criteria.

Our applications show that these tools yield new insights for even excellent empirical analyses. In Acemoglu et al. (2019), 28 percent of democracies contribute negatively to the estimated effect of democracy, five countries account for more than 40 percent of the estimate, and the positive mean effect comes from fewer episodes of declining GDP rather than more episodes of rapid growth. In Autor, Dorn and Hanson (2013), five commuting-zone-year observations account for a third of the estimated effect of Chinese import competition on US manufacturing. In Stephens and Yang (2014), the effective first stage is small enough to raise questions about how much compulsory schooling laws reveal about the returns to education, even though the first-stage F-statistic is large. These takeaways are not evident from commonly reported results, but they are straightforward to produce using the OW and TW. The transparency-generating analyses outlined in this paper are a natural complement to quasi-experimental estimates and have the potential to improve the credibility and understanding of such estimates.

References

Aaronson, Daniel, and Bhashkar Mazumder. 2011. "The Impact of Rosenwald Schools on Black Achievement." *Journal of Political Economy*, 119(5): 821–888.

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica*, 74(1): 235–267.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy*, 127(1): 47–100.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica*, 66(2): 249–288.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Chapter 23 - Empirical Strategies in Labor Economics." In . Vol. 3A of *Handbook of Labor Economics*, ed. Orley C. Ashenfelter and David Card, 1277–1366. Elsevier.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies*, 67(3): 499–527.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science*, 60(1): 250–267.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review*, 103(6): 2121–68.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley and Sons.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When is TSLS Actually LATE?" National Bureau of Economic Research Working Paper 29709.

Bleakley, Hoyt. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics*, 122: 73–117.

Borusyak, Kirill, and Peter Hull. 2024. "Negative Weights Are No Concern in Design-Based Specifications." *AEA Papers and Proceedings*, 114: 597–600.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-Experimental Shift-Share Research Designs." *Review of Economic Studies*, 89(1): 181–213.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *Review of Economic Studies*, 91(6): 3253–3285.

Broderick, Tamara, Ryan Giordano, and Rachael Meager. 2020. "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?" arXiv Working Paper arXiv:2011.14999.

Card, David, and Alan B. Krueger. 1992*a*. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy*, 100(1): 1–40.

Card, David, and Alan B. Krueger. 1992*b*. "School Quality and Black-White Relative Earnings: A Direct Assessment." *Quarterly Journal of Economics*, 107(1): 151–200.

Chattopadhyay, Ambarish, and José R. Zubizarreta. 2023. "On the Implied Weights of Linear Regression for Causal Inference." *Biometrika*, 110(3): 615–629.

Chen, Jiafeng. 2026. "Potential Weights and Implicit Causal Designs in Linear Regression." Working Paper.

Cook, R. Dennis. 1977. "Detection of Influential Observation in Linear Regression." *Technometrics*, 19(1): 15–18.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review*, 110(9): 2964–2996.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2023. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." *The Econometrics Journal*, 26(3): C1–C30.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. "Unconditional Quantile Regres-

sions." *Econometrica*, 77(3): 953–973.
Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica*, 1(4): 387–401.
Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik Instruments: What, When, Why, and How." *American Economic Review*, 110(8): 2586–2624.
Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. "Contamination Bias in Linear Regressions." *American Economic Review*, 114(12): 4015–51.
Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, 225(2): 254–277. Themed Issue: Treatment Effect 1.
Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 20(1): 25–46.
Heckman, James J., and Edward Vytlacil. 2001. "Policy-Relevant Treatment Effects." *American Economic Review*, 91(2): 107–111.
Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.
Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts." *Review of Economic Studies*, 64(4): 487–535.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.
Knaus, Michael C. 2024. "Treatment Effect Estimators as Weighted Outcomes." arXiv Working Paper arXiv:2411.11559.
Lovell, Michael C. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association*, 58(304): 993–1010.
Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2021. "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables." *American Economic Review*, 111(11): 3663–3698.
Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." *Econometrica*, 86(5): 1589–1619.
Stephens, Melvin, Jr., and Dou-Yan Yang. 2014. "Compulsory Education and the Benefits of Schooling." *American Economic Review*, 104(6): 1777–92.
Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, 225(2): 175–199. Themed Issue: Treatment Effect 1.
Słoczyński, Tymon. 2022. "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights." *Review of Economics and Statistics*, 104(3): 501–509.
Wald, Abraham. 1940. "The Fitting of Straight Lines if Both Variables are Subject to Error." *Annals of Mathematical Statistics*, 11: 284–300.
Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business and Economic Statistics*, 14(4): 478–486.
Young, Alwyn. 2022. "Consistency without Inference: Instrumental Variables in Practical Application." *European Economic Review*, 147: 104112.

Zubizarreta, José R. 2015. “Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data.” *Journal of the American Statistical Association*, 110(511): 910–922.

Online Appendix

A Proofs

Proof of Proposition 1

Proof of Proposition 1(A). The mean-invariance property means that:

$$\hat{\theta} = \sum_i \omega_i Y_i = \sum_i \omega_i (Y_i + \kappa),$$

for some constant $\kappa \neq 0$. Rewriting this yields the desired result:

$$\sum_i \omega_i Y_i = \sum_i \omega_i Y_i + \kappa \sum_i \omega_i \Rightarrow \sum_i \omega_i = 0.$$

□

Proof of Proposition 1(B). The result follows immediately from Proposition 1(A) and Definition 2:

$$0 = \sum_i \omega_i = \sum_{i \in \mathcal{T}_E} \omega_i - \sum_{i \in \mathcal{C}_E} |\omega_i| \quad \Rightarrow \quad \sum_{i \in \mathcal{T}_E} \omega_i = \sum_{i \in \mathcal{C}_E} |\omega_i|.$$

□

Proof of Proposition 2

Proof. By the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963), we can estimate θ in equation (7) using the following regression:

$$Y_i = \tilde{D}_i \theta + \epsilon_i, \tag{A.1}$$

where $\tilde{D}_i = D_i - \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}$ is the residual from regressing D_i on $\mathbf{X}_i$. The OLS estimator of θ from equation (A.1) is $\hat{\theta} = (\tilde{\mathbf{D}}'\tilde{\mathbf{D}})^{-1}\tilde{\mathbf{D}}'\mathbf{Y}$. From equation (5), we can calculate the OW as $\omega_i = \partial\hat{\theta}/\partial Y_i$. This gives us:

$$\omega_i = (\tilde{\mathbf{D}}'\tilde{\mathbf{D}})^{-1}\tilde{D}_i = \frac{\tilde{D}_i}{\sum_j \tilde{D}_j^2}.$$

□

Proof of Proposition 3

Proof. Plugging in the predicted first-stage value, we can estimate θ using OLS on the following equation:

$$Y_i = \hat{D}_i \theta + \mathbf{X}_i' \boldsymbol{\beta} + \epsilon_i.$$

By the Frisch-Waugh-Lovell theorem, we can estimate θ using OLS on the following regression:

$$Y_i = (\hat{D}_i - \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}})\theta + \epsilon_i. \tag{A.2}$$

Estimated OLS residuals are uncorrelated with control variables, which implies:

$$\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\hat{\mathbf{D}} + \hat{\mathbf{u}}) = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{D}},$$

where $\hat{\mathbf{u}} = \mathbf{D} - \hat{\mathbf{D}}$ is the stacked vector of residuals from the regression of D_i on $\mathbf{X}_i$ and $\mathbf{Z}_i$. The second equality holds because $\mathbf{X}$ is included in the first-stage regression, so $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$. This allows us to rewrite equation (A.2) as:

$$Y_i = (\hat{D}_i - \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D})\theta + \epsilon_i \equiv (\hat{D}_i - \breve{D}_i)\theta + \epsilon_i. \tag{A.3}$$

The 2SLS estimator of θ from equation (A.3) is $\hat{\theta} = ((\hat{\mathbf{D}} - \breve{\mathbf{D}})'(\hat{\mathbf{D}} - \breve{\mathbf{D}}))^{-1}(\hat{\mathbf{D}} - \breve{\mathbf{D}})'\mathbf{Y}$, which immediately implies

$$\omega_i = ((\hat{\mathbf{D}} - \breve{\mathbf{D}})'(\hat{\mathbf{D}} - \breve{\mathbf{D}}))^{-1}(\hat{D}_i - \breve{D}_i) = \frac{\hat{D}_i - \breve{D}_i}{\sum_j (\hat{D}_j - \breve{D}_j)^2}.$$

□

Statement and Proof of Proposition A.1

The following proposition is described informally in footnote 13.

Proposition A.1 (OLS and 2SLS Covariate Balance)**.** For all of the control variables in $\mathbf{X}$ in equation (7), the OW for an OLS estimate of $\hat{\theta}$ ensures that OW-weighted mean values for the effective treatment and control groups are equal:

$$\bar{x}_{\mathcal{T}_E} = \bar{x}_{\mathcal{C}_E}, \quad \forall x \in \mathbf{X},$$

where OW-weighted means are constructed as in equation (8). The same property applies to a 2SLS estimate of $\hat{\theta}$.

First, we prove a lemma.

Lemma A.1. Consider OLS or 2SLS estimation of θ in equation (7). For any $x \in \mathbf{X}$,

$$\sum_i \omega_i x_i = 0.$$

Proof. For OLS, we have:

$$\sum_i \omega_i x_i = \sum_i \frac{\tilde{D}_i}{\sum_j \tilde{D}_j^2} x_i = 0,$$

where the last equality holds because residuals from the regression of D_i on $\mathbf{X}_i$ are orthogonal to each element $x \in \mathbf{X}$ by construction.

For 2SLS, we similarly have:

$$\sum_i \omega_i x_i = \sum_i \frac{\hat{D}_i - \breve{D}_i}{\sum_j (\hat{D}_j - \breve{D}_j)^2} x_i = 0.$$

The last equality follows because every column of $\mathbf{X}$ is also a column of $\mathbf{Q} = (\mathbf{X}, \mathbf{Z})$, which implies that $\hat{D}_i - \breve{D}_i$ is orthogonal to every $x \in \mathbf{X}$ because both projections map any $x \in \mathbf{X}$ to itself. □

Now, we prove Proposition A.1.

Proof.

$$\bar{x}_{\mathcal{T}_E} - \bar{x}_{\mathcal{C}_E} = \sum_{i \in \mathcal{T}_E} \frac{\omega_i}{\Omega} x_i - \sum_{i \in \mathcal{C}_E} \frac{|\omega_i|}{\Omega} x_i = \frac{1}{\Omega} \sum_i \omega_i x_i = 0,$$

where the last equality comes from Lemma A.1. □

Proof of Proposition 4

First, we prove the statement in Proposition 4 that the weighted treatment sum condition ($\sum_i \omega_i D_i = 1$) is satisfied by OLS, 2SLS, WLS, matching, normalized IPW, regression-adjustment, and doubly-robust estimators. We prove this separately for each estimator, using the OWs derived in Appendix B for the estimators that are not covered in the main text.

Proof for OLS. Using the OW derived in Proposition 2, we have

$$\begin{aligned}
\sum_i \omega_i D_i &= \sum_i \frac{\tilde{D}_i D_i}{\sum_j \tilde{D}_j^2} \\
&= \sum_i \frac{\tilde{D}_i(\tilde{D}_i + \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D})}{\sum_j \tilde{D}_j^2} = \sum_i \frac{\tilde{D}_i^2 + \tilde{D}_i \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}}{\sum_j \tilde{D}_j^2} \\
&= \sum_i \frac{\tilde{D}_i^2}{\sum_j \tilde{D}_j^2} = 1,
\end{aligned}$$

where the second line uses the standard result that D_i can be written as the sum of the residual from a regression of D_i on $\mathbf{X}_i$ and the predicted value, and the third line uses the standard result that the predicted value and residual are uncorrelated.

□

Proof for 2SLS. Using the OW derived in Proposition 3, we have

$$\begin{aligned}\sum_i \omega_i D_i &= \sum_i \frac{(\hat{D}_i - \breve{D}_i)D_i}{\sum_j (\hat{D}_j - \breve{D}_j)^2} = \frac{\sum_i \hat{D}_i D_i - \breve{D}_i D_i}{\sum_j \hat{D}_j^2 - 2\hat{D}_j \breve{D}_j + \breve{D}_j^2} \\ &= \frac{\sum_i \hat{D}_i(\hat{D}_i + \hat{e}_i) - \breve{D}_i(\breve{D}_i + \breve{e}_i)}{\sum_j \hat{D}_j^2 - 2(\breve{D}_j + \breve{e}_j - \hat{e}_j)\breve{D}_j + \breve{D}_j^2} = \frac{\sum_i \hat{D}_i^2 + \hat{D}_i \hat{e}_i - \breve{D}_i^2 - \breve{D}_i \breve{e}_i}{\sum_j \hat{D}_j^2 - \breve{D}_j^2 - 2\breve{D}_j \breve{e}_j + 2\breve{D}_j \hat{e}_j} \\ &= \frac{\sum_i \hat{D}_i^2 - \breve{D}_i^2}{\sum_j \hat{D}_j^2 - \breve{D}_j^2} = 1,\end{aligned}$$

where $\hat{e}_i = D_i - \hat{D}_i$ and $\breve{e}_i = D_i - \breve{D}_i$ are the first-stage residuals corresponding to the predicted values $\hat{D}_i$ and $\breve{D}_i$ defined in Proposition 3. The second line uses the standard result that D_i can be written as the sum of a predicted value and a residual ($D_i = \hat{D}_i + \hat{e}_i = \breve{D}_i + \breve{e}_i$), and the third line uses the standard result that the predicted value and residual are uncorrelated ($0 = \sum_i \hat{D}_i \hat{e}_i = \sum_i \breve{D}_i \breve{e}_i$), along with $\sum_i \breve{D}_i \hat{e}_i = 0$, which follows because $\hat{e}_i$ is orthogonal to all variables in $\mathbf{X}_i$ and $\breve{D}_i$ is a linear combination of $\mathbf{X}_i$. □

Proof for WLS. Using the OW derived in Proposition B.1, and writing the treatment variable as the sum of the weighted residual and the weighted predicted value, $D_i = \mathring{D}_i + \mathbf{X}_i'(\mathbf{X}'\mathbf{\Phi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}\mathbf{D}$, we have:

$$\sum_i \omega_i D_i = \frac{\sum_i \phi_i \mathring{D}_i \left(\mathring{D}_i + \mathbf{X}_i'(\mathbf{X}'\mathbf{\Phi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}\mathbf{D}\right)}{\sum_j \phi_j (\mathring{D}_j)^2} = \frac{\sum_i \phi_i (\mathring{D}_i)^2 + \left(\mathring{\mathbf{D}}'\mathbf{\Phi}\mathbf{X}\right)(\mathbf{X}'\mathbf{\Phi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}\mathbf{D}}{\sum_j \phi_j (\mathring{D}_j)^2} = 1,$$

where the last equality uses $\mathbf{X}'\mathbf{\Phi}\mathring{\mathbf{D}} = \mathbf{X}'\mathbf{\Phi}\mathbf{D} - \mathbf{X}'\mathbf{\Phi}\mathbf{X}(\mathbf{X}'\mathbf{\Phi}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Phi}\mathbf{D} = \mathbf{0}$, which is the weighted analog of the orthogonality between residuals and regressors used in the proof for OLS. □

Proof for Matching. Using the OW derived in Proposition B.2, we have

$$\sum_i \omega_i D_i = \sum_{i:D_i=1} \frac{1}{n_1} = 1.$$

□

Proof for Normalized IPW. Using the OW derived in Proposition B.3 for the ATE, we have

$$\sum_i \omega_i D_i = \sum_{i:D_i=1} \frac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} = 1.$$

For the ATT, we have

$$\sum_i \omega_i D_i = \sum_{i:D_i=1} \frac{D_i}{\sum_j D_j} = 1.$$

□

The proofs for the regression-adjustment and doubly-robust estimators use the following lemma.

Lemma A.2 (Projection of a Constant)**.** Let $\mathbf{X}_d$ denote the $n_d \times (k-1)$ matrix that stacks $\mathbf{X}_i'$ across the n_d observations with $D_i = d$, where $\mathbf{X}_i$ includes a constant, and let $\boldsymbol{\iota}_d$ denote the $n_d \times 1$ vector of ones, so that $\mathbf{X}_d'\boldsymbol{\iota}_d = \sum_{i:D_i=d} \mathbf{X}_i$. Then

$$(\mathbf{X}_d'\mathbf{X}_d)^{-1}\mathbf{X}_d'\boldsymbol{\iota}_d = \mathbf{e}_c,$$

where $\mathbf{e}_c$ is the $(k-1) \times 1$ vector with a one in the position of the constant term and zeros elsewhere. Consequently, $\mathbf{a}'\mathbf{e}_c$ equals the element of $\mathbf{a}$ corresponding to the constant term, for any conformable vector $\mathbf{a}$.

Proof of Lemma A.2. The expression is the coefficient vector from an OLS regression of a vector of ones on $\mathbf{X}_d$. Because $\mathbf{X}_i$ includes a constant, this regression fits perfectly with an intercept of one and a coefficient of zero on every other variable. This coefficient vector is unique because $\mathbf{X}_d$ has full rank. □

Proof for Regression Adjustment. Using the OW derived in Proposition B.4 for the ATE, we have

$$\begin{aligned}
\sum_i \omega_i D_i &= \sum_{i:D_i=1} \frac{1}{n} \sum_j \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_i = \frac{1}{n} \sum_j \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1} \sum_{i:D_i=1} \mathbf{X}_i \\
&= \frac{1}{n} \sum_j \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\iota}_1 = \frac{1}{n} \sum_j \mathbf{X}_j'\mathbf{e}_c = \frac{1}{n} \sum_j 1 = 1,
\end{aligned}$$

where the third equality relies on the definitions of $\mathbf{X}_1$ and $\boldsymbol{\iota}_1$, the fourth equality uses Lemma A.2, and the fifth equality uses the fact that the element of $\mathbf{X}_j$ corresponding to the constant term equals one for every observation. For the ATT, the same steps give

$$\sum_i \omega_i D_i = \frac{1}{n_1} \sum_{j:D_j=1} \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\iota}_1 = \frac{1}{n_1} \sum_{j:D_j=1} \mathbf{X}_j'\mathbf{e}_c = \frac{1}{n_1} \sum_{j:D_j=1} 1 = 1.$$

□

Proof for Doubly-Robust. Using the OW derived in Proposition B.5, and again using $\sum_{i:D_i=1} \mathbf{X}_i = \mathbf{X}_1'\boldsymbol{\iota}_1$, we have

$$\begin{aligned}
\sum_i \omega_i D_i &= \sum_{i:D_i=1} \left[\frac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} + \left(\frac{1}{n} \sum_j \mathbf{X}_j' - \frac{\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j'}{\sum_j \frac{D_j}{\hat{P}_j}} \right) (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_i \right] \\
&= \frac{\sum_{i:D_i=1} \frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} + \left(\frac{1}{n} \sum_j \mathbf{X}_j' - \frac{\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j'}{\sum_j \frac{D_j}{\hat{P}_j}} \right) (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\iota}_1 \\
&= 1 + \left(\frac{1}{n} \sum_j \mathbf{X}_j' - \frac{\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j'}{\sum_j \frac{D_j}{\hat{P}_j}} \right) \mathbf{e}_c = 1 + (1-1) = 1.
\end{aligned}$$

The second equality collects terms as in the proof for regression adjustment. The first term in the third equality holds because D_j is zero for untreated observations, and the second term uses Lemma A.2. The final equality uses the fact that the element of $\mathbf{X}_j$ corresponding to the constant term equals one for every observation, so that $\frac{1}{n}\sum_j \mathbf{X}_j'\mathbf{e}_c = 1$ and $\left(\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j'\mathbf{e}_c\right) / \left(\sum_j \frac{D_j}{\hat{P}_j}\right) = 1$. □

Next, we prove a lemma that is of some independent interest and is used in the proof of the proposition.

Lemma A.3 (Mean Difference in Treatment Variable)**.** For estimators that satisfy the weighted treatment sum condition ($\sum_i \omega_i D_i = 1$), the average difference in the treatment variable between the effective treatment and control groups can be written:

$$\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E} = \frac{1}{\Omega}.$$

Proof.

$$\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E} = \sum_{i\in\mathcal{T}_E} \frac{\omega_i}{\Omega} D_i - \sum_{i\in\mathcal{C}_E} \frac{|\omega_i|}{\Omega} D_i = \frac{1}{\Omega}\left(\sum_i \omega_i D_i\right) = \frac{1}{\Omega},$$

where the first equality uses the definitions from equation (8) and the last equality uses the weighted treatment sum condition. □

Finally, we prove Proposition 4.

Proof.

$$\hat{\theta} = \sum_{i\in\mathcal{T}_E} \omega_i Y_i - \sum_{i\in\mathcal{C}_E} |\omega_i| Y_i = \Omega\left(\sum_{i\in\mathcal{T}_E} \frac{\omega_i}{\Omega} Y_i - \sum_{i\in\mathcal{C}_E} \frac{|\omega_i|}{\Omega} Y_i\right) = \Omega\left(\bar{Y}_{\mathcal{T}_E} - \bar{Y}_{\mathcal{C}_E}\right) = \frac{\bar{Y}_{\mathcal{T}_E} - \bar{Y}_{\mathcal{C}_E}}{\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E}},$$

where the first equality uses the fact that the outcomes of observations with $\omega_i = 0$ do not contribute to $\hat{\theta}$, the third equality uses the definitions from equation (8), and the last equality uses Lemma A.3. □

The weighted treatment sum condition is not required for the entire argument above, only for the final step. For an estimator that does not satisfy it, the proof of Lemma A.3 gives $\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E} = \Omega^{-1}\sum_i \omega_i D_i$, so the same steps yield a Wald representation up to a scale factor:

$$\hat{\theta} = \left(\sum_i \omega_i D_i\right) \frac{\bar{Y}_{\mathcal{T}_E} - \bar{Y}_{\mathcal{C}_E}}{\bar{D}_{\mathcal{T}_E} - \bar{D}_{\mathcal{C}_E}}. \tag{A.4}$$

The kernel regression estimator in Appendix B illustrates why the scale factor can differ from one. Using the OW from Proposition B.6, $\sum_i \omega_i D_i$ is the difference between the kernel-weighted mean of the treatment variable near D_T and near D_U, which is approximately $D_T - D_U$ rather than one. This is as expected, because the corresponding parameter is the difference in expected outcomes between two treatment levels rather than the effect of a one-unit increase in treatment.

Proof of Proposition 5

Proof of Proposition 5. Starting from equation (5), we can use the potential outcomes notation to write the estimated treatment effect as:

$$\hat{\theta} = \sum_i \omega_i Y_i(0) + \sum_i \omega_i D_i \tau_i = \Omega\left(\sum_{i\in\mathcal{T}_E} \frac{\omega_i}{\Omega} Y_i(0) - \sum_{i\in\mathcal{C}_E} \frac{|\omega_i|}{\Omega} Y_i(0)\right) + \sum_i \omega_i D_i \tau_i \equiv \Delta + \sum_i TW_i \tau_i$$

□

Proof of Corollary 1

Proof. Starting from Proposition 5, we have:

$$\hat{\theta} = \Delta + \sum_{g=1}^{G}\left(\sum_{i:g_i=g} TW_i\tau_i\right) = \Delta + \sum_{g=1}^{G}\left(\left[\sum_{i:g_i=g} TW_i\right]\sum_{i:g_i=g}\frac{TW_i\tau_i}{\sum_{i':g_{i'}=g} TW_{i'}}\right) \equiv \Delta + \sum_{g=1}^{G} s_g \bar{\tau}_g,$$

where the first equality partitions the sum in Proposition 5 across the mutually-exclusive and collectively-exhaustive groups, and the second equality multiplies and divides the gth term by $s_g = \sum_{i:g_i=g}\omega_i D_i$. The second step requires $s_g \neq 0$; a group with $s_g = 0$ makes no contribution to $\hat{\theta}$, so its term can be dropped from the sum and $\bar{\tau}_g$ need not be defined.

□

Proof of Proposition 6

Proof. There are two key elements of the proof. First, we provide an alternative representation of the OLS OW. Second, we show that the solution to the constrained minimization problem in Proposition 6 coincides with this alternative representation.

Characterizing the OLS OW. Let $\mathbf{C}$ be a $k \times 1$ vector, where the first element is 1 and all other elements are 0. Let $\mathbf{W}_i = (D_i, \mathbf{X}_i')'$ be the $k\times 1$ vector containing the treatment variable of interest and the vector of controls, with $\mathbf{W}$ denoting the $n \times k$ matrix that is stacked across observations. The estimated coefficient on D_i from a regression of Y_i on $\mathbf{W}_i$ is:

$$\hat{\theta} = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}. \tag{A.5}$$

Given this, the OLS OW for $\hat{\theta}$ can be written:

$$\omega_i = \frac{\partial\hat{\theta}}{\partial Y_i} = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i. \tag{A.6}$$

The constrained minimization problem. The Lagrangian for the constrained optimization problem in Proposition 6 can be written:

$$\mathcal{L} = \frac{1}{2}\sum_i \omega_i^2 - \lambda_D\left(\sum_i \omega_i D_i - 1\right) - \boldsymbol{\lambda}_X'\left(\sum_i \omega_i \mathbf{X}_i\right), \tag{A.7}$$

where $\boldsymbol{\lambda}_X$ has dimension $(k-1)\times 1$. The first-order condition produces an expression for the OW that includes the Lagrange multipliers:

$$\frac{\partial \mathcal{L}}{\partial \omega_i} = \omega_i - \lambda_D D_i - \boldsymbol{\lambda}_X'\mathbf{X}_i = 0 \Rightarrow \omega_i = \lambda_D D_i + \boldsymbol{\lambda}_X'\mathbf{X}_i. \tag{A.8}$$

Let $\boldsymbol{\Lambda} = (\lambda_D, \boldsymbol{\lambda}_X')'$, which is $k \times 1$. We can write the $n \times 1$ vector of OWs as

$$\boldsymbol{\omega} = \mathbf{W}\boldsymbol{\Lambda}. \tag{A.9}$$

The objective function is strictly convex and the constraints are linear, so the first-order condition is necessary and sufficient for a unique global minimum, provided that the feasible set is nonempty. This holds under the maintained assumption that $\mathbf{W}$ has full rank. Writing the two constraints in matrix notation, we also have:

$$\boldsymbol{\omega}'\mathbf{W} = \mathbf{C}'. \tag{A.10}$$

Plugging equation (A.9) into equation (A.10), we have:

$$\boldsymbol{\Lambda}'\mathbf{W}'\mathbf{W} = \mathbf{C}' \Rightarrow \boldsymbol{\Lambda}' = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1} \Rightarrow \boldsymbol{\Lambda} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{C} \tag{A.11}$$

Plugging this back into equation (A.9) yields:

$$\boldsymbol{\omega} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C}, \tag{A.12}$$

which gives us:

$$\omega_i = \mathbf{W}_i'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C} = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i. \tag{A.13}$$

This is identical to the expression for the OLS OW in equation (A.6).

□

Statement and Proof of Proposition A.2

The following proposition provides an analogous result to Proposition 6 for 2SLS.

Proposition A.2 (Constrained Optimization Representation of 2SLS)**.** The 2SLS OW in Proposition 3 can be obtained by solving the following problem:

$$\min_{\{\omega_i\},\gamma_X,\gamma_Z} \sum_i \omega_i^2 \quad s.t. \sum_i \omega_i D_i = 1\ ,\ \sum_i \omega_i \mathbf{X}_i = \mathbf{0}\ ,\ \omega_i = \boldsymbol{\gamma}_X'\mathbf{X}_i + \boldsymbol{\gamma}_Z'\mathbf{Z}_i.$$

Proof. First, we provide an alternative representation of the 2SLS OW. Second, we show that the solution to the constrained minimization problem in Proposition A.2 is the same as this representation.

Characterizing the 2SLS OW. Let $\mathbf{C}$ be a $k \times 1$ vector, where the first element is 1 and all other elements are 0. Let $\mathbf{W}_i = (D_i, \mathbf{X}_i')'$ be a $k \times 1$ vector containing the treatment variable of interest and the vector of controls, with $\mathbf{W}$ denoting the $n \times k$ matrix that is stacked across observations. Let $\mathbf{Q}_i = (\mathbf{X}_i', \mathbf{Z}_i')'$ be a $(k-1+p) \times 1$ vector containing the exogenous controls and excluded instruments, with $\mathbf{Q}$ denoting the $n \times (k-1+p)$ matrix that is stacked across observations. The estimated coefficient on D_i from a 2SLS regression is:

$$\hat{\theta} = \mathbf{C}'(\mathbf{W}'\mathbf{P_Q}\mathbf{W})^{-1}\mathbf{W}'\mathbf{P_Q}\mathbf{Y},$$

where $\mathbf{P_Q} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$. Given this, the 2SLS OW for $\hat{\theta}$ can be written:

$$\omega_i = \frac{\partial \hat{\theta}}{\partial Y_i} = \mathbf{C}'(\mathbf{W}'\mathbf{P_Q}\mathbf{W})^{-1}\mathbf{W}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}_i. \tag{A.14}$$

The constrained minimization problem. The Lagrangian for the constrained optimization problem in Proposition A.2 can be written:

$$\mathcal{L} = \frac{1}{2}\sum_i \omega_i^2 - \lambda_D\left(\sum_i \omega_i D_i - 1\right) - \boldsymbol{\lambda}_X'\left(\sum_i \omega_i \mathbf{X}_i\right) - \sum_i \lambda_i(\omega_i - \boldsymbol{\gamma}_X'\mathbf{X}_i - \boldsymbol{\gamma}_Z'\mathbf{Z}_i), \tag{A.15}$$

where $\boldsymbol{\lambda}_X$ has dimension $(k-1)\times 1$ and λ_D and λ_i are scalars. The first-order condition produces an expression for the OW that includes the Lagrange multipliers:

$$\frac{\partial \mathcal{L}}{\partial \omega_i} = \omega_i - \lambda_D D_i - \boldsymbol{\lambda}_X'\mathbf{X}_i - \lambda_i = 0 \Rightarrow \omega_i = \lambda_D D_i + \boldsymbol{\lambda}_X'\mathbf{X}_i + \lambda_i.$$

The first-order conditions for $\boldsymbol{\gamma}_X$ and $\boldsymbol{\gamma}_Z$ imply:

$$\sum_i \lambda_i \mathbf{X}_i = \mathbf{0}\ ,\ \sum_i \lambda_i \mathbf{Z}_i = \mathbf{0}.$$

Let $\boldsymbol{\Lambda} = (\lambda_D, \boldsymbol{\lambda}_X')'$. Let $\boldsymbol{\lambda}$ be the $n \times 1$ vector of stacked λ_i and $\boldsymbol{\omega}$ be the $n \times 1$ vector of OWs. We can write all three first-order conditions as:

$$\boldsymbol{\omega} = \mathbf{W}\boldsymbol{\Lambda} + \boldsymbol{\lambda} \tag{A.16}$$
$$\boldsymbol{\lambda}'\mathbf{Q} = \mathbf{0}. \tag{A.17}$$

Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_X', \boldsymbol{\gamma}_Z')'$. We can also rewrite the constraints in matrix notation:

$$\boldsymbol{\omega}'\mathbf{W} = \mathbf{C}' \tag{A.18}$$
$$\boldsymbol{\omega} = \mathbf{Q}\boldsymbol{\gamma}. \tag{A.19}$$

Pre-multiplying equation (A.16) by $\mathbf{P_Q}$ yields:

$$\mathbf{P_Q}\boldsymbol{\omega} = \mathbf{P_Q}\mathbf{W}\boldsymbol{\Lambda} + \mathbf{P_Q}\boldsymbol{\lambda} = \mathbf{P_Q}\mathbf{W}\boldsymbol{\Lambda}, \tag{A.20}$$

where the last equality follows from equation (A.17).

Next, equation (A.19) implies that $\mathbf{P_Q}\boldsymbol{\omega} = \boldsymbol{\omega}$, which simplifies the above equation to:

$$\boldsymbol{\omega} = \mathbf{P_Q}\mathbf{W}\boldsymbol{\Lambda}. \tag{A.21}$$

Finally, pre-multiplying by $\mathbf{W}'$ yields:

$$\mathbf{W}'\boldsymbol{\omega} = \mathbf{W}'\mathbf{P_Q}\mathbf{W}\boldsymbol{\Lambda}. \tag{A.22}$$

Plugging equation (A.22) into equation (A.18), we have:

$$\mathbf{W}'\mathbf{P_Q}\mathbf{W}\boldsymbol{\Lambda} = \mathbf{C} \Rightarrow \boldsymbol{\Lambda} = (\mathbf{W}'\mathbf{P_Q}\mathbf{W})^{-1}\mathbf{C}. \tag{A.23}$$

This step assumes that $\mathbf{W}'\mathbf{P_Q}\mathbf{W}$ is invertible, which is implied by standard rank conditions in 2SLS. These rank conditions also ensure that the feasible set is nonempty, and the objective function is strictly convex with linear constraints, so the first-order conditions characterize a unique global minimum.

Plugging this back into equation (A.21) yields:

$$\boldsymbol{\omega} = \mathbf{P_Q}\mathbf{W}(\mathbf{W}'\mathbf{P_Q}\mathbf{W})^{-1}\mathbf{C}, \tag{A.24}$$

which gives us:

$$\omega_i = \mathbf{C}'(\mathbf{W}'\mathbf{P_Q}\mathbf{W})^{-1}\mathbf{W}'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}_i. \tag{A.25}$$

This is identical to the expression for the 2SLS OW in equation (A.14).

□

Statement and Proof of Equivalence between Interacted OLS Estimator and Corollary 2

The following proposition is referenced in Section 4.2 as the basis for implementing the estimator in Corollary 2.

Proposition A.3 (Interacted OLS Estimator)**.** The treatment effect estimate implied by the solution to the minimization problem in Corollary 2 is identical to the estimate obtained by augmenting the OLS regression of interest by including interactions between the treatment variable, D_i, and indicators for the mutually exclusive groups of interest, and then estimating a linear combination of these treatment-group interaction terms evaluated at the desired shares $(a_1, \ldots, a_G)$.

Proof. This proof follows the structure of the proof of Proposition 6. Let $\mathbf{C} = (a_1, \ldots, a_G, 0, \ldots, 0)'$ be a $(G + k - 1) \times 1$ vector, where the first G elements are each a_g, the weights placed on the desired contribution of each group, and the last $k - 1$ elements are 0. Let $\mathbf{G}_i$ be a $G \times 1$ vector of indicators for being in each group. Let $\mathbf{W}_i = (D_i \mathbf{G}_i', \mathbf{X}_i')'$ be the $(G+k-1) \times 1$ vector containing interactions between the treatment variable of interest and indicators for being in each group and the vector of controls. $\mathbf{W}$ denotes the $n \times (G+k-1)$ matrix that is stacked across observations. A regression of Y_i on $\mathbf{W}_i$ produces coefficients:

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}. \tag{A.26}$$

We are interested in the treatment effect evaluated at a specific proportion of each group given by the values in $(a_1, \ldots, a_G)$. This is given by $\hat{\theta} = \mathbf{C}'\hat{\boldsymbol{\beta}}$. Given this, the interacted OLS OW for $\hat{\theta}$ can be written:

$$\omega_i = \frac{\partial \hat{\theta}}{\partial Y_i} = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i. \tag{A.27}$$

<u>The constrained minimization problem.</u> The Lagrangian for the constrained optimization problem in Corollary 2 can be written:

$$\mathcal{L} = \frac{1}{2}\sum_i \omega_i^2 - \lambda_D \left(\sum_i \omega_i D_i - 1\right) - \boldsymbol{\lambda}_X' \left(\sum_i \omega_i \mathbf{X}_i\right) - \boldsymbol{\lambda}_G' \left(\sum_i \omega_i D_i \mathbf{G}_i - (a_1, \ldots, a_G)'\right), \tag{A.28}$$

where $\boldsymbol{\lambda}_X$ has dimension $(k-1) \times 1$ and $\boldsymbol{\lambda}_G$ has dimension $G \times 1$. The first-order condition produces an expression for the OW that includes the Lagrange multipliers:

$$\frac{\partial \mathcal{L}}{\partial \omega_i} = \omega_i - \lambda_D D_i - \boldsymbol{\lambda}_X' \mathbf{X}_i - \boldsymbol{\lambda}_G' D_i \mathbf{G}_i = 0 \Rightarrow \omega_i = \lambda_D D_i + \boldsymbol{\lambda}_X' \mathbf{X}_i + \boldsymbol{\lambda}_G' D_i \mathbf{G}_i. \tag{A.29}$$

Because the groups are mutually exclusive and collectively exhaustive, we have $D_i = \boldsymbol{\iota}_G' D_i \mathbf{G}_i$, where $\boldsymbol{\iota}_G$ is a $G \times 1$ vector of ones. The first-order condition can therefore be written using only the $G+k-1$ regressors in $\mathbf{W}_i$:

$$\omega_i = (\boldsymbol{\lambda}_G + \lambda_D \boldsymbol{\iota}_G)' D_i \mathbf{G}_i + \boldsymbol{\lambda}_X' \mathbf{X}_i. \tag{A.30}$$

This step reflects the fact that the constraint $\sum_i \omega_i D_i = 1$ is implied by the group constraints $\sum_{i:g_i=g} \omega_i D_i = a_g$ together with $\sum_{g=1}^G a_g = 1$, so the Lagrange multipliers λ_D and $\boldsymbol{\lambda}_G$ are only identified up to the combination $\boldsymbol{\lambda}_G + \lambda_D \boldsymbol{\iota}_G$.

Let $\boldsymbol{\Lambda} = ((\boldsymbol{\lambda}_G + \lambda_D \boldsymbol{\iota}_G)', \boldsymbol{\lambda}_X')'$, which is $(G+k-1) \times 1$. We can write the $n \times 1$ vector of OWs as:

$$\boldsymbol{\omega} = \mathbf{W}\boldsymbol{\Lambda}. \tag{A.31}$$

As in the proof of Proposition 6, the first-order condition is necessary and sufficient for a unique global minimum as long as $\mathbf{W}$ has full rank, which requires that each group contains variation in the treatment variable after partialling out $\mathbf{X}_i$. Writing the constraints in matrix notation, we also have:

$$\boldsymbol{\omega}'\mathbf{W} = \mathbf{C}'. \tag{A.32}$$

Plugging equation (A.31) into equation (A.32), we have:

$$\boldsymbol{\Lambda}'\mathbf{W}'\mathbf{W} = \mathbf{C}' \Rightarrow \boldsymbol{\Lambda}' = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1} \Rightarrow \boldsymbol{\Lambda} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{C}. \tag{A.33}$$

Plugging this back into equation (A.31) yields:

$$\boldsymbol{\omega} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C}, \tag{A.34}$$

which gives us:

$$\omega_i = \mathbf{W}_i'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{C} = \mathbf{C}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i. \tag{A.35}$$

This is identical to the expression for the OLS OW in equation (A.27).

□

B Constructing Outcome Weights for Other Estimators

This appendix derives OWs for estimators besides OLS and 2SLS. For each estimator, we define the estimator and then derive the OW.

B.1 Weighted Least Squares

Consider estimation of equation (7) by WLS, where observation i receives a known, strictly positive weight ϕ_i. Let $\boldsymbol{\Phi}$ denote the $n \times n$ diagonal matrix with ϕ_i as its ith diagonal element. The WLS estimator solves:

$$\min_{\theta,\boldsymbol{\beta}} \sum_i \phi_i \left(Y_i - D_i\theta - \mathbf{X}_i'\boldsymbol{\beta}\right)^2. \tag{B.1}$$

Proposition B.1 (WLS OW)**.** The OW for observation i in a WLS estimate of $\hat{\theta}$ from equation (B.1) is

$$\omega_i = \frac{\phi_i \mathring{D}_i}{\sum_j \phi_j \mathring{D}_j^2},$$

where $\mathring{D}_i = D_i - \mathbf{X}_i'(\mathbf{X}'\boldsymbol{\Phi}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}\mathbf{D}$ is the residual from a WLS regression of D_i on $\mathbf{X}_i$ using the weights ϕ_i.

Proof. Define the transformed variables $\dot{Z}_i \equiv \phi_i^{1/2} Z_i$ for each variable Z, with $\dot{\mathbf{Y}}$, $\dot{\mathbf{D}}$, and $\dot{\mathbf{X}}$ denoting the stacked transformed outcome, treatment, and controls. The WLS objective function in equation (B.1) equals the OLS objective function for the transformed data, $\sum_i(\dot{Y}_i - \dot{D}_i\theta - \dot{\mathbf{X}}_i'\boldsymbol{\beta})^2$, so WLS estimation of equation (B.1) is numerically identical to OLS estimation of:

$$\dot{Y}_i = \dot{D}_i\theta + \dot{\mathbf{X}}_i'\boldsymbol{\beta} + \dot{\epsilon}_i. \tag{B.2}$$

Following the same steps as in the proof of Proposition 2, the Frisch-Waugh-Lovell theorem implies that the OLS estimator of θ in equation (B.2) is:

$$\hat{\theta} = \frac{\sum_i \tilde{\dot{D}}_i \dot{Y}_i}{\sum_j \tilde{\dot{D}}_j^2}, \tag{B.3}$$

where $\tilde{\dot{D}}_i \equiv \dot{D}_i - \dot{\mathbf{X}}_i'(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\dot{\mathbf{D}}$ is the residual from the regression of $\dot{D}_i$ on $\dot{\mathbf{X}}_i$.

The transformation implies that $\dot{\mathbf{X}}'\dot{\mathbf{X}} = \mathbf{X}'\boldsymbol{\Phi}\mathbf{X}$, $\dot{\mathbf{X}}'\dot{\mathbf{D}} = \mathbf{X}'\boldsymbol{\Phi}\mathbf{D}$, and $\dot{\mathbf{X}}_i = \phi_i^{1/2}\mathbf{X}_i$. Substituting these expressions into equation (B.3) shows that the residualized transformed treatment variable is proportional to the weighted residual:

$$\tilde{\dot{D}}_i = \phi_i^{1/2} D_i - \phi_i^{1/2}\mathbf{X}_i'(\mathbf{X}'\boldsymbol{\Phi}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}\mathbf{D} = \phi_i^{1/2}\mathring{D}_i. \tag{B.4}$$

Substituting this expression and $\dot{Y}_i = \phi_i^{1/2} Y_i$ into equation (B.3) yields:

$$\hat{\theta} = \frac{\sum_i \phi_i^{1/2}\mathring{D}_i \phi_i^{1/2} Y_i}{\sum_j \left(\phi_j^{1/2}\mathring{D}_j\right)^2} = \frac{\sum_i \phi_i \mathring{D}_i Y_i}{\sum_j \phi_j (\mathring{D}_j)^2}. \tag{B.5}$$

From equation (5), the OW is $\omega_i = \partial\hat{\theta}/\partial Y_i$, which gives the result.

□

The WLS OW inherits the properties of the OLS OW. The stacked vector of weighted residuals $\mathring{\mathbf{D}}$ satisfies $\mathbf{X}'\boldsymbol{\Phi}\mathring{\mathbf{D}} = \mathbf{0}$, which implies $\sum_i \omega_i x_i = 0$ for every $x \in \mathbf{X}$. Because $\mathbf{X}_i$ includes a constant, setting $x_i = 1$ delivers the zero-sum property of Proposition 1(A), and the remaining elements deliver an analog of the covariate balance property of Proposition A.1. Writing $D_i = \mathring{D}_i + \mathbf{X}_i'(\mathbf{X}'\boldsymbol{\Phi}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}\mathbf{D}$ and using the same orthogonality condition also gives $\sum_i \omega_i D_i = 1$, so the Wald representation in Proposition 4 applies to WLS as well.

B.2 Matching

Let D_i be a binary treatment variable. Let $w(i, i')$ be a matching weight, where i indexes a treated unit and i' indexes an untreated unit. Each treated unit is matched to some weighted average of untreated units, and the sum of weights across untreated units is 1: $\sum_{i':D_{i'}=0} w(i, i') = 1$. The specific value of $w(i, i')$ depends on the match criteria. For example, in nearest-neighbor matching with one neighbor, $w(i, i') = 1$ if treated unit i is matched to untreated unit i' and $w(i, i') = 0$ otherwise. The number of treated units is n_1.

The representation of a matching estimator as a weighted average of observed outcomes is standard (e.g., Abadie and Imbens, 2006). The resulting matching estimator can be written:

$$\hat{\theta} = \frac{1}{n_1} \sum_{i:D_i=1} \left(Y_i - \sum_{i':D_{i'}=0} w(i,i') Y_{i'} \right) = \frac{1}{n_1} \sum_{i:D_i=1} Y_i - \sum_{i':D_{i'}=0} \frac{1}{n_1} \sum_{i:D_i=1} w(i,i') Y_{i'}. \tag{B.6}$$

Proposition B.2 (Matching OW)**.** The OW for observation i in a matching estimate of $\hat{\theta}$ from equation (B.6) is

$$\omega_i = \begin{cases} \dfrac{1}{n_1} & \text{if } D_i = 1 \\ -\dfrac{1}{n_1} \displaystyle\sum_{j:D_j=1} w(j,i) & \text{if } D_i = 0 \end{cases}$$

Proof. This follows immediately from Definition 1 and equation (B.6). □

B.3 Inverse Propensity Weighting

Let D_i be a binary treatment variable and $\hat{P}_i$ be the estimated propensity score. The un-normalized IPW estimator (also referred to as the Horvitz-Thompson estimator) of the average treatment effect (ATE) is:

$$\hat{\theta} = \frac{1}{n} \sum_i \left(\frac{D_i}{\hat{P}_i} - \frac{1-D_i}{1-\hat{P}_i} \right) Y_i. \tag{B.7}$$

The un-normalized estimator in equation (B.7) is not guaranteed to satisfy the mean-invariance property described in Proposition 1. As a result, we do not consider this estimator. Instead, we consider the normalized IPW estimator (also referred to as the Hajek estimator) of the ATE:

$$\hat{\theta} = \sum_i \left(\frac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} - \frac{\frac{1-D_i}{1-\hat{P}_i}}{\sum_j \frac{1-D_j}{1-\hat{P}_j}} \right) Y_i. \tag{B.8}$$

The Hajek estimator satisfies the mean-invariance property, which implies that Proposition 1 holds for it.

The normalized IPW estimator of the average treatment effect on the treated (ATT) is:

$$\hat{\theta} = \sum_i \left(\frac{D_i}{\sum_j D_j} - \frac{\frac{(1-D_i)\hat{P}_i}{1-\hat{P}_i}}{\sum_j \frac{(1-D_j)\hat{P}_j}{1-\hat{P}_j}} \right) Y_i. \tag{B.9}$$

Proposition B.3 (Normalized IPW OW)**.** The OW for observation i in a normalized-IPW estimate of ATE $\hat{\theta}$ from equation (B.8) is

$$\omega_i = \begin{cases} \dfrac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} & \text{if } D_i = 1 \\ -\dfrac{\frac{1-D_i}{1-\hat{P}_i}}{\sum_j \frac{1-D_j}{1-\hat{P}_j}} & \text{if } D_i = 0. \end{cases}$$

The OW for observation i in a normalized-IPW estimate of ATT $\hat{\theta}$ from equation (B.9) is

$$\omega_i = \begin{cases} \dfrac{D_i}{\sum_j D_j} & \text{if } D_i = 1 \\ -\dfrac{\frac{(1-D_i)\hat{P}_i}{1-\hat{P}_i}}{\sum_j \frac{(1-D_j)\hat{P}_j}{1-\hat{P}_j}} & \text{if } D_i = 0 \end{cases}$$

Proof. This follows immediately from Definition 1 and equations (B.8) and (B.9). □

B.4 Regression-Adjustment Estimator

Let D_i be a binary treatment variable and $\mathbf{X}_i$ be a vector of exogenous variables, including a constant. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ be OLS estimates from regressing Y_i on $\mathbf{X}_i$ for observations with $D_i = 1$ and $D_i = 0$. Letting $\mathbf{Y}_d$ denote the vector of stacked outcomes for observations with $D_i = d$ and $\mathbf{X}_d$ denote the corresponding matrix of exogenous variables, we have $\hat{\boldsymbol{\beta}}_d = (\mathbf{X}_d'\mathbf{X}_d)^{-1}\mathbf{X}_d'\mathbf{Y}_d$.

The regression-adjustment estimator of the ATE can be written:

$$\hat{\theta} = \frac{1}{n}\sum_i \mathbf{X}_i'(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) = \frac{1}{n}\sum_i \mathbf{X}_i'((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1 - (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y}_0). \tag{B.10}$$

The regression-adjustment estimator of the ATT can be written:

$$\hat{\theta} = \frac{1}{n_1}\sum_{i:D_i=1} \mathbf{X}_i'(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) = \frac{1}{n_1}\sum_{i:D_i=1} \mathbf{X}_i'((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1 - (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y}_0). \tag{B.11}$$

Proposition B.4 (Regression-Adjustment OW)**.** The OW for observation i in a regression-adjustment estimate of ATE $\hat{\theta}$ from equation (B.10) is

$$\omega_i = \begin{cases} \dfrac{1}{n}\sum_j \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_i & \text{if } D_i = 1 \\ -\dfrac{1}{n}\sum_j \mathbf{X}_j'(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_i & \text{if } D_i = 0. \end{cases}$$

The OW for observation i in a regression-adjustment estimate of ATT $\hat{\theta}$ from equation (B.11) is

$$\omega_i = \begin{cases} \dfrac{1}{n_1} \displaystyle\sum_{j:D_j=1} \mathbf{X}_j'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_i & \text{if } D_i = 1 \\ -\dfrac{1}{n_1} \displaystyle\sum_{j:D_j=1} \mathbf{X}_j'(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_i & \text{if } D_i = 0. \end{cases}$$

Proof. This follows immediately from Definition 1 and equations (B.10) and (B.11). □

B.5 Doubly-Robust Estimator

Let D_i be a binary treatment variable, $\hat{P}_i$ be the estimated propensity score, and $\mathbf{X}_i$ be a vector of control variables including a constant. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ be OLS estimates from regressing Y_i on $\mathbf{X}_i$ for observations with $D_i = 1$ and $D_i = 0$.

The doubly-robust normalized estimator of the ATE can be written:

$$\hat{\theta} = \frac{1}{n}\sum_i \mathbf{X}_i'(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) + \sum_i \frac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}}(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}_1) - \sum_i \frac{\frac{1-D_i}{1-\hat{P}_i}}{\sum_j \frac{1-D_j}{1-\hat{P}_j}}(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}_0). \tag{B.12}$$

Proposition B.5 (Doubly-Robust OW)**.** The OW for observation i in a doubly-robust estimate of ATE $\hat{\theta}$ from equation (B.12) is

$$\omega_i = \begin{cases} \dfrac{\frac{D_i}{\hat{P}_i}}{\sum_j \frac{D_j}{\hat{P}_j}} + \left(\dfrac{1}{n}\displaystyle\sum_j \mathbf{X}_j' - \dfrac{\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j'}{\sum_j \frac{D_j}{\hat{P}_j}}\right)(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_i & \text{if } D_i = 1 \\ -\dfrac{\frac{1-D_i}{1-\hat{P}_i}}{\sum_j \frac{1-D_j}{1-\hat{P}_j}} - \left(\dfrac{1}{n}\displaystyle\sum_j \mathbf{X}_j' - \dfrac{\sum_j \frac{1-D_j}{1-\hat{P}_j}\mathbf{X}_j'}{\sum_j \frac{1-D_j}{1-\hat{P}_j}}\right)(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_i & \text{if } D_i = 0. \end{cases}$$

Proof. This follows immediately from Definition 1 and equation (B.12). □

The doubly-robust OW differs from the normalized IPW OW in Proposition B.3 only through the second term in each branch. Each of these terms multiplies the difference between the unweighted sample mean of the control variables, $\frac{1}{n}\sum_j \mathbf{X}_j$, and their IPW-reweighted mean, and so measures the extent to which IPW reweighting fails to reproduce the sample mean of the control variables. If the estimated propensity score balances the control variables exactly, so that

$$\frac{\sum_j \frac{D_j}{\hat{P}_j}\mathbf{X}_j}{\sum_j \frac{D_j}{\hat{P}_j}} = \frac{1}{n}\sum_j \mathbf{X}_j = \frac{\sum_j \frac{1-D_j}{1-\hat{P}_j}\mathbf{X}_j}{\sum_j \frac{1-D_j}{1-\hat{P}_j}},$$

then both terms equal zero and the doubly-robust OW coincides with the normalized IPW OW.

B.6 Kernel Regression

Let D_i be a continuous treatment variable. We take the parameter of interest as the difference in expected outcomes from moving from D_U to D_T: $\theta = E[Y|D = D_T] - E[Y|D = D_U]$. For kernel $K_h(\cdot)$, the kernel regression estimator of the ATE can be written:

$$\hat{\theta} = \frac{\sum_i K_h(D_i - D_T)Y_i}{\sum_j K_h(D_j - D_T)} - \frac{\sum_i K_h(D_i - D_U)Y_i}{\sum_j K_h(D_j - D_U)}. \tag{B.13}$$

Proposition B.6 (Kernel Regression OW)**.** The OW for observation i in a kernel regression estimate of ATE $\hat{\theta}$ from equation (B.13) is

$$\omega_i = \frac{K_h(D_i - D_T)}{\sum_j K_h(D_j - D_T)} - \frac{K_h(D_i - D_U)}{\sum_j K_h(D_j - D_U)}.$$

Proof. This follows immediately from Definition 1 and equation (B.13). □

B.7 Marginal Treatment Effects

Let D_i be a binary treatment variable, $\mathbf{X}_i$ be a $(k-1) \times 1$ vector of exogenous control variables including a constant, $\mathbf{Z}_i$ be a vector of instruments, and $\hat{P}_i$ be the estimated propensity score estimated (typically with a logit or probit model) from a first stage of D_i on $\mathbf{X}_i$ and $\mathbf{Z}_i$. If the MTE is estimated through a polynomial regression, then the second stage regression is:

$$Y_i = \mathbf{X}_i'\boldsymbol{\alpha} + \sum_{m=1}^{M} \beta_m \hat{P}_i^m + \epsilon_i. \tag{B.14}$$

The estimated MTE at p, denoted $\hat{\theta}(p)$, is the derivative of the second stage regression with respect to the propensity score:

$$\hat{\theta}(p) = \sum_{m=1}^{M} \hat{\beta}_m m p^{m-1}. \tag{B.15}$$

Because the first stage does not use the outcome variable, the estimated propensity score $\hat{P}_i$ depends only on $(\mathbf{D}, \mathbf{X}, \mathbf{Z})$. Each $\hat{\beta}_m$ is therefore an OLS coefficient from a regression whose right-hand-side variables do not depend on $\mathbf{Y}$, so each $\hat{\beta}_m$ is a linear combination of outcomes, and so is any parameter constructed from equation (B.15). The exact formula for the OW depends on the choice of functional form in the second stage regression and on the target parameter that the researcher constructs from the MTE. Rather than write out one particular case, we describe the steps through which the OW can be calculated.

1. Let $\mathbf{V}_i = (\mathbf{X}_i', \hat{P}_i, \hat{P}_i^2, \ldots, \hat{P}_i^M)'$ be the $(k-1+M) \times 1$ vector of second stage regressors from equation (B.14), with $\mathbf{V}$ denoting the $n \times (k-1+M)$ matrix that is stacked across observations. The second stage coefficients are:

$$\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\alpha}}', \hat{\beta}_1, \ldots, \hat{\beta}_M)' = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{Y}. \tag{B.16}$$

2. Let $\mathbf{C}_m$ be the $(k-1+M) \times 1$ vector with a 1 in the position corresponding to β_m and zeros elsewhere, so that $\hat{\beta}_m = \mathbf{C}_m'(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{Y}$. Following the same steps used to derive equation (A.6), the OW for $\hat{\beta}_m$ is:

$$\omega_{i,m} = \frac{\partial \hat{\beta}_m}{\partial Y_i} = \mathbf{C}_m'(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}_i. \tag{B.17}$$

3. Repeating step 2 for $m = 1, \ldots, M$ and combining the resulting OWs using equation (B.15) gives the OW for the MTE at p:

$$\omega_i(p) = \frac{\partial \hat{\theta}(p)}{\partial Y_i} = \sum_{m=1}^{M} mp^{m-1}\omega_{i,m} = \mathbf{C}(p)'(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}_i, \tag{B.18}$$

where $\mathbf{C}(p) \equiv \sum_{m=1}^{M} mp^{m-1}\mathbf{C}_m = (\mathbf{0}', 1, 2p, \ldots, Mp^{M-1})'$ and $\mathbf{0}$ is a $(k-1) \times 1$ vector of zeros.

4. Most parameters of interest are weighted integrals of the MTE, $\hat{\theta}_\mu = \int_0^1 \mu(p)\hat{\theta}(p)dp$, where p indexes the level of unobserved resistance at which the MTE is evaluated, following the local instrumental variables representation $\hat{\theta}(p) = \partial E[Y_i \mid \mathbf{X}_i, \hat{P}_i = p]/\partial p$, and the weight function $\mu(\cdot)$ is known and depends on the target parameter and the distribution of $\hat{P}_i$ (Heckman and Vytlacil, 2001, 2005; Mogstad, Santos and Torgovitsky, 2018). For example, $\mu(p) = 1$ for the ATE; $\mu(p) = n^{-1}\sum_i \mathbf{1}(\hat{P}_i > p)/(n^{-1}\sum_i \hat{P}_i)$ for the ATT; and $\mu(p) = \mathbf{1}(p_0 < p < p_1)/(p_1 - p_0)$ for the LATE between propensity scores p_0 and p_1. Because equation (B.18) is linear in $\mathbf{C}(p)$, the OW for the target parameter is:

$$\omega_i^\mu = \int_0^1 \mu(p)\omega_i(p)dp = \bar{\mathbf{C}}_\mu'(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}_i, \quad \text{where } \bar{\mathbf{C}}_\mu \equiv \int_0^1 \mu(p)\mathbf{C}(p)dp. \tag{B.19}$$

The first $k-1$ elements of $\bar{\mathbf{C}}_\mu$ are zero and its mth polynomial element is $\int_0^1 \mu(p)mp^{m-1}dp$, which can be computed analytically for common weight functions and numerically otherwise. For the ATE, $\int_0^1 mp^{m-1}dp = 1$ for every m, so $\hat{\theta}_\mu = \sum_{m=1}^{M}\hat{\beta}_m$ and $\bar{\mathbf{C}}_\mu = (\mathbf{0}', 1, \ldots, 1)'$.

Two features of this construction are worth noting. First, the resulting OW satisfies the zero-sum property of Proposition 1(A). Because $\mathbf{X}_i$ includes a constant, the $n \times 1$ vector of ones $\boldsymbol{\iota}$ is a column of $\mathbf{V}$, so $\mathbf{V}'\boldsymbol{\iota} = \mathbf{V}'\mathbf{V}\mathbf{e}_c$, where $\mathbf{e}_c$ selects that column. Equation (B.19) then gives $\sum_i \omega_i^\mu = \bar{\mathbf{C}}_\mu'(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\boldsymbol{\iota} = \bar{\mathbf{C}}_\mu'\mathbf{e}_c = 0$, because the elements of $\bar{\mathbf{C}}_\mu$ corresponding to $\mathbf{X}_i$ are zero. The same argument applies to $\omega_i(p)$ at each p.

Second, only steps 1 and 2 depend on the parametric form of the second stage. Any second stage estimator of $E[Y_i|\mathbf{X}_i, \hat{P}_i = p]$ that is linear in $\mathbf{Y}$—including series, spline, and local polynomial estimators—produces $\hat{\theta}(p) = \sum_i \omega_i(p)Y_i$; only the design matrix $\mathbf{V}$ and the vector $\mathbf{C}(p)$ change. Similarly, if the second stage includes interactions between $\mathbf{X}_i$ and the polynomial terms, so that the MTE varies with covariates, then $\mathbf{C}(p)$ is replaced by a vector $\mathbf{C}(p, \mathbf{x})$ that also depends on the evaluation point $\mathbf{x}$, and step 4 additionally averages over the distribution of $\mathbf{X}_i$.

C Details on Relationship with Existing Measures

This appendix provides more details on the relationship between the outcome weight (OW) and treatment effect weight (TW) studied in this paper and several existing measures used for regression diagnostics and influence analysis. We discuss leverage, Cook's distance, the change in a coefficient when leaving out one observation, and the multiple regression weight of Aronow and Samii (2016, hereafter, AS). Throughout, we consider an OLS regression of outcome Y_i on treatment D_i and controls (including a constant) $\mathbf{X}_i$. The combined vector of treatment and controls is $\mathbf{W}_i = (D_i, \mathbf{X}_i')'$, with $\mathbf{W}$ denoting the matrix that stacks $\mathbf{W}_i$ across all observations.

C.1 Leverage

Definition. Following Belsley, Kuh and Welsch (1980), the leverage of observation i is:

$$h_i = \mathbf{W}_i'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i.$$

Leverage describes how the fitted value $\hat{Y}_i$ depends on Y_i: $h_i = \partial\hat{Y}_i/\partial Y_i$. Intuitively, this captures how far observation i's covariates lie from the center of the predictor space. Leverage satisfies $1/n \leq h_i \leq 1$ and $\sum_i h_i = k$, where $k = \dim(\mathbf{W}_i)$ is the number of regressors.

Comparison to OW. Both leverage and the OW are functions of the design matrix, $\mathbf{W}$, alone—neither uses the observed outcomes Y_i. However, they differ in scope and sign. Leverage summarizes how unusual observation i is in terms of the full design matrix $\mathbf{W}_i$, while the OW is specific to the treatment variable: it measures residual variation in D_i after partialling out $\mathbf{X}_i$. This can also be seen by noting that leverage is $\partial\hat{Y}_i/\partial Y_i$, while the OW is $\partial\hat{\theta}/\partial Y_i$. Leverage is always positive, while the OW can be positive or negative. Consequently, leverage does not provide a means of understanding the effective treatment and control groups.

C.2 Cook's Distance

Definition. Cook (1977) measures the aggregate shift in fitted values when observation i is deleted:

$$\text{Cook}_i = \frac{\sum_j(\hat{Y}_j - \hat{Y}_{j(-i)})^2}{k \cdot \widehat{\sigma}^2},$$

where $\hat{Y}_j$ is the fitted value for observation j from the model estimated using all observations, $\hat{Y}_{j(-i)}$ is the fitted value from the model estimated without observation i, k is the number of regressors, and $\widehat{\sigma}^2$ is the mean squared error. An equivalent representation is:

$$\text{Cook}_i = \frac{e_i^2}{k \cdot \widehat{\sigma}^2} \cdot \frac{h_i}{(1-h_i)^2},$$

where $e_i = Y_i - \hat{Y}_i$ is the residual. Cook's distance is thus a product of a residual term (how poorly the model fits observation i) and a leverage term (how extreme observation i's covariates are).

Comparison to OW. Cook's distance captures how much observation i shifts the entire vector of fitted values $\hat{\mathbf{Y}}$, while the OW captures how much observation i's outcome Y_i contributes to a single coefficient $\hat{\theta}$. Cook's distance is similar to leverage in providing a nonnegative scalar that does not identify effective treatment or control groups. Cook's distance also differs from the OW because it depends on the outcome Y_i (via e_i^2). In contrast, the OW depends only on the design matrix.

C.3 Leave-One-Out Change in Coefficient

Definition. The change in the estimate of $\hat{\theta}$ when removing observation i from estimation is $\hat{\theta} - \hat{\theta}_{(-i)}$, where $\hat{\theta}_{(-i)}$ denotes the estimate obtained without observation i. This can be written as:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = \frac{(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i\, e_i}{1 - h_i},$$

where $e_i = Y_i - \hat{Y}_i$ is the residual for observation i and h_i is its leverage (Cook, 1977).

Because $\hat{\theta}$ is the first element of $\hat{\boldsymbol{\beta}}$, premultiplying by $\mathbf{e}_1' = (1, 0, \ldots, 0)'$ gives

$$\hat{\theta} - \hat{\theta}_{(-i)} = \frac{\mathbf{e}_1'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}_i\, e_i}{1 - h_i} = \frac{\omega_i\, e_i}{1 - h_i},$$

where the second equality follows from equation (A.6).

Comparison to OW. The above formula decomposes the leave-one-out change in estimate into three factors: the OW ω_i, the residual e_i, and a term $1/(1 - h_i)$ that inflates high-leverage observations. Leaving out an observation leads to a large change in the estimate only when $|\omega_i|$ and $|e_i|$ are simultaneously large, with the effect further amplified when h_i is close to one. Chattopadhyay and Zubizarreta (2023) identify influential observations using the sample influence curve of observation i, which equals $\hat{\theta} - \hat{\theta}_{(-i)}$ multiplied by a positive constant that depends only on sample sizes.

There are several differences between the OW and the leave-one-out change in coefficient, $\hat{\theta} - \hat{\theta}_{(-i)}$. First, the OW permits identification of the effective treatment and control groups. In contrast, the sign of the leave-one-out change in coefficient also depends on the sign of the residual e_i. Relatedly, the OW depends only on the design matrix, while the leave-one-out change also depends on the outcome Y_i through the residual.

C.4 Aronow and Samii (2016)

Definition. AS define the "multiple regression weight" as:

$$w_i^{AS} = (D_i - E[D_i|\mathbf{X}_i])^2. \tag{C.1}$$

They use this weight to define the "mean of a covariate Z_i in the effective sample" (p. 255) as:

$$\mu^{AS}(Z_i) = \frac{E[w_i^{AS} Z_i]}{E[w_i^{AS}]}. \tag{C.2}$$

Comparison to OW. The AS multiple regression weight is always positive, unlike the OW and TW. As a result, the AS multiple regression weight cannot be used to identify observations in the effective treatment and control groups or conduct the range of analyses built on this distinction.

However, the AS weight is closely related to a variant of the TW in which the treatment effect of observation i is defined relative to the linear projection $L(D_i \mid \mathbf{X}_i)$ rather than the untreated state, so that the treatment effect weight is $TW_i^{LP} = \omega_i(D_i - L(D_i \mid \mathbf{X}_i))$ rather than $TW_i = \omega_i D_i$ from Definition 3, for which the equivalence does not hold. Under the assumption that the expected value of D_i given $\mathbf{X}_i$ is equal to the linear projection—$E(D \mid \mathbf{X}) = L(D \mid \mathbf{X})$—this variant of the TW and the AS weight yield identical effective sample means for any covariate Z_i.

The mean of Z_i weighted by this variant of the TW is

$$\mu^{TW-LP}(Z_i) = \frac{E[\omega_i\,(D_i - L(D_i \mid \mathbf{X}_i))\,Z_i]}{E[\omega_i\,(D_i - L(D_i \mid \mathbf{X}_i))]}.$$

Substituting the OLS OW from Proposition 2 into this expression yields:

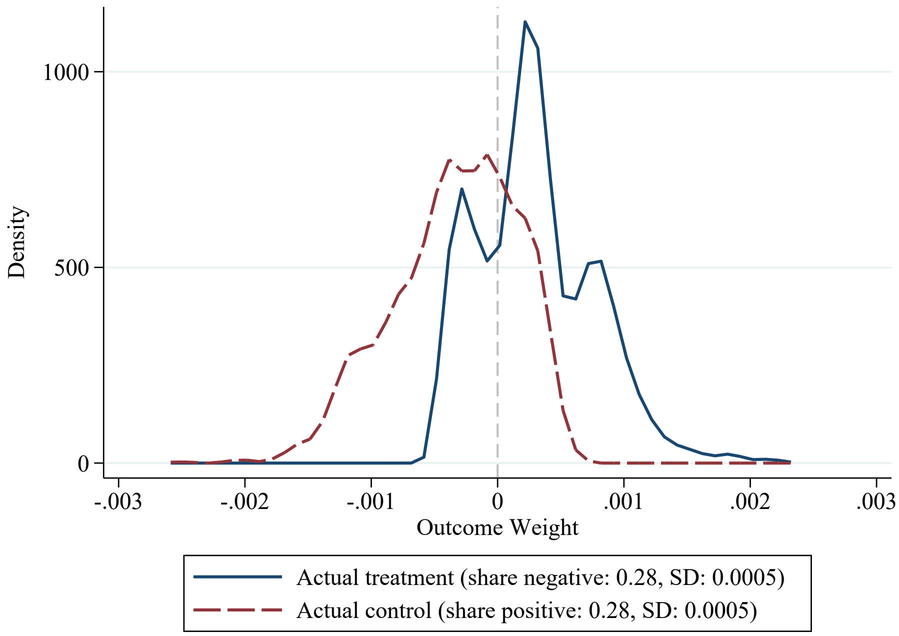
$$\mu^{TW-LP}(Z_i) = \frac{E\big[(D_i - L(D_i \mid \mathbf{X}_i))^2\,Z_i\big]}{E\big[(D_i - L(D_i \mid \mathbf{X}_i))^2\big]} = \frac{E\big[(D_i - E[D_i \mid \mathbf{X}_i])^2\,Z_i\big]}{E\big[(D_i - E[D_i \mid \mathbf{X}_i])^2\big]} = \mu^{AS}(Z_i),$$

where the first equality plugs in the OLS OW, the second equality uses the assumption made here that $E(D \mid \mathbf{X}) = L(D \mid \mathbf{X})$, and the third equality follows immediately from combining the AS expressions (C.1) and (C.2).

In the case where $E(D \mid \mathbf{X}) \neq L(D \mid \mathbf{X})$, the equality between $\mu^{TW-LP}(Z_i)$ and $\mu^{AS}(Z_i)$ breaks down. Our approach does not rely on the linearity assumption made by AS. As a result, our approach can be used in settings where linearity does not hold, like two-way fixed effects regressions (e.g., Acemoglu et al., 2019).

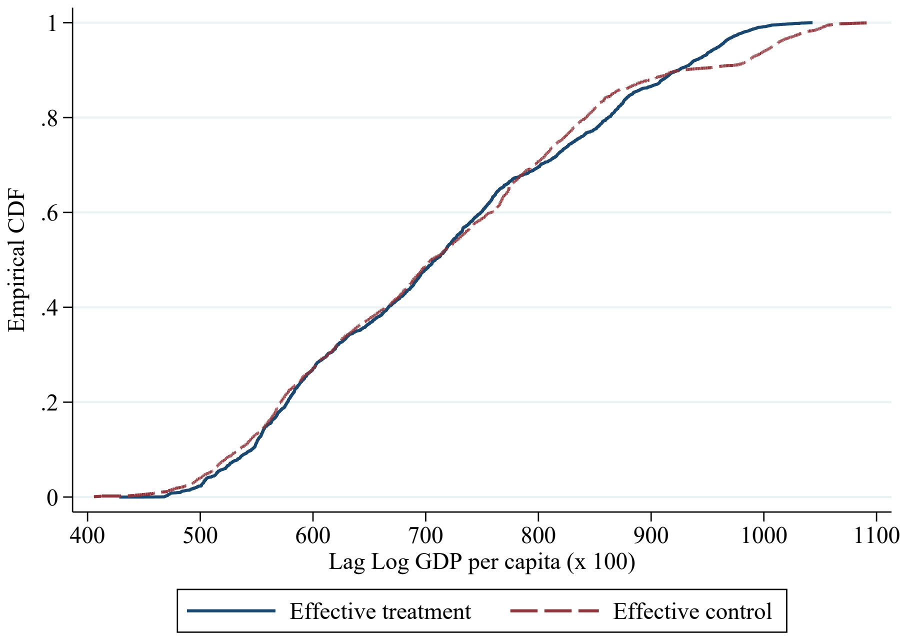
D Additional Tables and Figures

Figure D.1: Distribution of Outcome Weights by Democracy Status, Acemoglu et al. (2019)



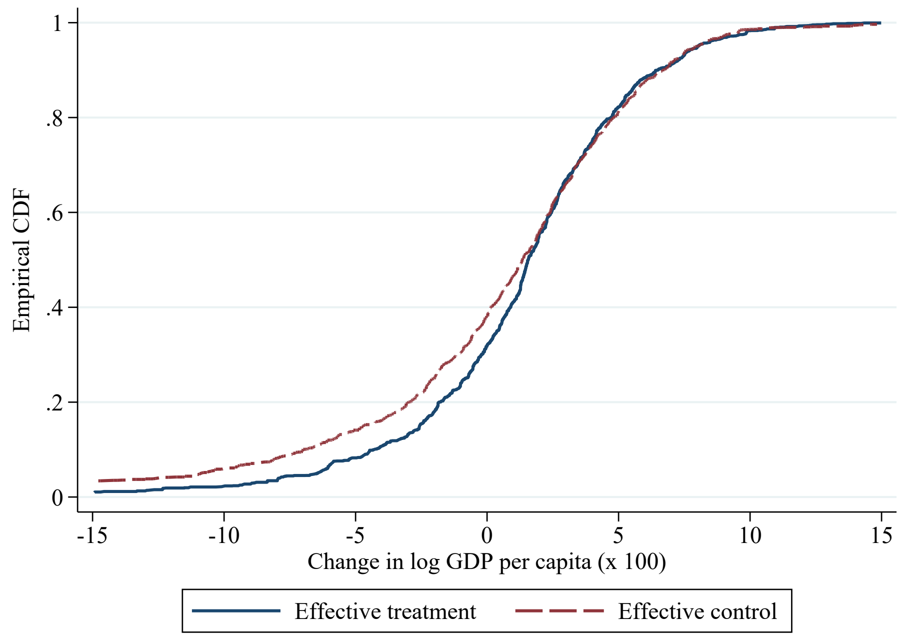
Notes: Figure displays kernel densities of outcome weights for observations in the actual treatment group (democracies) and actual control group (nondemocracies). The legend reports summary statistics on OWs.
Source: Authors' analysis of data from Acemoglu et al. (2019).

Figure D.2: Distribution of One-Year Lag of Log GDP per Capita, Effective Treatment and Control Groups, Acemoglu et al. (2019)



Notes: Figure displays empirical CDFs of the one-year lag of log GDP per capita (times 100) for the effective treatment and control groups weighted by the OW. We use the absolute value of the OW as a weight for the effective control group.
Source: Authors' analysis of data from Acemoglu et al. (2019).

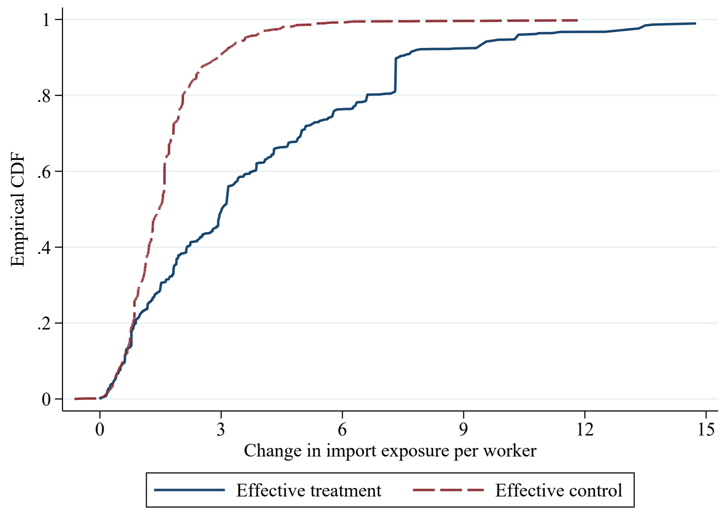
Figure D.3: Distribution of Outcomes, Nonconvex Weights, Acemoglu et al. (2019)



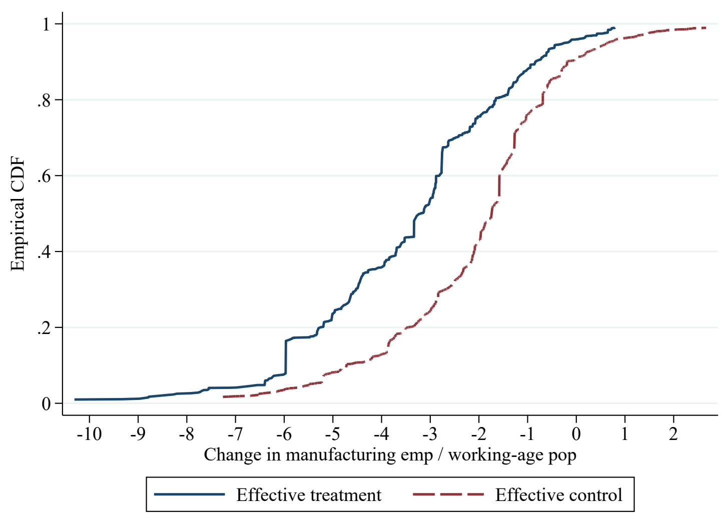
Notes: Figure displays empirical CDFs of the one-year change in log GDP per capita (times 100) for the effective treatment and control groups weighted by the OW. Unlike in Panel B of Figure 1, this figure uses OWs generated by the constrained estimator that restricts effective treatment to be only treated observations and effective control to be only control observations. We only show results between changes of -15 and 15 log points to make the difference between the CDFs in the middle of the distribution more visible.
Source: Authors' analysis of data from Acemoglu et al. (2019).

Figure D.4: OW-Weighted Empirical CDFs of Treatment and Outcome, Effective Treatment and Control, Autor, Dorn and Hanson (2013)

(a) OW-Weighted Empirical CDFs of Treatment



(b) OW-Weighted Empirical CDFs of Outcome



Notes: Panel A displays empirical CDFs of the change in import exposure per worker for the effective treatment and control groups weighted by the OW. We only show observations up to a value of 15. Panel B displays empirical CDFs of the change in the manufacturing employment share of the working-age population for the effective treatment and control groups weighted by the OW. We only show results between percentiles 1 and 99 to make the difference between the CDFs in the middle of the distribution more visible.
Source: Authors' analysis of data from Autor, Dorn and Hanson (2013).