NBER WORKING PAPER SERIES

CONTAMINATION BIAS IN LINEAR REGRESSIONS

Paul Goldsmith-Pinkham
Peter Hull
Michal Kolesár

Working Paper 30108
http://www.nber.org/papers/w30108

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2022, Revised February 2024

We thank Alberto Abadie, Jason Abaluck, Isaiah Andrews, Josh Angrist, Tim Armstrong, Kirill Borusyak, Kyle Butts, Clément de Chaisemartin, Peng Ding, Len Goff, Jin Hahn, Xavier D'Haultfoeuille, Simon Lee, Bernard Salanié, Pedro Sant'Anna, Tymon Słoczyński, Isaac Sorkin, Jonathan Roth, Jacob Wallace, Stefan Wager, and numerous seminar participants for helpful comments. Hull acknowledges support from National Science Foundation Grant SES-2049250. Kolesár acknowledges support by the Sloan Research Fellowship and by the National Science Foundation Grant SES-22049356. Mauricio Cáceres Bravo, Jerray Chang, William Cox, and Dwaipayan Saha provided expert research assistance. An earlier draft of this paper circulated under the title "On Estimating Multiple Treatment Effects with Regression". The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Contamination Bias in Linear Regressions
Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár
NBER Working Paper No. 30108
June 2022, Revised February 2024
JEL No. C14,C21,C22,C90

ABSTRACT

We study regressions with multiple treatments and a set of controls that is flexible enough to purge omitted variable bias. We show that these regressions generally fail to estimate convex averages of heterogeneous treatment effects—instead, estimates of each treatment's effect are contaminated by non-convex averages of the effects of other treatments. We discuss three estimation approaches that avoid such contamination bias, including the targeting of easiest-to-estimate weighted average effects. A re-analysis of nine empirical applications finds economically and statistically meaningful contamination bias in observational studies; contamination bias in experimental studies is more limited due to idiosyncratic effect heterogeneity.

Paul Goldsmith-Pinkham
Yale School of Management
165 Whitney Avenue
New Haven, CT 06511
and NBER
paulgp@gmail.com

Michal Kolesár
Department of Economics
278 Julis Romo Rabinowitz Building
Princeton University
Princeton, NJ 08544-1021
mkolesar@princeton.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.edu

A Stata package for multiple treatment effect regression is available at https://github.com/gphk-metrics/stata-mu

1 Introduction

Consider a linear regression of an outcome Y_i on a vector of treatments X_i and a vector of flexible controls W_i. The treatments are assumed to be as good as randomly assigned conditional on the controls. For example, X_i may indicate the assignment of individuals i to different interventions in a stratified randomized control trial (RCT), with the randomization protocol varying across some experimental strata indicators in W_i. Or, in an education value-added model (VAM), X_i might indicate the matching of students i to different teachers or schools with W_i including measures of student demographics and lagged achievement which yield a credible selection-on-observables assumption. The regression might also be the first stage of an instrumental variables (IV) regression leveraging the assignment of multiple decision-makers (e.g. bail judges) indicated in X_i, which is as-good-as-random conditional on some controls W_i. These sorts of regressions are widely used across many fields in economics.[1]

This paper shows that such multiple-treatment regressions generally fail to estimate convex weighted averages of heterogeneous causal effects, and discusses solutions to this problem. The problem may be surprising given an influential result in Angrist (1998), showing that regressions on a single binary treatment D_i and flexible controls W_i estimate a convex average of treatment effects whenever D_i is conditionally as good as randomly assigned. We show that this result does not generalize to multiple treatments: regression estimates of each treatment's effect are generally contaminated by a non-convex average of the effects of other treatments. Thus, the regression coefficient for a given treatment arm incorporates the effects of *all* arms.

We first derive a general characterization of such *contamination bias* in multiple-treatment regressions.[2] We show the core problem by focusing on the special case of a set of mutually exclusive treatment indicators, though our characterization applies even when the treatments are not restricted to be binary or mutually exclusive. To separate the problem from the typical challenge of omitted variables bias (OVB), we assume a best-case scenario where the covariate parametrization is flexible enough to include the treatment propensity scores (e.g., with a linear covariate adjustment, we assume that the propensity scores are linear in the covariates). This condition holds trivially if the only covariates are strata indicators. Under these conditions, we show that the regression coefficient on each treatment identifies a

[1] Prominent RCTs where randomization probabilities vary across strata include Project STAR (Krueger, 1999) and the RAND Health Insurance Experiment (Manning et al., 1987). Prominent VAM examples include studies of teachers (Kane & Staiger, 2008; Chetty et al., 2014), schools (Angrist et al., 2017; Angrist et al., 2021; Mountjoy & Hickman, 2020), and healthcare institutions (Abaluck et al., 2021; Geruso et al., 2020). Prominent "judge IV" examples include Kling (2006), Maestas et al. (2013), and Dobbie and Song (2015).

[2] Our use of the term "contamination" follows Sun and Abraham (2021), and differs from its use in some analyses of clinical trials (e.g. Keogh-Brown et al., 2007) to describe settings where members of one treatment group receive the treatment of another group—what economists typically call "non-compliance". Our "bias" terminology refers to an implication of our result: if a given treatment has constant effects, but the other treatment effects are heterogeneous, the regression estimand is inconsistent for the given treatment effect.

convex weighted average of its causal effects plus a contamination bias term given by a linear combination of the causal effects of other treatments, with weights that sum to zero. Thus, each treatment effect estimate will generally incorporate the effects of other treatments, unless the effects are uncorrelated with the contamination weights. Since these weights sum to zero some are necessarily negative—further complicating the interpretation of the coefficients.

Contamination bias arises because regression adjustment for the confounders in W_i is generally insufficient for making the other treatments ignorable when estimating a given treatment's effect, even when this adjustment is flexible enough to avoid OVB. To see this intuition clearly, suppose the only controls are strata indicators. OVB is avoided when the treatments are as good as randomly assigned within strata. But because the treatments enter the regression linearly, the Angrist (1998) result implies that the causal interpretation of a *given* treatment's coefficient is only guaranteed when its assignment depends linearly on both the strata indicators *and* the other treatment indicators. With mutually exclusive treatments, this condition fails because the dependence is inherently nonlinear—the probability of assignment to a given treatment is zero if an individual is assigned to one of the other treatments, regardless of their stratum, but strata indicators affect the treatment probability otherwise. Such dependence generates contamination bias.[3]

Contamination bias also arises under an alternative "model-based" identifying assumption that—rather than making assumptions on the treatment's "design" (i.e. propensity scores)—posits that the covariate specification spans the conditional mean of the potential outcome under no treatment, $Y_i(0)$. In a linear model with unit and time fixed effects, this reduces to the parallel trends restriction often used in difference-in-differences (DiD) and event study regressions. It is common for X_i to include multiple indicators in such settings—for example, the leads and lags relative to a treatment adoption date used to support the parallel trends assumption or estimate treatment effect dynamics.[4] We show that replacing the restriction on propensity scores in our characterization with an assumption on $Y_i(0)$ generates an additional issue: the own-treatment weights are negative whenever the implicit propensity score model used by the regression to partial out the covariates and the other treatments fits probabilities greater than one. This result shows that the negative weighting and contamination bias issues documented previously in the context of two-way fixed effects regressions (e.g., Goodman-Bacon, 2021; Sun & Abraham, 2021; de Chaisemartin & D'Haultfœuille, 2020, 2022; Callaway & Sant'Anna, 2021; Borusyak et al., 2022; Wooldridge, 2021; Hull, 2018b) are more general—and conceptually distinct—problems.[5] Negative weighting arises because

[3]This issue is distinct from the Freedman (2008b, 2008a) critique of using regression to analyze randomized trials, which concerns estimation, not identification.

[4]Alternatively X_i may indicate multiple contemporaneous treatments, as in certain "mover" regressions.

[5]Our analysis also relates to issues with interpreting multiple-treatment IV estimates (Behaghel et al., 2013; Kirkeboen et al., 2016; Kline & Walters, 2016; Hull, 2018a; Lee & Salanié, 2018; Bhuller & Sigstad, 2022).

regressions leveraging model-based restrictions on $Y_i(0)$ may fit treatment probabilities exceeding one. Contamination bias arises because additive covariate adjustments don't account for the non-linear dependence of a given treatment on the other treatments and covariates. This generates a different form of propensity score misspecification: a non-zero fitted probability of a given treatment, even when one of the other treatments is known to be non-zero.[6]

We then discuss three solutions to the contamination bias problem, and their trade-offs. These solutions apply when the propensity scores are non-degenerate, such as in an RCT or other "design-based" regression specification.[7] First, a conceptually principled solution is to adapt approaches to estimating the average treatment effect (ATE) of a conditionally ignorable binary treatment to the multiple treatment case (e.g. Cattaneo, 2010; Chernozhukov et al., 2018, 2021; Graham & Pinto, 2022). For example, one could run a regression that includes interactions between the treatments and demeaned controls, or combine such regression with inverse propensity score weighting for doubly-robust estimation. Such ATE estimators work well under strong overlap of the covariate distribution for units in each treatment arm. But they may be imprecise under limited overlap or be outright infeasible with overlap failures—common scenarios in observational studies (Crump et al., 2009).

This practical consideration motivates an alternative approach: estimating a weighted average of treatment effects, as regression does in the binary treatment case, while avoiding the contamination bias of multiple-treatment regressions. We derive the weights that are easiest to estimate, in the sense of minimizing a semiparametric efficiency bound under homoskedasticity. These easiest-to-estimate weights are always convex. They also coincide with the implicit linear regression weights when the treatment is binary (i.e. the Angrist (1998) case), formalizing a virtue of regression adjustment. In the multiple treatment case, the easiest-to-estimate weighting can be implemented by a simple second solution: a linear regression which restricts estimation to the individuals who are either in the control group or the treatment group of interest. While trivial to implement, effects estimated using these one-treatment-at-a-time regressions are not directly comparable, since the weighting is treatment-specific. The third solution we discuss is to impose common weights across treatments in our optimization; we show how these weights can be implemented using a weighted regression approach. We give guidance for how researchers can gauge the extent of contamination bias in practice and

[6]While our results are framed in the context of a causal model, we show how analogous results apply to descriptive regressions which seek to estimate averages of conditional group contrasts without assuming a causal framework—as in studies of outcome disparities across multiple racial or ethnic groups, studies of regional variation in healthcare utilization or outcomes, or studies of industry wage gaps.

[7]Solving the contamination bias problem under model-based identification approaches requires either targeting subpopulations of the treated or applying substantive restrictions on the conditional means of potential outcomes under treatment. We do not explore this case as it has already been studied extensively in the DiD context (e.g. de Chaisemartin & D'Haultfœuille, 2022; Sun & Abraham, 2021; Callaway & Sant'Anna, 2021; Borusyak et al., 2022; Wooldridge, 2021).

implement these solutions in a new R and Stata package, `multe`.[8]

We study the empirical relevance of contamination bias in nine applications: six RCTs with stratified randomization and three observational studies of racial disparities. We find economically and statistically meaningful contamination bias in several of these applications. Notably, the largest contamination bias is found in the observational studies while the smallest is bias is found in the experimental studies. In a detailed analysis of one of the experiments (the Project STAR trial) we show that the lack of contamination bias reflects limited correlation between treatment effects and contamination weights and small variation in the contamination weights, rather than limited effect heterogeneity. This analysis highlights the importance of conducting contamination bias diagnostics, particularly in observational studies where propensity score variation may cause large variability in the contamination weights.

We structure the rest of the paper as follows. Section 2 illustrates contamination bias in a simple stylized setting. Section 3 characterizes the general problem, and discusses connections to previous analyses. Section 4 three solutions, and gives guidance for measuring and avoiding contamination bias in practice. Section 5 illustrates these tools using nine applications. Section 6 concludes. Appendix A collects all proofs and extensions. Appendix B discusses the connection between our contamination bias characterization and that in the DiD literature. Details on the applications and additional exhibits are given in Appendices C and D.

2 Motivating Example

We build intuition for the contamination bias problem in two simple examples. We first review how regressions on a single randomized binary treatment and binary controls identify a convex average of heterogeneous treatment effects. We then show how this result fails to generalize when we introduce an additional treatment arm. We base these examples on a stylized version of the Project STAR experiment, which we return to as an application in Section 5.1. The simple structure of these examples helps isolate the core mechanisms of contamination bias. Later sections consider non-experimental settings with richer control specifications, both theoretically and empirically.

2.1 Convex Weights with One Randomized Treatment

Consider the regression of an outcome Y_i on a single treatment indicator $D_i \in \{0, 1\}$, a single binary control $W_i \in \{0, 1\}$, and an intercept:

$$Y_i = \alpha + \beta D_i + \gamma W_i + U_i. \tag{1}$$

[8]The packages are avaiable at CRAN and https://github.com/gphk-metrics/stata-multe, respectively.

By definition, U_i is a mean-zero regression residual that is uncorrelated with D_i and W_i. For example, analysing the Project STAR trial, Krueger (1999) primarily studied the effect of small class size D_i on the test scores Y_i of kindergartners indexed by i. Project STAR randomized students to classes within schools, with the fraction of students assigned to small classes varying by school due to the varying number of total students in each school. To account for this, Krueger (1999) included school fixed effects as controls. Such specifications are often found in stratified RCTs with varying treatment assignment rates across a set of pre-treatment strata. If we imagine two such strata, demarcated by a binary indicator W_i, then eq. (1) corresponds to a stylized two-school version of a Project STAR regression.

We wish to interpret the coefficient β in terms of the causal effects of D_i on Y_i. For this we use potential outcome notation, letting $Y_i(d)$ denote the test score of student i when $D_i = d$. Individual i's treatment effect is then given by $\tau_{1i} = Y_i(1) - Y_i(0)$, and we can write realized achievement as $Y_i = Y_i(0) + \tau_{1i}D_i$. Since treatment assignment is random within schools, D_i is conditionally independent of potential outcomes given W_i: $(Y_i(0), Y_i(1)) \perp D_i \mid W_i$.

Angrist (1998) showed that regression coefficients like β identify a convexly-weighted average of within-strata ATEs. In our Project STAR example, this result shows that:

$$\beta = \phi\tau_1(0) + (1-\phi)\tau_1(1), \quad \text{where} \quad \phi = \frac{\operatorname{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1}\operatorname{var}(D_i \mid W_i = w)\Pr(W_i = w)} \in [0,1] \tag{2}$$

gives a convex weighting scheme, and $\tau_1(w) = E[Y_i(1) - Y_i(0) \mid W_i = w]$ is the ATE in school $w \in \{0,1\}$. Thus, in our example the coefficient β identifies a weighted average of school-specific small classroom effects $\tau_1(w)$ across the two schools.

Equation (2) can be derived by applying the Frisch-Waugh-Lovell (FWL) Theorem. The multivariate regression coefficient β can be written as a univariate regression coefficient from regressing Y_i onto the population residual $\tilde{D}_i$ from regressing D_i onto W_i and a constant:

$$\beta = \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{E[\tilde{D}_i Y_i(0)]}{E[\tilde{D}_i^2]} + \frac{E[\tilde{D}_i D_i \tau_{1i}]}{E[\tilde{D}_i^2]}, \tag{3}$$

where we substitute the potential outcome model for Y_i in the second equality. Since W_i is binary, the propensity score $E[D_i \mid W_i]$ is linear and the residual $\tilde{D}_i$ is mean independent of W_i (not just uncorrelated with it): $E[\tilde{D}_i \mid W_i] = 0$. Therefore,

$$E[\tilde{D}_i Y_i(0)] = E[E[\tilde{D}_i Y_i(0) \mid W_i]] = E[E[\tilde{D}_i \mid W_i]E[Y_i(0) \mid W_i]] = 0. \tag{4}$$

The first equality in eq. (4) follows from the law of iterated expectations, the second equality follows by the conditional random assignment of D_i and the third equality uses $E[\tilde{D}_i \mid W_i] = 0$.

Hence, the first summand in eq. (3) is zero. Analogous arguments show that

$$E[\tilde{D}_i D_i \tau_{1i}] = E[E[\tilde{D}_i D_i \tau_{1i} \mid W_i]] = E[E[\tilde{D}_i D_i \mid W_i] E[\tau_{1i} \mid W_i]] = E[\text{var}(D_i \mid W_i)\tau_1(W_i)],$$

where $\text{var}(D_i \mid W_i) = E[\tilde{D}_i^2 \mid W_i]$ gives the conditional variance of the small-class treatment within schools. Since $E[\text{var}(D_i \mid W_i)] = E[E[\tilde{D}_i^2 \mid W_i]] = E[\tilde{D}_i^2]$, it follows that we can write the second summand in eq. (3) as

$$\beta = \frac{E[\text{var}(D_i \mid W_i)\tau_1(W_i)]}{E[\text{var}(D_i \mid W_i)]} = \phi\tau_1(0) + (1-\phi)\tau_1(1),$$

proving the representation of β in eq. (2).

The key fact underlying this derivation is that the residual $\tilde{D}_i$ from the auxiliary regression of the treatment D_i on the other regressors W_i is mean-independent of W_i. By the FWL theorem, treatment coefficients like β can always be represented as in eq. (3) even without this property. We next show, however, that the remaining steps in the derivation of eq. (2) fail when an additional treatment arm is included. This failure can be attributed to the fact that the auxiliary FWL regression delivers a treatment residual that is uncorrelated with—but not mean-independent of—the other regressors. The lack of mean independence leads to an additional term in the expression for the regression coefficient.

2.2 Contamination Bias with Two Randomized Treatments

In reality, Project STAR randomized students to three mutually exclusive conditions within schools: a control group with a regular class ($D_i = 0$), a treatment that reduced class size ($D_i = 1$), and a treatment that introduced full-time teaching aides ($D_i = 2$). We incorporate this extension of our stylized example by considering a regression of student achievement Y_i on a vector of two treatment indicators, $X_i = (X_{i1}, X_{i2})'$, where $X_{ik} = \mathbb{1}\{D_i = k\}$ indicates assignment to treatment $k = 1, 2$. We continue to include a constant and the school indicator W_i as controls, yielding the regression

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma W_i + U_i. \tag{5}$$

The observed outcome is now given by $Y_i = Y_i(0) + \tau_{i1} X_{i1} + \tau_{i2} X_{i2}$, with $\tau_{i1} = Y_i(1) - Y_i(0)$ and $\tau_{i2} = Y_i(2) - Y_i(0)$ denoting the potentially heterogeneous effects of a class size reduction and introduction of a teaching aide, respectively. As before, we analyze this regression by assuming X_i is conditionally independent of the potential achievement outcomes $Y_i(d)$ given the school indicator W_i: $(Y_i(0), Y_i(1), Y_i(2)) \perp X_i \mid W_i$.

To analyze the coefficient on X_{i1}, we again use the FWL theorem to write

$$\beta_1 = \frac{E[\tilde{\tilde{X}}_{i1} Y_i]}{E[\tilde{\tilde{X}}_{i1}^2]} = \frac{E[\tilde{\tilde{X}}_{i1} Y_i(0)]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i1} \tau_{i1}]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}]}{E[\tilde{\tilde{X}}_{i1}^2]}, \tag{6}$$

where $\tilde{\tilde{X}}_{i1}$ again denotes a population residual, but now from regressing X_{i1} on W_i, a constant, *and* X_{i2}. Unlike before, this residual is uncorrelated with but *not* mean-independent of the remaining regressors (W_i, X_{i2}) because the dependence between X_{i1} and X_{i2} is non-linear. When $X_{i2} = 1$, X_{i1} must be zero regardless of the value of W_i (because they are mutually exclusive) while if $X_{i2} = 0$ the mean of X_{i1} does depend on W_i unless the treatment assignment is completely random. Thus, in general, $\tilde{\tilde{X}}_{i1} \neq X_{i1} - E[X_{i1} \mid W_i, X_{i2}]$.

Because $\tilde{\tilde{X}}_{i1}$ does not coincide with a conditionally de-meaned X_{i1}, we can not generally reduce eq. (6) to an expression involving only the effects of the first treatment arm, τ_{i1}. It turns out that we nevertheless still have $E[\tilde{\tilde{X}}_{i1} Y_i(0)] = 0$, as in eq. (4), since the auxilliary regression residuals are still uncorrelated with any individual characteristic like $Y_i(0)$.[9] The regression thus does not suffer from OVB. However, we do not generally have $E[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}] = 0$. Instead, simplifying eq. (6) by the same steps as before leads to the expression

$$\beta_1 = E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)] \tag{7}$$

as a generalization of eq. (2). Here $\lambda_{11}(W_i) = E[\tilde{\tilde{X}}_{i1} X_{i1} \mid W_i]/E[\tilde{\tilde{X}}_{i1}^2]$ can be shown to be non-negative and to average to one, similar to the ϕ weight in eq. (2). Thus, if not for the second term in eq. (7), β_1 would similarly identify a convex average of the conditional ATEs $\tau_1(W_i) = E[Y_i(1) - Y_i(0) \mid W_i]$. But precisely because $\tilde{\tilde{X}}_{i1} \neq X_{i1} - E[X_{i1} \mid W_i, X_{i2}]$, this second term is generally present: $\lambda_{12}(W_i) = E[\tilde{\tilde{X}}_{i1} X_{i2} \mid W_i]/E[\tilde{\tilde{X}}_{i1}^2]$ is generally non-zero, complicating the interpretation of β_1 by including the conditional effects of the other treatment $\tau_2(W_i) = E[Y_i(2) - Y_i(0) \mid W_i]$.

The second *contamination bias* term in eq. (7) arises because the residualized small class treatment $\tilde{\tilde{X}}_{i1}$ is not conditionally independent of the second full-time aide treatment X_{i2} within schools, despite being uncorrelated with X_{i2} by construction. This can be seen by viewing $\tilde{\tilde{X}}_{i1}$ as the result of an equivalent two-step residualization. First, both X_{i1} and X_{i2} are de-meaned within schools: $\tilde{X}_{i1} = X_{i1} - E[X_{i1} \mid W_i] = X_{i1} - p_1(W_i)$ and $\tilde{X}_{i2} = X_{i2} - E[X_{i2} \mid W_i] = X_{i2} - p_2(W_i)$ where $p_j(W_i) = E[X_{ij} \mid W_i]$ gives the propensity score for treatment j. Second, a bivariate regression of $\tilde{X}_{i1}$ on $\tilde{X}_{i2}$ is used to generate the residuals $\tilde{\tilde{X}}_{i1}$. When the propensity scores vary across the schools (i.e. $p_j(0) \neq p_j(1)$), the relationship

[9] To see this, note that in the auxiliary regression $X_{i1} = \mu_0 + \mu_1 X_{i2} + \mu_2 W_i + \tilde{\tilde{X}}_{i1}$ we can partial out W_i and the constant from both sides to write $\tilde{X}_{i1} = \mu_1 \tilde{X}_{i2} + \tilde{\tilde{X}}_{i1}$. Thus, $\tilde{\tilde{X}}_{i1} = \tilde{X}_{i1} - \mu_1 \tilde{X}_{i2}$ is a linear combination of residuals which, per eq. (4), are both uncorrelated with $Y_i(0)$. It follows that $E[\tilde{\tilde{X}}_{i1} Y_i(0)] = 0$.

between these residuals varies by school, and the line of best fit between $\tilde{X}_{i1}$ and $\tilde{X}_{i2}$ averages across this relationship. As a result, the line of best fit does not isolate the conditional (i.e. within-school) variation in X_{i1}: the remaining variation in $\tilde{\tilde{X}}_{i1}$ will tend to predict X_{i2} within schools, making the *contamination weight* $\lambda_{12}(W_i)$ non-zero.

2.3 Illustration and Intuition

A simple numerical example helps make the contamination bias problem concrete. Suppose in the previous setting that school 0 (indicated by $W_i = 0$) assigned only 5 percent of the students to the small classroom treatment, with 45 percent of the students assigned to the full-time aide treatment and the rest assigned to the control group. In school 1 (indicated by $W_i = 1$), there was a substantially larger push for students to be placed into treatment groups with 45 percent of students assigned to a small classroom, 45 percent assigned to a classroom with a full-time aide, and only 10 percent assigned to the control group. Therefore, $p_1(0) = 0.05$ and $p_2(0) = 0.45$ while $p_1(1) = p_2(1) = 0.45$. Suppose that the schools have the same number of students, so that $\Pr(W_i = 1) = 0.5$. It then follows from the above formulas that $\lambda_{12}(0) = 99/106$ and $\lambda_{12}(1) = -99/106$.

As reasoned above, the contamination weights are non-zero here because the within-school correlation between the residualized treatments, $\tilde{X}_{i1}$ and $\tilde{X}_{i12}$, is heterogeneous: in school 0 it is about -0.2, so that the value of the demeaned class aide treatment is only weakly predictive of the small classroom treatment, while in school 1 it is highly predictive with correlation -0.8. Figure D.1 in Appendix D illustrates this graphically, showing that because the overall regression of $\tilde{X}_{i1}$ on $\tilde{X}_{i2}$ averages over these two correlations, the regression residuals are predictive of the value of the class aide treatment.

To illustrate the potential magnitude of bias in this example, suppose that classroom reductions have no effect on student achievement (so $\tau_1(0) = \tau_1(1) = 0$), but that the effect of a teaching aide varies across schools. In school 1 the aide is highly effective, $\tau_2(1) = 1$, (which may be the reason for the higher push in this school to place students into treatment groups) but in school 0, the aide has no effect, $\tau_2(0) = 0$. By eq. (7), the regression coefficient on the first treatment identifies

$$\beta_1 = E[\lambda_{11}(W_i) \cdot 0] + E[\lambda_{12}(W_i)\tau_2(W_i)] = 0 + (-99/106 \times 1 + 99/106 \times 0)/2 \approx -0.47.$$

Thus, in this example, a researcher would conclude that small classrooms have a sizable negative effect on student achievement—equal in magnitude to around half of the true teaching aide effect in school 1—despite the true small-classroom effect being zero for all students. This treatment effect coefficient can be engineered to match an arbitrary magnitude and sign by varying the heterogeneity of the teaching aide effects across schools.

To build further intuition for eq. (7), it is useful to consider two cases where the contamination bias term is zero. First, note that since regression residuals are by construction uncorrelated with the included regressors, $E[\lambda_{12}(W_i)] = E[\tilde{\tilde{X}}_{i1}X_{i2}]/E[\tilde{\tilde{X}}_{i1}^2] = 0$. Therefore, $E[\lambda_{12}(W_i)\tau_2(W_i)] = E[\lambda_{12}(W_i)\tau_2(W_i)] - E[\lambda_{12}(W_i)]E[\tau_2(W_i)] = \text{cov}(\lambda_{12}(W_i), \tau_2(W_i))$. If the average effects of the teaching aide treatment are constant across the two schools, $\tau_2(1) = \tau_2(0)$, then $\tau_2(W_i)$ is constant, and this covariance is zero such that contamination bias disappears. More generally, when the average teaching aide treatment effects across schools $\tau_2(W_i)$ exhibit idiosyncratic variation, in the sense that they have a weak covariance with the contamination weights across schools, the contamination bias term will be small.

Second, consider the case where X_{i1} and X_{i2} are independent conditional on W_i—such as when the small classroom and teacher aid interventions are independently assigned within schools, in contrast to the previously assumed mutual exclusivity of these treatments. In this case the conditional expectation $E[X_{i1} \mid W_i, X_{i2}] = E[X_{i1} \mid W_i]$ will be linear, since X_{i1} and X_{i2} are unrelated given W_i, and will thus be identified by the auxiliary regression of X_{i1} on W_i, X_{i2}, and a constant. Consequently, the $\tilde{\tilde{X}}_{i1}$ residuals will coincide with $X_{i1} - E[X_{i1} \mid W_i]$. The coefficient on X_{i1} in eq. (5) can therefore be shown to be equivalent to the previous eq. (2), identifying the same convex average of $\tau_1(w)$. This case highlights that dependence across treatments is necessary for the contamination bias to arise.

3 General Problem

We now derive a general characterization of the contamination bias problem, in regressions of an outcome Y_i on a K-dimensional treatment vector X_i and flexible transformations of a control vector W_i. We focus on the case of mutually exclusive indicators $X_{ik} = \mathbb{1}\{D_i = k\}$ for values of an underlying treatment $D_i \in \{0, \ldots, K\}$ (with the $\mathbb{1}\{D_i = 0\}$ indicator omitted). We extend the characterization to a general (i.e. potentially non-binary) X_i in Appendix A.1.

We suppose the effects of X_i on Y_i are estimated by a partially linear model:

$$Y_i = X_i'\beta + g(W_i) + U_i, \tag{8}$$

where β and g are defined as the minimizers of expected squared residuals $E[U_i^2]$:

$$(\beta, g) = \underset{\tilde{\beta} \in \mathbb{R}^K, \tilde{g} \in \mathcal{G}}{\text{argmin}} \; E[(Y_i - X_i'\tilde{\beta} - \tilde{g}(W_i))^2] \tag{9}$$

for some linear space of functions $\mathcal{G}$. This setup nests linear covariate adjustment by setting $\mathcal{G} = \{\alpha + w'\gamma \colon [\alpha, \gamma']' \in \mathbb{R}^{1+\dim(W_i)}\}$, in which case eq. (8) gives a linear regression of Y_i on X_i, W_i, and a constant. The setup also allows for more flexible covariate adjustments—such

as by specifying $\mathcal{G}$ to be a large class of "nonparametric" functions (e.g. Robinson, 1988).

Two examples highlight the generality of this setup:

Example 1 (*Multi-Armed RCT*)**.** W_i is a vector of mutually-exclusive indicators for experimental strata, within which X_i is randomly assigned to individuals i. g is linear.

Example 2 (*Two-Way Fixed Effects*)**.** $i = (j, t)$ indexes panel data, with a fixed set of units $j = 1, \ldots, n$ observed over periods $t = 1, \ldots, T$. $W_i = (J_i, T_i)$ where $J_i = j$ and $T_i = t$ denote the underlying unit and period, and $g(W_i) = \alpha + (\mathbb{1}\{J_i = 2\}, \ldots, \mathbb{1}\{J_i = n\}, \mathbb{1}\{T_i = 2\}, \ldots, \mathbb{1}\{T_i = T\})'\gamma$ includes unit and period indicators. X_i contains indicators for leads and lags relative to a deterministic treatment adoption date, $A(j) \in \{1, \ldots, T\}$ (with at least one lead excluded to prevent collinearity).

Example 1 nests the motivating RCT example in Section 2, allowing for an arbitrary number of experimental strata in W_i and multiple treatment arms in X_i. Example 2 shows that our setup can also nest the kind of regressions considered in a recent literature on DiD and related regression specifications (e.g. Goodman-Bacon, 2021; Hull, 2018b; Sun & Abraham, 2021; de Chaisemartin & D'Haultfœuille, 2020, 2022; Callaway & Sant'Anna, 2021; Borusyak et al., 2022; Wooldridge, 2021). We elaborate on the connections to this literature in Appendix B by considering general two-way fixed effects (TWFE) specifications with non-random treatments. These include specifications with multiple static treatment indicators, as in "mover regressions" that leverage over-time transitions, as well as dynamic event study specifications.[10]

As a first step towards characterizing the β treatment coefficient vector, we solve the minimization problem in eq. (9). Let $\tilde{X}_i$ denote the residuals from projecting X_i onto the control specification, with elements $\tilde{X}_{ik} = X_{ik} - \operatorname{argmin}_{\tilde{g} \in \mathcal{G}} E[(X_{ik} - \tilde{g}(W_i))^2]$. It follows from the projection theorem (e.g. van der Vaart, 1998, Theorem 11.1) that

$$\beta = E[\tilde{X}_i \tilde{X}_i']^{-1} E[\tilde{X}_i Y_i]. \tag{10}$$

Applying the FWL theorem, each treatment coefficient can be written $\beta_k = E[\tilde{\tilde{X}}_{ik} Y_i] / E[\tilde{\tilde{X}}_{ik}^2]$ where $\tilde{\tilde{X}}_{ik}$ is the residual from regressing $\tilde{X}_{ik}$ on $\tilde{X}_{i,-k} = (\tilde{X}_{i1}, \ldots, \tilde{X}_{i,k-1}, \tilde{X}_{i,k+1}, \ldots, \tilde{X}_{iK})'$. Letting $E^*[X_{ik} \mid X_{i,-k}, W_i]$ denote the projection of X_{ik} onto the space $\{X_{i,-k}'\tilde{\delta} + \tilde{g}(W_i) \colon \tilde{\delta} \in \mathbb{R}^{K-1}, \tilde{g} \in \mathcal{G}\}$, we may write these residuals as $\tilde{\tilde{X}}_{ik} = X_{ik} - E^*[X_{ik} \mid X_{i,-k}, W_i]$.

[10] Some papers in this DiD literature study issues we do not consider, such as when researchers fail to include indicators for all relevant treatment states. This specification of X_i will generally add bias terms to our decomposition of β, below. Similarly, we do not consider multicollinearity issues like in Borusyak et al. (2022) by implicitly assuming a unique solution to eq. (9). For event studies this means we assume some units are never treated, with $A(j) = \infty$.

3.1 Causal Interpretation

We now consider the interpretation of each treatment coefficient β_k in terms of causal effects. Let $Y_i(k)$ denote the potential outcome of unit i when $D_i = k$. Observed outcomes are given by $Y_i = Y_i(D_i) = Y_i(0) + X_i'\tau_i$ where τ_i is a vector of treatment effects with elements $\tau_{ik} = Y_i(k) - Y_i(0)$. We denote the conditional expectation of the vector of treatment effects given the controls by $\tau(W_i) = E[\tau_i \mid W_i]$, so that $\tau_k(W_i)$ is the conditional ATE for the kth treatment. We let $p(W_i) = E[X_i \mid W_i]$ denote the vector of propensity scores, so that $p_k(W_i) = \Pr(D_i = k \mid W_i)$. Our characterization of contamination bias doesn't require the propensity scores to be bounded away from 0 and 1 and in fact allows them to be degenerate, i.e. $p_k(w) \in \{0, 1\}$ for all w. This is the case in Example 2, since X_i is a non-random function of W_i. We return to practical questions of propensity score support in Section 4.

We make two assumptions to interpret β_k in terms of the effects τ_i. First, we assume mean-independence of the potential outcomes and treatment, conditional on the controls:

Assumption 1. *$E[Y_i(k) \mid D_i, W_i] = E[Y_i(k) \mid W_i]$ for all k.*

A sufficient condition for this assumption is that the treatment is randomly assigned conditional on the controls, making it conditionally independent of the potential outcomes:

$$(Y_i(0), \ldots, Y_i(K)) \perp D_i \mid W_i. \tag{11}$$

Such conditional random assignment appears in Example 1. In Example 2, where treatment is a non-random function of the unit and time indices in W_i, Assumption 1 holds trivially.

Second, we assume $\mathcal{G}$ is specified such that that one of two conditions holds:

Assumption 2. *Let $\mu_0(w) = E[Y_i(0) \mid W_i = w]$ and recall $p_k(w) = E[X_{ik} \mid W_i = w]$. Either*

$$p_k \in \mathcal{G} \tag{12}$$

for all k, or

$$\mu_0 \in \mathcal{G}. \tag{13}$$

The first condition requires the covariate adjustment to be flexible enough to capture each treatment's propensity score. For example, with a linear specification for g, eq. (12) requires the propensity scores to be linear in W_i (cf. eq. (30) in Angrist & Krueger, 1999). This condition holds trivially in Example 1, since W_i is a vector of indicators for groups within which X_i is randomly assigned. When this condition holds, the projection of the treatment onto the covariates coincides with the vector of propensity scores, and the projection residuals coincide with the conditionally demeaned treatment vector $\tilde{X}_i = X_i - p(W_i)$.

In Example 2, with X_i being a deterministic function of unit and time indices and $g(W_i)$ including unit and time fixed effects, eq. (12) fails because the propensity scores are binary—they cannot be captured by a linear combination of the TWFEs. However, eq. (13) is satisfied by a parallel trends assumption: that the average untreated potential outcomes $Y_i(0)$ are linear in the unit and time effects. We elaborate on this setup in Appendix B.[11]

Under either condition in Assumption 2, the specification of controls is flexible enough to avoid OVB. To see this formally, suppose all treatment effects are constant: $\tau_{ik} = \tau_k$ for all k. This restriction lets us write $Y_i = Y_i(0) + X_i'\tau$, where τ is a vector collecting the constant effects. The only source of bias when regressing Y_i on X_i and controls is then the unobserved variation in the untreated potential outcomes $Y_i(0)$. But it follows from the definition of β in eq. (10) that there is no such OVB when Assumption 2 holds; the coefficient vector identifies the constant effects:

$$\begin{aligned}\beta = E[\tilde{X}_i\tilde{X}_i']^{-1}E[\tilde{X}_iY_i] = E[\tilde{X}_i\tilde{X}_i']^{-1}(E[\tilde{X}_iY_i(0)] + E[\tilde{X}_i\tilde{X}_i']\tau) \\ = E[\tilde{X}_i\tilde{X}_i']^{-1}\underbrace{E[\tilde{X}_iE[Y_i(0) \mid W_i]]}_{=0} + \tau = \tau.\end{aligned}$$

Here the first line uses the fact that $E[\tilde{X}_iX_i'] = E[\tilde{X}_i\tilde{X}_i']$ because $\tilde{X}_i$ is a vector of projection residuals, and the second line uses the law of iterated expectations and Assumption 1. Under eq. (12), $E[\tilde{X}_i \mid W_i] = 0$, so that the term in braces is zero by another application of the law of iterated expectations: $E[\tilde{X}_iE[Y_i(0) \mid W_i]] = E[E[\tilde{X}_i \mid W_i]E[Y_i(0) \mid W_i]] = 0$. It is likewise zero under eq. (13) since $\tilde{X}_i$ is by definition of projection orthogonal to any function in $\mathcal{G}$ such that $E[\tilde{X}_iE[Y_i(0) \mid W_i]] = E[\tilde{X}_i\mu_0(W_i)] = 0$. Hence, OVB is avoided in the constant-effects case so long as either the propensity scores or the untreated potential outcomes are spanned by the control specification. Versions of this robustness property have been previously observed in, for instance, Robins et al. (1992).

When treatment effects are heterogeneous but X_i contains a *single* treatment indicator, β identifies a weighted average of the conditional effects $\tau(W_i)$. Specifically, since by the previous argument we still have $E[\tilde{X}_iY_i(0)] = 0$, it follows from eq. (10) that

$$\beta = \frac{E[\tilde{X}_iX_i\tau_i]}{E[\tilde{X}_i^2]} = E[\lambda_{11}(W_i)\tau(W_i)], \quad \text{with} \quad \lambda_{11}(W_i) = \frac{E[\tilde{X}_iX_i \mid W_i]}{E[\tilde{X}_iX_i]}, \tag{14}$$

where the second equality uses iterated expectations and the identity $E[\tilde{X}_i^2] = E[\tilde{X}_iX_i]$. Under eq. (12), $E[\tilde{X}_iX_i \mid W_i] = E[\tilde{X}_i^2 \mid W_i] = \text{var}(X_i \mid W_i)$, so the weights further simplify

[11] Identification based on eq. (12) can be seen as "design-based" in that it only restricts the treatment assignment process. Identification based on eq. (13) can be seen as "model-based" in that it makes no assumptions on the treatment assignment process but specifies a model for the unobserved untreated potential outcomes.

to $\lambda_{11}(W_i) = \frac{\text{var}(X_i|W_i)}{E[\text{var}(X_i|W_i)]} \geq 0$. This extends the Angrist (1998) result to a general control specification; versions of this extension appear in, for instance, Angrist and Krueger (1999), Angrist and Pischke (2009, Chapter 3.3), and Aronow and Samii (2016).

This result provides a robustness rationale for estimating the effect of a single as-good-as-randomly assigned treatment with a partially linear model (8): so long as the specification of $\mathcal{G}$ is rich enough to make eq. (12) hold, β will identify a convex average of heterogeneous treatment effects. In Section 4 we will derive another rationale for targeting β in this model, showing that the weights $\lambda_{11}(W_i)$ minimize the semiparametric efficiency bound (conditional on the controls) for estimating some weighted-average treatment effect.

Our first proposition shows that with multiple treatments, the interpretation of β becomes more complicated because of contamination bias:

Proposition 1. *Under Assumptions 1 and 2, the treatment coefficients in (8) identify*

$$\beta_k = E[\lambda_{kk}(W_i)\tau_k(W_i)] + \sum_{\ell \neq k} E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)], \tag{15}$$

where, recalling that $E^*[X_{ik} \mid X_{i,-k}, W_i]$ gives the projection of X_{ik} onto the space $\{X'_{i,-k}\tilde{\delta} + \tilde{g}(W_i)\colon \tilde{\delta} \in \mathbb{R}^{K-1}, \tilde{g} \in \mathcal{G}\}$,

$$\lambda_{kk}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik}X_{ik} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]} = \frac{p_k(W_i)(1 - E^*[X_{ik} \mid X_{i,-k} = 0, W_i])}{E[\tilde{\tilde{X}}_{ik}^2]}, \qquad \textit{and}$$
$$\lambda_{k\ell}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik}X_{i\ell} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]} = -\frac{p_\ell(W_i)E^*[X_{ik} \mid X_{i\ell} = 1, W_i]}{E[\tilde{\tilde{X}}_{ik}^2]}$$

with $E[\lambda_{kk}(W_i)] = 1$ and $E[\lambda_{k\ell}(W_i)] = 0$. Furthermore, if eq. (12) holds, $\lambda_{kk}(W_i) \geq 0$.

Proposition 1 shows that the coefficient on X_{ik} in eq. (8) is a sum of two terms. The first term is a weighted average of conditional ATEs $\tau_k(W_i)$, with *own treatment weights* $\lambda_{kk}(W_i)$ that average to one—generalizing the characterization of the single-treatment case, eq. (14). The expression for λ_{kk} implies that these weights are convex if the implicit linear probability model used to compute $\tilde{\tilde{X}}_{ik}$ fits probabilities that lie below one, $E^*[X_{ik} \mid X_{i,-k} = 0, W_i] \leq 1$. The second term is a weighted average of treatment effects for *other* treatments $\tau_\ell(W_i)$, with *contamination weights* $\lambda_{k\ell}(W_i)$ that average to zero. Because the contamination weights are zero on average, they must be negative for some values of the controls unless they are all identically zero.[12] This is the case when the implicit linear probability model correctly predicts that $X_{ik} = 0$ if $X_{i\ell} = 1$.

[12] Proposition 1 complements an algebraic result in Chattopadhyay and Zubizarreta (2021, Section 7.1), which shows that the regression estimator of β_k can be written in terms of weighted sample averages of outcomes among units in different treatment arms (regardless of whether Assumptions 1 and 2 hold). In

Hence, if the linear probability model is correctly specified, i.e. $E[X_{ik} \mid X_{i,-k}, W_i] = X'_{i,-k}\alpha + g_k(W_i)$ for some vector α and $g_k \in \mathcal{G}$, the contamination weights $\lambda_{k\ell}(W_i)$ are zero and the own treatment weights $\lambda_{kk}(W_i)$ are positive. This is the analog of condition (12) if we interpret X_{ik} as a binary treatment of interest and $X'_{i,-k}\alpha + g_k(W_i)$ as a specification for the controls. In other words, the assignment of treatment k must be additively separable between $X_{i,-k}$ and W_i. However, with mutually exclusive treatments, this won't be the case unless treatment assignment is unconditionally random. In particular, since X_{ik} must equal zero if the unit is assigned to one of the other treatments regardless of the value of W_i, under correct specification it must be the case that $\alpha_\ell = -g_k(W_i)$ for all elements α_ℓ of α. This in turn implies that the assignment of treatment k doesn't depend on W_i, which is impossible unless the propensity score $p_k(W_i)$ is constant.

Thus, misspecification in the linear probability model will generally yield nonsensical fitted probabilities $E^*[X_{ik} \mid X_{i\ell} = 1, W_i] \neq 0$ that generate non-zero contamination weights $\lambda_{k\ell}(W_i)$. Furthermore, if the misspecification also yields fitted probabilities $E^*[X_{ik} \mid X_{i,-k} = 0, W_i] > 1$, we will have negative own treatment weights. The last part of Proposition 1 shows that such nonsensible predictions are ruled out if eq. (12) holds.

We make four further remarks on our general characterization of contamination bias:

Remark 1. Since the contamination weights are mean zero, we may write the contamination bias term as $E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)] = \operatorname{cov}(\lambda_{k\ell}(W_i), \tau_\ell(W_i))$. Thus, the treatment coefficient β_k does not suffer from contamination bias if the contamination weights $\lambda_{k\ell}(W_i)$ are uncorrelated with the conditional ATEs $\tau_\ell(W_i)$. This is trivially true if the other treatments are homogeneous, i.e. when $\tau_\ell(W_i) = \tau_\ell$. More generally, contamination bias will be small if the contamination weight exhibits weak covariance with the conditional ATEs. Since $\operatorname{cov}(\lambda_{k\ell}(W_i), \tau_\ell(W_i)) = \operatorname{cor}(\lambda_{k\ell}(W_i), \tau_\ell(W_i)) \operatorname{sd}(\lambda_{k\ell}(W_i)) \operatorname{sd}(\tau_\ell(W_i))$, this is the case when (i) the factors influencing treatment effect heterogeneity are largely unrelated to the factors influencing the treatment assignment process in the sense that $\operatorname{cor}(\lambda_{k\ell}(W_i), \tau_\ell(W_i))$ is close to zero, (ii) the contamination weights display limited variability, and/or (iii) treatment effect heterogeneity in the other treatments $\ell \neq k$ is limited.

Remark 2. Since the weights in eq. (15) are functions of the variances $E[\tilde{\tilde{X}}_{ik}^2]$ and covariances $E[\tilde{\tilde{X}}_{ik}X_{i\ell}]$ and $E[\tilde{\tilde{X}}_{ik}X_{ik}]$, they are identified and can be used to further characterize each β_k coefficient. For example, the contamination bias term can be bounded by the identified contamination weights $\lambda_{k\ell}(W_i)$ and bounds on the heterogeneity in conditional ATEs $\tau_\ell(W_i)$.

Remark 3. The results in Proposition 1 are stated for the case when X_i are mutually exclusive

contrast, our analysis interprets regression *estimands* in terms of weighted averages of conditional ATEs under a broad class of identifying assumptions. In a finite-population setting, Abadie et al. (2020) show that β identifies matrix-weighted averages of individual treatment effect vectors τ_i; however, they do not discuss the interpretation of the estimand.

treatment indicators. In Appendix A.1 we relax this assumption to allow for combinations of non-mutually exclusive treatments (either discrete or continuous). In this case, the own-treatment weights $\lambda_{kk}(W_i)$ may be negative even if eq. (12) holds.

Remark 4. While we derived Proposition 1 in the context of a causal model, an analogous result follows for descriptive regressions that do not assume potential outcomes or impose Assumption 1. Consider, specifically, the goal of estimating an average of conditional group contrasts $E[Y_i \mid D_i = k, W_i = w] - E[Y_i \mid D_i = 0, W_i = w]$ with a partially linear model eq. (8) and replace condition (13) with an assumption that $E[Y_i \mid D_i = 0, W_i = w] \in \mathcal{G}$. The steps that lead to Proposition 1 then show that such regressions also generally suffer from contamination bias: the coefficient on a given group indicator averages the conditional contrasts across all other groups, with non-convex weights. Furthermore, the weights on own-group conditional contrasts are not necessarily positive. These sorts of conditional contrast comparisons are therefore not generally robust to misspecification of the conditional mean, $E[Y_i \mid D_i, W_i]$.

3.2 Implications

Proposition 1 shows that treatment effect heterogeneity can induce two conceptually distinct issues in flexible regression estimates of treatment effects. First, with either single or multiple treatments, there is a negative weighting of a treatment's *own* effects when projecting the treatment indicator onto other treatment indicators and covariates yields fitted values exceeding one, i.e. when $E^*[X_{ik} \mid X_{i,-k} = 0, W_i] > 1$. This issue is relevant in various DiD regressions and related approaches which rely on a model of untreated potential outcomes that ensures eq. (13) holds (e.g. parallel trends assumptions) but which potentially misspecify the assignment model in eq. (12). Although the recent DiD literature focuses on TWFE regressions, Proposition 1 shows such negative weighing can arise more generally—such as when researchers allow for linear trends, interacted fixed effects, or other extensions of the basic parallel trends model. None of these alternative specifications for g are in general flexible enough to capture the degenerate propensity scores and hence ensure that $E^*[X_{ik} \mid X_{i,-k} = 0, W_i] \leq 1$.

Second, in the multiple treatment case, there is a potential for contamination bias from *other* treatment effects—regardless of which condition in Assumption 2 holds. This form of bias is relevant whenever one uses an additive covariate adjustment, no matter how flexibly the covariates are specified. Versions of this problem have been noted in, for example, the Sun and Abraham (2021) analysis of DiD regressions with treatment leads and lags or the Hull (2018b) analysis of mover regressions (see Appendix B).[13] Proposition 1 shows such

[13]The negative weights issue raised in de Chaisemartin and D'Haultfœuille (2020) (when $K = 1$), and the related issue that own-treatment weights may be negative in Sun and Abraham (2021) and de Chaisemartin and D'Haultfœuille (2022) (when $K > 1$), arise because the treatment probability is not linear in the unit and time effects. If eq. (12) holds with $K = 1$, Proposition 1 shows β estimates a convex combination of treatment

contamination bias arises much more broadly, however.

The characterization in Proposition 1 also relates to concerns in interpreting multiple-treatment IV estimates with heterogeneous effects (Behaghel et al., 2013; Kirkeboen et al., 2016; Kline & Walters, 2016; Hull, 2018a; Lee & Salanié, 2018; Bhuller & Sigstad, 2022). This connection comes from viewing eq. (8) as the second stage of an IV model estimated by a control function approach; in the linear IV case, for example, $g(W_i)$ can be interpreted as giving the residuals from a first-stage regression of X_i on a vector of valid instruments Z_i. In the single-treatment case, the resulting β coefficient has an interpretation of a weighted average of conditional local average treatment effects under the appropriate first-stage monotonicity condition (Imbens & Angrist, 1994). But as in Proposition 1 this interpretation fails to generalize when X_i includes multiple mutually-exclusive treatment indicators: each β_k combines the local effects of treatment k with a non-convex average of the effects of other treatments.

Finally, Proposition 1 has implications for single-treatment IV estimation with multiple instruments and flexible controls if the first stage has the form of eq. (8), where now Y_i is interpreted as the treatment and X_i gives the vector of instruments. Proposition 1 shows that the first-stage coefficients on the instruments β_k will not generally be convex weighted average of the true first-stage effects τ_{ik}. Because of this non-convexity, the regression specification may fail to satisfy the effective monotonicity condition even when τ_{ik} is always positive: the cross-instrument contamination of causal effects may cause monotonicity violations, even when specifications with individual instruments do not. This issue is distinct from previous concerns over monotonicity failures in multiple-instrument designs (Mueller-Smith, 2015; Frandsen et al., 2019; Norris, 2019; Mogstad et al., 2021), which are generally also present in such just-identified specifications. It is also distinct from concerns about insufficient flexibility in the control specification when monotonicity holds unconditionally (Blandhol et al., 2022).

This new monotonicity concern may be especially important in "examiner" IV designs, which exploit the conditional random assignment to multiple decision-makers. Many studies leverage such variation by computing average examiner decision rates, often with a leave-one-out correction, and use this "leniency" measure as a single instrument with linear controls. These IV estimators can be thought of as implementing versions of a jackknife IV estimator (Angrist et al., 1999), based on a first stage that uses examiner indicators as instruments, similar to eq. (8). Proposition 1 thus raises a new concern with these IV analyses when controls (such as time fixed effects) are needed to ensure ignorable treatment assignment.

effects. This covers the setting considered in Theorem 1(iv) in Athey and Imbens (2022). In their Comment 2, Athey and Imbens (2022) say that "the sum of the weights [used in Theorem 1(iv)] is one, although some of the weights may be negative". Proposition 1 shows these weights are, in fact, non-negative.

4 Solutions

We now discuss three solutions to the contamination bias problem raised by Proposition 1, each targeting a distinct causal parameter. First, in Section 4.1, we discuss estimation of unweighted ATEs. The other two solutions, discussed in Section 4.2, estimate weighted averages of individual treatment effects using an easiest-to-estimate weighting (EW) scheme in that the weights minimize the semiparametric efficiency bound for estimating weighted ATEs under homoskedasticity. If the weights are allowed to vary across treatments, the EW scheme for each treatment k is recovered by estimating the partially linear model in eq. (8) but in a sample restricted to individuals in the control group and to those receiving treatment k. If the weights are constrained to be common across treatments, this leads to a weighted regression estimator. In Section 4.3, we outline our proposed guidance to researchers in measuring contamination bias and applying these solutions.

Implementing the first solution requires strong overlap (i.e. that treatment propensity scores are bounded away from zero and one) while the other two solutions are not well-defined if the propensity score is fully degenerate. Solutions allowing for degenerate propensity scores require either targeting subpopulations of treated observations or adding substantive restrictions on conditional means of treated potential outcomes (beyond eq. (13), which only restricts untreated potential outcomes). We refer readers to de Chaisemartin and D'Haultfœuille (2022), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Borusyak et al. (2022), and Wooldridge (2021) for such solutions in the context of DiD regressions.

4.1 Estimating Average Treatment Effects

Many estimators exist for the ATE of binary treatments—see Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018) for reviews. Several of these approaches extend naturally to multiple treatments: including matching on covariates or the propensity score, inverse propensity score weighting, interacted regression, or doubly-robust methods (see, among others, Cattaneo (2010), Chernozhukov et al. (2021), and Graham and Pinto (2022)). Here we summarize the last two approaches.

For the interacted regression solution, we adapt the implementation for the binary treatment case discussed in Imbens and Wooldridge (2009, Section 5.3) to multiple treatments. Specifically, consider the specification:

$$Y_i = X_i'\beta + q_0(W_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(W_i) - E[q_k(W_i)]\right) + \dot{U}_i, \tag{16}$$

where $q_k \in \mathcal{G}$, $k = 0, \ldots, K$ and we continue to define β and the functions q_k as minimizers of

$E[\dot{U}_i^2]$. When $\mathcal{G}$ consists of linear functions, eq. (16) specifies a linear regression of Y_i on X_i, W_i, a constant, and the interactions between each treatment indicator X_{ik} and the demeaned control vector $W_i - E[W_i]$. Define $\mu_k(w) = E[Y_i(k) \mid W_i = w]$ for $k = 0, \ldots, K$, so that $\tau_k(w) = \mu_k(w) - \mu_0(w)$. If Assumption 1 holds and $\mathcal{G}$ is furthermore rich enough to ensure $\mu_k \in \mathcal{G}$ for $k = 0, \ldots, K$ then $\beta = \tau$. Moreover, $q_k(w) = \tau_k(w)$ for $k = 1, \ldots, K$, such that the regression identifies both the unconditional and conditional ATEs.

The added interactions in eq. (16) ensure that each treatment coefficient β_k is determined only by the outcomes in treatment arms with $D_i = 0$ and $D_i = k$, avoiding the contamination bias in Proposition 1. Demeaning the $q_k(W_i)$ in the interactions ensures they are appropriately centered to interpret the coefficients on the uninteracted X_{ik} as ATEs.

Estimation of eq. (16) is conceptually straightforward for parametric q_k. In particular, if $\mathcal{G}$ consists of linear functions, one simply estimates

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\tau_k + W_i'\alpha_{W,0} + \sum_{k=1}^{K} X_{ik}(W_i - \overline{W})'\gamma_{W,k} + \dot{U}_i. \tag{17}$$

by ordinary least squares (OLS), where $\overline{W} = \frac{1}{N}\sum_i W_i$ is the sample average of the covariate vector. More generally, to increase the plausibility of the key assumption that $\mu_k \in \mathcal{G}$, one may constrain $\mathcal{G}$ only by nonparametric smoothness assumptions. Given a sequence of basis functions $\{b_j(W_i)\}_{j=1}^{\infty}$, such as polynomials or splines, one then approximates q_k with a linear combination of the first J terms, with J increasing with the sample size, thus tailoring the model complexity to data availability. Given a choice of J, estimation and inference can proceed exactly as in the parametric case; the only difference is that the baseline covariates W_i in eq. (17) are replaced by the basis vector $(b_1(W_i), \ldots, b_J(W_i))'$ and $\overline{W}$ is replaced by the sample average of this expansion. This estimator has been studied in the binary treatment case by Chen et al. (2008) and Imbens et al. (2007), with the latter providing a detailed analysis of how to choose J and the former showing that this sieve estimator achieves the semiparametric efficiency bound under strong overlap: it is impossible to construct another regular estimator of the ATE with smaller asymptotic variance.

An attractive alternative approach combines the interacted regression with inverse propensity score weighting. Instead of using OLS to estimate eq. (16) one uses weighted least squares, weighting observations by the inverse of some estimate $\hat{p}_{D_i}(W_i)$ of the propensity score (see, e.g., Robins et al. (1994), Wooldridge (2007), and Słoczyński and Wooldridge (2018)). An advantage of this approach is that it is doubly-robust: the estimator is consistent so long as either the propensity score estimator is consistent or the outcome model is correct (i.e. $\mu_k \in \mathcal{G}$). A recent literature shows how the double-robustness property, when combined with cross-fitting, reduces the sensitivity of the ATE estimate to overfitting or regularization bias

in estimating the nuisance functions p_k and μ_k. Cross-fitting also allows for using more flexible methods to approximate p_k and μ_k, including modern machine learning methods (see, e.g. Chernozhukov et al., 2018, 2022, 2021).

Either approach should work reliably in stratified RCTs and other settings with strong overlap. But under weak overlap, when propensity scores are not bounded away from zero and one, all of these ATE estimators may be imprecise and have poor finite-sample behavior. This is not a shortcoming of the specific estimator; indeed, Khan and Tamer (2010) show that identification of the ATE is irregular under weak overlap and that it is not possible to estimate it at a $\sqrt{N}$-rate. When overlap fails entirely, with some propensity scores obtaining values of zero or one, the ATE is no longer point-identified. These results formalize the intuition that it is difficult or impossible to reliably estimate the counterfactual outcomes for units with extreme propensity scores.[14] Such extreme propensity scores are common in observational settings. The solutions we discuss next downweight these difficult-to-estimate counterfactuals to address this practical challenge.

4.2 Easiest-to-Estimate Averages of Treatment Effects

Suppose in a sample of observations $i = 1, \ldots, N$ we wish to estimate a weighted average of conditional potential outcome contrasts $\sum_{i=1}^{N} \lambda(W_i) \sum_{k=0}^{K} c_k \mu_k(W_i) / \sum_{i=1}^{N} \lambda(W_i)$, where $\mu_k(W_i) = E[Y_i(k) \mid W_i]$, c is a $(K+1)$-dimensional contrast vector with elements c_k, and $\lambda(W_i)$ is some weighting scheme.[15] We focus on two specifications for the contrast vector, leading to two alternatives to estimating the ATE using eq. (16). First, for separately estimating the effect of each treatment k, we set $c_k = 1$, $c_0 = -1$ and set the remaining entries of c to 0. The contrast of interest then becomes $\sum_{i=1}^{N} \lambda(W_i)\tau_k(W_i) / \sum_{i=1}^{N} \lambda(W_i)$, the weighted ATE of treatment k across different strata. Second, we specify c so as to allow us to simultaneously contrast the effects of all K treatments—we discuss this further below. For each contract vector c, we find the easiest-to-estimate weighting (EW) scheme $\lambda(W_i)$ that leads to the smallest possible standard errors under homoskedasticity.

This optimization problem has four motivations. The first is a robustness motivation: a researcher would like to estimate a given contrast as preceisely as possible, at least under the benchmark of constant treatment effects, while being robust to the possibility that the effects are heterogeneous. While the optimization problem does not impose convexity it turns out that the EW scheme is convex. Hence, the resulting estimand retains an inter-

[14] One approach to limited overlap is trimming: i.e., dropping observations with extreme propensity scores (Crump et al., 2006, 2009; Yang et al., 2016). As with the estimators we derive next, trimming estimators shift the estimand from ATE to easier-to-estimate weighted averages of conditional ATEs.

[15] In a slight abuse of notation relative to Section 3, the weights λ here are not required to average to one. Instead, we scale the estimand by the sum of the weights, $\sum_{i=1}^{N} \lambda(W_i)$.

pretation of identifying a convex average of conditional contrasts when treatment effects are heterogeneous while avoiding the contamination bias displayed by the regression estimator per Proposition 1. This robustness property presumably underlies the popularity of regression as a tool for estimating the effect of a binary treatment: the regression estimator is efficient under homoskedasticity and constant treatment effects while, by the Angrist (1998) result, retaining a causal interpretation under heterogeneous effects.[16]

Second, the EW scheme can be seen as giving a bound on the information available in the data: if the scheme nonetheless yields overly large standard errors, inference on other treatment effects (such as the unweighted ATE) as least as uninformative. Computing the EW standard errors thus reveals whether informative conclusions (regardless of how one specifies the treatment effect of interest) are only possible under additional assumptions or with the aid of additional data. If the EW scheme yields small standard errors even though the standard errors for, say, the unweighted ATE are large, one can conclude that the data is informative about *some* treatment effects—even if it is not informative about the unweighted average.

In fact, our solution below shows that in the binary treatment case the EW scheme is exactly the same as the weights used by regression. To illustrate the second justification in this special case, recall that the treatment weights are proportional to the conditional variance of treatment, $\text{var}(D_i \mid W_i) = p_1(W_i)(1 - p_1(W_i))$, which tend to zero as $p_1(W_i)$ tends to zero or one. Regression thus downweights observations with extreme propensity scores where the estimation of counterfactual outcomes is difficult, avoiding the poor finite-sample behavior of ATE estimators under weak overlap and allowing regression to be informative even in cases when it is not possible to precisely estimate the unweighted ATE. More generally, since under binary treatment regression gives the EW scheme, it establishes the extent to which internally valid and informative inference for *any* causal effect are possible with the data at hand.

Third, the EW scheme can be viewed as offering an intermediate point along a particular robustness-precision "possibility frontier". The ATE estimator based on the interacted specification in eq. (16) lies on one end of this frontier, being the most robust to treatment effect heterogeneity (i.e. retaining a clear interpretation regardless of the form of $\tau(w)$ or how it relates to the propensity scores). But this robustness comes at the potential cost of imprecision and non-standard inference under weak overlap. The regression estimator based on eq. (8) lies on the other end of the frontier: it is likely to be precise even when overlap is weak (and is efficient under homoskedasticity if the partly linear model in eq. (8) is correct, such that treatment effects are constant). But this precision comes at the cost of contamination

[16]There are at least two ways to motivate the interest in convex weights. First, $\lambda(W_i) \geq 0$ ensures the estimand captures average effects for *some* well-defined (and characterizable) subpopulation. Second, it prevents what Small et al. (2017) call a sign-reversal: that if $\tau_k(w)$ has the same sign for all w ($+, 0$ or $-$), then the estimand will also have this sign. Blandhol et al. (2022) call such estimands "weakly causal".

bias under heterogeneous treatment effects. The EW scheme lies in between these extremes, purging contamination bias and retaining good performance under weak overlap by giving up explicit control over the treatment effect weighting, letting it be data-determined.[17]

Finally, while the derivation of the EW scheme is motivated by statistical precision concerns, the resulting estimand can be seen as identifying the impact of a policy that manipulates the treatment via a particular incremental propensity score intervention. We discuss this interpretation in Remark 6 below.

We derive the EW scheme in two steps. First, we establish a precision benchmark—a semiparametric efficiency bound—for estimation of a given weighted average of treatment effects under the idealized scenario that the propensity score is known. Second, we determine which weights λ minimize the semiparametric efficiency bound. We discuss estimation when the propensity score is not known in Section 4.3.

The following proposition establishes the first step of our derivation:

Proposition 2. *Suppose eq. (11) holds in an i.i.d. sample of size N, with known non-degenerate propensity scores $p_k(W_i)$. Let $\sigma_k^2(W_i) = \text{var}(Y_i(k) \mid W_i)$. Consider the problem of estimating the weighted average of contrasts*

$$\theta_{\lambda,c} = \frac{1}{\sum_{i=1}^N \lambda(W_i)} \sum_{i=1}^N \lambda(W_i) \sum_{k=0}^K c_k \mu_k(W_i),$$

where the weighting function λ and contrast vector c are both known. Suppose the weighting function satisfies $E[\lambda(W_i)] \neq 0$, and that the second moments of $\lambda(W_i)$ and $\mu(W_i)$ are bounded. Then, conditional on the controls $W_1, \ldots, W_N$, the semiparametric efficiency bound is almost-surely given by

$$\mathcal{V}_{\lambda,c} = \frac{1}{E[\lambda(W_i)]^2} E\left[\sum_{k=0}^K \frac{\lambda(W_i)^2 c_k^2 \sigma_k^2(W_i)}{p_k(W_i)}\right]. \tag{18}$$

As formalized in the Appendix A.2 proof, $\mathcal{V}_{\lambda,c}$ establishes the lower bound on the asymptotic variance of any regular estimator of $\theta_{\lambda,c}$ under the idealized case of known propensity scores.[18]

To establish the second step, we minimize eq. (18) over λ. Simple algebra shows that the

[17] There are other approaches to resolving the robustness-precision tradeoff, such as seeking precise estimates subject to the weights λ remaining "close" to one, or placing some restrictions on the form of effect heterogeneity, in contrast to leaving it completely unrestricted as we do here (see Mogstad et al. (2018) for an example of this approach in an IV setting). We leave these alternatives to future research.

[18] The efficiency bound for the population analog $\theta^*_{\lambda,c} = E[\lambda(W_i) \sum_{k=0}^K c_k \mu_k(W_i)]/E[\lambda(W_i)]$ has an additional term, $E[\lambda(W_i)^2(\sum_{k=0}^K c_k \mu_k(W_i) - \theta^*_{\lambda,c})^2]/E[\lambda(W_i)]^2$, reflecting the variability of the conditional average contrast. The variance-minimizing weights for $\theta^*_{\lambda,c}$ thus depend on the nature of treatment effect heterogeneity. By focusing on $\theta_{\lambda,c}$, we avoid this term, which allows us give the characterization in eq. (19) without any assumptions about heterogeneity in treatment effects.

EW scheme is (up to an arbitrary constant) given by

$$\lambda_c^*(W_i) = \left(\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(W_i)}{p_k(W_i)} \right)^{-1}. \tag{19}$$

Note that this weighting scheme delivers convex weights, $\lambda_c^* \geq 0$, even though convexity was not imposed in the optimization. Hence, there is no cost in precision if we restrict attention to convex weighted averages of conditional ATEs.

When the contrast vector is selected to estimate the weighted average effect of a particular treatment k, a corollary to Proposition 2 is that regression weights are the easiest-to-estimate:

Corollary 1. *For some $k \geq 1$, let c^k be a vector with elements $c_j^k = 1$ if $j = k$, $c_j^k = -1$ if $j = 0$, and $c_j^k = 0$ otherwise. Suppose that the conditional variance of relevant potential outcomes is homoskedastic: $\sigma_k^2(W_i) = \sigma_0^2(W_i) = \sigma^2$. Then the variance-minimizing weighting scheme is given by $\lambda_{c^k}^* = \lambda^k$, where*

$$\lambda^k(W_i) = \frac{p_0(W_i) p_k(W_i)}{p_0(W_i) + p_k(W_i)}. \tag{20}$$

Per eq. (14), the weighting λ^k coincides with the weighting of conditional ATEs from the partially linear model (8) when it is fit only on observations with $D_i \in \{0, k\}$, provided $p_k/(p_k + p_0) \in \mathcal{G}$.[19] When the treatment D_i is binary, this simply amounts to running a regression on the binary treatment indicator with an additive covariate adjustment.

Corollary 1 thus gives a precision justification for estimating the effect of any given treatment k by a partially linear regression with an additive covariate adjustment in the subsample with $D_i \in \{0, k\}$ under a homoskedasticity benchmark, complementing the robustness motivation discussed earlier.[20] To estimate the effects of all treatments one can run K such regressions, restricting the sample to one treatment arm and the control group.

This precision justification builds on earlier results in Crump et al. (2006, Corollary 5.2) (a working paper version of Crump et al., 2009) and Li et al. (2018, Corollary 1) who show, in the context of a binary treatment, that the weighting $p_1(W_i)(1 - p_1(W_i))$ minimizes the asymptotic variance of a particular class of inverse propensity score weighted estimators. Our Corollary 1 extends the property to all regular estimators, as well as to multiple treatments.

Remark 5. The one-treatment-at-a-time regression can also be motivated as a direct solution to contamination bias in the partially linear regression in eq. (8). In particular, as discussed in

[19] This follows since the propensity score in the subsample is given by $\Pr(D_i = k \mid W_i, D_i \in \{0, k\}) = \frac{p_k(W_i)}{p_0(W_i)+p_k(W_i)}$, so that $\lambda^k(W_i)$ in eq. (20) equals the conditional variance of the treatment indicator times the probability of being in the subsample.

[20] As usual, homoskedasticity is a tractable baseline: the arguments in favor of OLS following Corollary 1 can be extended to favor a (feasible) weighted least squares regression when $\sigma^2(W_i)$ is consistently estimable.

Section 3.1, contamination bias arises because the implicit linear probability model $E^*[X_{ik} \mid X_{i,-k}, W_i]$ incorrectly imposes additive separability between $X_{i,-k}$ and W_i. To solve this issue, one can include interactions between the controls and $X_{i,-k}$. This is analogous to the interacted regression in eq. (16), except we exclude the interaction $X_{ik}(q_k(W_i) - E[q_k(W_i)])$. Running this regression is equivalent to the one-treatment-at-a-time regression.[21]

Remark 6. The population analog of the estimand implied by the weighting in Corollary 1, $E[\lambda_k(W_i)\tau_k(W_i)/E[\lambda_k(W_i)]$, also identifies the effect of a particular marginal policy intervention. Consider the effects of a class of policies indexed by a scalar δ that restrict treatments to $\{0, k\}$ by increasing the propensity score of treatment k to $p_k^\delta(W_i)$ and setting $p_0^\delta(W_i) = 1 - p_k^\delta(W_i)$.[22] Then the marginal effect of the increasing the policy intensity δ per unit treated at $\delta = 0$ is given by $E[\partial p_k^\delta(W_i)/\partial\delta \cdot \tau(W_i)]/E[\partial p_k^\delta(W_i)/\partial\delta]$ (see Zhou & Opacic, 2022, for derivation and discussion). Thus, the weights $\lambda_k(W_i) = \frac{p_0(W_i)p_k(W_i)}{p_0(W_i)+p_k(W_i)}$ identify the marginal policy effect if they correspond to the derivative $\partial p_k^\delta(W_i)/\partial\delta$. This is approximately the case for policies under which individuals with more extreme propensity scores are less likely to exhibit a behavioral response, which as argued in Kennedy (2019) is the case for many policies. If the treatment is binary to begin with, Zhou and Opacic (2022) show that the approximation is exact for policies that increase the log odds of a treatment by a constant δ—such as by increasing the intercept in a logit model for treatment.

A shortcoming of the EW scheme in Corollary 1 is that it is treatment-specific, precluding comparisons of the weighted-average effects across treatments.[23] This issue is especially salient when the control group is arbitrarily chosen, such as in teacher VAM regressions which omit an arbitrary teacher from estimation and seek causal comparisons across all teachers.

We thus turn to the question of how Proposition 2 can be used to select the easiest-to-estimate weighting scheme which allows for simultaneous comparisons across all treatment arms. Suppose that the contrast of interest is drawn at random from a given marginal treatment distribution $\Pr(D_i = k) = \pi_k$, so that $c_j = 1$ with probability $\pi_j(1 - \pi_j)/(1 - \sum_{k=0}^K \pi_k^2)$

[21] To see this, observe that the coefficient on X_{ik} is unchanged if we don't demean the controls and if we replace $q_0(W_i)$ with $(X_{ik} + X_{i0})q_0(W_i)$. That is, if we regress Y_i onto $X_{ik}, (X_{ik} + X_{i0})q_0(W_i)$, $X_{i,-k}$, and the interactions $X_{i\ell}q_\ell(W_i)$ for $\ell \neq 0$. The regressor matrix is block-diagonal: $(X_{ik}, (X_{ik} + X_{i0})q_0(W_i))$ is non-zero iff $D_i \in \{0, k\}$ and the remaining regressors are nonzero iff $D_i \notin \{0, k\}$. Hence, the coefficient on X_{ik} can equivalently be computed by regressing Y_i onto X_{ik} and $q_0(W_i)$ in the sample with $D_i \in \{0, k\}$.

[22] With multiple treatments, policy relevance of any contrast only involving two treatments will generally require the policy to restrict the number of treatments to preclude flows in and out of multiple treatment states. For instance, the ATE gives the effect of comparing two policies: one makes only treatment k available, while the other makes only treatment 0 available.

[23] Formally, for treatments 1 and 2, we estimate the weighted averages $\sum_i \lambda^1(W_i)\tau_1(W_i)/\sum_i \lambda^1(W_i)$ and $\sum_i \lambda^2(W_i)\tau_2(W_i)/\sum_i \lambda^2(W_i)$. Because the weights λ^1 and λ^2 differ, the difference between these estimands cannot generally be written as a convex combination of conditional treatment effects $\tau_1(W_i) - \tau_2(W_i)$. This critique also applies to the own-treatment weights in Proposition 1. Thus even without contamination bias one may find the implicit multiple-treatment regression weighting deficient.

and $c_j = -1$ with the same probability.[24] Let F_π denote this distribution over the (now random) contrasts. If the researcher wishes to report an accurate contrast estimate but needs to commit to a weighting scheme before knowing the contrast of interest, it is optimal to minimize the expected variance

$$\int \mathcal{V}_{\lambda,c} dF_\pi(c) = \frac{1}{E[\lambda(W_i)]^2(1-\sum_{k=0}^K \pi_k^2)} \sum_{k=0}^K E\left[\frac{\lambda(W_i)^2 2\pi_k(1-\pi_k)\sigma_k^2(W_i)}{p_k(W_i)}\right].$$

Minimizing this expression over λ is equivalent to minimizing eq. (18) with $c_k^2 = 2\pi_k(1-\pi_k)$, which yields eq. (19) with this contrast specification as the optimal weighting. Thus, the optimal weights are proportional to $\left(\sum_{k=0}^K \frac{\pi_k(1-\pi_k)\sigma_k^2(W_i)}{p_k(W_i)}\right)^{-1}$. Specializing to the homoskedastic case leads to the following result:

Corollary 2. *Let F_π denote the distribution over possible contrast vectors such that $P_{F_\pi}(c_k = 1) = P_{F_\pi}(c_k = -1) = \pi_j(1-\pi_j)/(1-\sum_{k=0}^K \pi_k^2)$. Suppose that $\sigma_k^2(W_i) = \sigma^2$ for all k. Then the weighting scheme minimizing the average variance bound $\int \mathcal{V}_{\lambda,c} dF_\pi(c)$ is given by:*

$$\lambda^{\text{CW}}(W_i) = \left(\sum_{k=0}^K \frac{\pi_k(1-\pi_k)}{p_k(W_i)}\right)^{-1}.$$

The easiest-to-estimate common weighting (CW) scheme λ^{CW} generalizes the intuition behind the single binary treatment (Corollary 1), placing higher weight on covariate strata where the treatments are evenly distributed, and putting less weight on strata with limited overlap. When the treatment is binary, $K = 1$, the π_k's do not matter and the CW scheme reduces to that in Corollary 1: $\lambda^{\text{CW}}(W_i) = \lambda^1(W_i) = \lambda^0(W_i) = p_1(W_i)p_0(W_i)$. With multiple treatments, however, the weights λ^{CW} remain the same for every treatment—allowing for simultaneous comparisons across all treatment pairs (k, ℓ).

There are two natural choices for the marginal treatment probabilities π. First, if one is equally interested in all contrasts, one can set $\pi_k = 1/(K+1)$. This uniform probability scheme was previously proposed by Li and Li (2019); our characterization in terms of optimizing a semiparametric efficiency bound is, to our knowledge, novel. Second, if more common treatments are of greater interest, we may set π_k to equal to the empirical treatment probabilities $N^{-1}\sum_i X_{ik}$. Treatment arms that have low prevalence would then have little impact on the weighting. This weighting targets precise estimation of contrasts involving more common treatments at the expense of contrasts involving less common treatments. We use this choice in our empirical applications in Section 5. In Section 4.3 below, we show how to implement

[24] Formally, we draw two treatments at random from the given marginal distribution, discarding the draw if the two treatments are equal.

the CW scheme λ^{CW} using a weighted regression approach.

4.3 Practical Guidance in Measuring and Avoiding Contamination Bias

A researcher interested in estimating the effects of multiple mutually exclusive treatments with regression can use Proposition 1 to measure the extent of contamination bias in their estimates. When the propensity score is not fully degenerate, they can further compute one the alternative estimators discussed in the previous subsections. Here we provide practical guidance on both procedures, which we illustrate empirically in the next section.

For simplicity, we focus on the case where g is linear and eq. (8) is estimated by OLS. We assume Assumption 1 and both conditions in Assumption 2 hold, such that all propensity scores p_k and potential outcome conditional expectation functions μ_k are linearly spanned by the controls W_i. These conditions hold, for example, when W_i contains a set of mutually exclusive group indicators. When $\mathcal{G}$ is unrestricted, the recommendations in this section would require non-parametric approximations for g analogous to those discussed in Section 4.1.

Under this setup, we can decompose the OLS estimator $\hat{\beta}$ from the uninteracted regression

$$Y_i = \alpha + \sum_{k=1}^{K} X_{ik}\beta_k + W_i'\gamma + U_i, \tag{21}$$

to obtain a sample analog of the decomposition in Proposition 1. To this end, note that the own-treatment and contamination bias weights in Proposition 1 are identified by the linear regression of X_i on the residuals $\tilde{X}_i$. Specifically, $\lambda_{k\ell}(W_i)$ is given by the (k,ℓ)th element of the $K \times K$ matrix $\Lambda(W_i) = E[\tilde{X}_i\tilde{X}_i']^{-1}E[\tilde{X}_iX_i' \mid W_i]$, which can be estimated by its sample analog $\hat{\Lambda}_i = (\dot{X}'\dot{X})^{-1}\dot{X}_iX_i'$, where $\dot{X}_i$ is the sample residual from an OLS regression of X_i on W_i and a constant and $\dot{X}$ is a matrix collecting these sample residuals. The (k,ℓ)th element of $\hat{\Lambda}_i$ estimates the weight that observation i puts on the ℓth treatment effect in the kth treatment coefficient. For $k = \ell$ this is an estimate of the own-treatment weight in Proposition 1; for $k \neq \ell$ this is an estimate of a contamination weight.

Under linearity, the kth conditional ATE may be written as $\tau_k(W_i) = \gamma_{0,k} + W_i'\gamma_{W,k}$, where $\gamma_{0,k}$ and $\gamma_{W,k}$ are coefficients in the interacted regression specification

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\gamma_{0,k} + W_i'\alpha_{W,0} + \sum_{k=1}^{K} X_{ik}W_i'\gamma_{W,k} + \dot{U}_i. \tag{22}$$

Estimating eq. (22) by OLS yields estimates $\hat{\tau}_k(W_i) = \hat{\gamma}_{0,k} + W_i'\hat{\gamma}_{W,k}$. For each observation i, we stack the set of conditional ATE estimates in a $K \times 1$ vector $\hat{\tau}(W_i)$.

Using the OLS normal equations, we then obtain a sample analog of the population

decomposition in Proposition 1:

$$\hat{\beta} = \sum_{i=1}^{N} \text{diag}(\hat{\Lambda}_i)\hat{\tau}(W_i) + \sum_{i=1}^{N} [\hat{\Lambda}_i - \text{diag}(\hat{\Lambda}_i)]\hat{\tau}(W_i). \tag{23}$$

The first term estimates the own-treatment effect components, $E[\lambda_{kk}(W_i)\tau_k(W_i)]$, while the second term estimates the contamination bias components, $\sum_{\ell \neq k} E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)]$. If the contamination bias term is large for some $\hat{\beta}_k$, it suggests the estimate of the kth treatment effect is substantially impacted by the effects of other treatments. Researchers can also compare the first term of eq. (23) to other weighted averages of own-treatment effects, including the ones discussed next, to gauge the impact of the regression weighting $\text{diag}(\hat{\Lambda}_i)$.[25]

Further analysis of the estimated weights $\hat{\lambda}_{k\ell}(w) = \frac{\sum_{i=1}^{N} \mathbb{1}\{W_i = w\}\hat{\Lambda}_{i,k\ell}}{\sum_{i=1}^{N} \mathbb{1}\{W_i = w\}}$ can shed more light on the regression estimates in $\hat{\beta}$. For example, the contamination weights for $\ell \neq k$ can be plotted against the treatment effect estimates $\hat{\tau}_\ell(W_i)$ to visually assess the sources of contamination bias. Low bias may arise from limited treatment effect heterogeneity, small contamination weights, or a low correlation between the two.

Implementing the alternative estimators is also straightforward under the linearity assumptions. First, estimating eq. (17) by OLS yields estimates of the unweighted ATEs $\tau_k = E[\tau_k(W_i)]$. The estimates are numerically equivalent to $\hat{\tau}_k = \hat{\gamma}_{0,k} + \overline{W}'\hat{\gamma}_{W,k}$, where $\hat{\gamma}_{0,k}$ and $\hat{\gamma}_{W,k}$ are OLS estimates of eq. (22).

The second alternative is to estimate the uninteracted regression,

$$Y_i = \ddot{\alpha}_k + X_{ik}\ddot{\beta}_k + W_i'\ddot{\gamma}_k + \ddot{U}_{ik} \tag{24}$$

among observations assigned either to treatment k or the control group, $D_i \in \{0, k\}$, for each k. These one-treatment-at-a-time regressions estimate the EW scheme from Corollary 1.

The third solution is to estimate the CW scheme λ^{CW} from Corollary 2. If the propensity scores $p(W_i)$ were known, one could run a weighted regression of Y_i onto X_i and a constant, with each observation weighted by $\lambda^{\text{CW}}(W_i)/p_{D_i}(W_i)$. When the weights are unknown, we replace λ^{CW} with its estimate

$$\hat{\lambda}^{\text{CW}}(W_i) = \left(\sum_{k=0}^{K} \frac{\pi_k(1-\pi_k)}{\hat{p}_k(W_i)} \right)^{-1}, \tag{25}$$

[25] When the covariates are not saturated, it is possible that the estimated weighting function $\hat{\Lambda}(w) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{W_i = w\}\hat{\Lambda}_i$ is not positive-definite for some or all w. In particular, the diagonal elements of $\hat{\Lambda}(w)$ need not all be positive. However, it is guaranteed that the diagonal of $\hat{\Lambda}(w)$ sums to one and the non-diagonal weights sum to zero, since $\sum_{i=1}^{N} \hat{\Lambda}_i = I_k$.

where $\hat{p}_k(W_i) = X_{ik} - \dot{X}_{ik}$ denotes estimated propensity scores. We then regress Y_i on X_i, weighting by $\hat{\lambda}^{\text{CW}}(W_i)/\hat{p}_{D_i}(W_i)$. In our applications below we estimate the propensity scores using a multinomial logit model. When the weights π are uniform, this estimator reduces to the estimator studied in Li and Li (2019). The resulting estimator can be written as

$$\hat{\beta}_{\hat{\lambda}^{\text{CW}},k} = \frac{1}{\sum_{i=1}^N \frac{\hat{\lambda}^{\text{CW}}(W_i)}{\hat{p}_k(W_i)} X_{ik}} \sum_{i=1}^N \frac{\hat{\lambda}^{\text{CW}}(W_i)}{\hat{p}_k(W_i)} X_{ik} Y_i - \frac{1}{\sum_{i=1}^N \frac{\hat{\lambda}^{\text{CW}}(W_i)}{\hat{p}_0(W_i)} X_{i0}} \sum_{i=1}^N \frac{\hat{\lambda}^{\text{CW}}(W_i)}{\hat{p}_0(W_i)} X_{i0} Y_i. \tag{26}$$

When the treatment is binary and $\hat{p}_k$ is obtained via a linear regression, this weighted regression estimator coincides with the usual (unweighted) regression estimator that regresses Y_i onto D_i and W_i.[26] Proposition 3 in Appendix A shows that the estimator $\hat{\beta}_{\hat{\lambda}^{\text{CW}}}$ is efficient in the sense that it achieves the semiparametric efficiency bound for estimating $\beta_{\lambda^{\text{CW}}} = \sum_i \lambda^{\text{CW}}(W_i)\tau(W_i)/\sum_i \lambda^{\text{CW}}(W_i)$.[27]

Remark 7. Under homoskedasticity, the second and third solutions yield estimates with smaller asymptotic variance than the estimator of the unweighted ATE. These gains in precision are achieved by changing the estimand to a different convex average of conditional treatment effects. In particular, covariate values w where the propensity score $p_k(w)$ is close to zero for some k will be effectively discarded. In practice, explicitly plotting the treatment weights λ^{CW} and λ^k may help to identify the types of individuals who are down-weighted by these solutions, and to assess the variation in these weights. Plotting them against treatment effect estimates $\hat{\tau}_k$ can help visually assess the extent to which differences in weighting schemes drive differences in between estimates. In particular, the difference between the ATE and any weighted ATE estimand of the effect of treatment k with weights $\lambda(W_i)$, normalized such that $E[\lambda(W_i)] = 1$ is given by $E[\lambda(W_i)\tau_k(W_i)] - E[\tau_k(W_i)] = E[\lambda(W_i)\tau_k(W_i)] - E[\lambda(W_i)]E[\tau_k(W_i)] = \text{cov}(\lambda(W_i), \tau_k(W_i))$. Thus, if the *own* treatment weights λ display only a weak covariance with *own* treatment effect, the weighting will have little effect on the estimand. This is analogous to the observation in Remark 1 that contamination bias reflects the covariance between the contamination weights and treatment effects of the *other* treatments.

[26] To see this, note that in this case $\hat{\lambda}(W_i) = \hat{p}_1(W_i)\hat{p}_0(W_i)$, so that $\hat{\beta}_{\hat{\lambda}^{\text{CW}},1} = \frac{\sum_{i=1}^N (1-\hat{p}_1(W_i))D_iY_i}{\sum_{i=1}^N (1-\hat{p}_1(W_i))D_i} - \frac{\sum_{i=1}^N \hat{p}_1(W_i)(1-D_i)Y_i}{\sum_{i=1}^N \hat{p}_1(W_i)(1-D_i)} = \frac{\sum_{i=1}^N (D_i-\hat{p}_1(W_i))Y_i}{\sum_{i=1}^N (D_i-\hat{p}_1(W_i))^2}$, where the second equality uses the least-squares normal equations $\sum_{i=1}^N X_{i1} = \sum_{i=1}^N \hat{p}_1(W_i)$ and $\sum_i X_{i1}\hat{p}_1(W_i) = \sum_{i=1}^N \hat{p}_1(W_i)^2$.

[27] Similar to the discussion in Section 4.1, it may be attractive to consider a version of $\hat{\beta}_{\hat{\lambda}^{\text{CW}},k}$ that combines the propensity score weighting with a regression adjustment using an estimate of μ_k; we leave detailed study of such an approach to future research.

5 Applications

5.1 Project STAR Application

We first illustrate our framework for analyzing and addressing contamination bias with data from Project STAR, as studied in Krueger (1999). The Project STAR RCT randomized 11,600 students in 79 public Tennessee elementary schools to one of three types of classes: regular-sized (20–25 students), small (target size 13–17 students), or regular-sized with a teaching aide. The proportion of students randomized to the small class size and teaching aide treatment varied over schools, due to school size and other constraints on classroom organization. Students entering kindergarten in the 1985–1986 school year participated in the experiment through the third grade. Other students entering a participating school in grades 1–3 during these years were similarly randomized between the three class types. We focus on kindergarten effects, where differential attrition and other complications with the experimental analysis are minimal.[28]

Column 1 of Panel A in Table 1 reports estimates of kindergarten treatment effects in a sample of 5,868 students initially randomized to the small class size and teaching aide treatments. Specifically, we estimate the partially linear regression (eq. (21)) where Y_i is student i's test score achievement at the end of kindergarten, $X_i = (X_{i1}, X_{i2})$ are indicators for the initial experimental assignment to a small kindergarten class and a regular-sized class with a teaching aide, respectively, and W_i is a vector of school fixed effects. We follow Krueger (1999) in computing Y_i as the average percentile of student i's math, reading, and word recognition score on the Stanford Achievement Test in the experimental sample. As in the original analysis (Krueger, 1999, column 6 of Table V, panel A), we obtain a small class size effect of 5.36 with a heteroskedasticity-robust standard error of 0.78 and a teaching aide effect of 0.18 (standard error: 0.72).[29]

As discussed in Section 2, treatment assignment probabilities vary across the schools indicated by the fixed effects in W_i. If treatment effects also vary across schools in a way that covaries with the contamination weights $\lambda_{k\ell}(W_i)$, we expect the estimated effect of small class sizes to be partly contaminated by the effect of teaching aides (and vice versa). Panel B reports the contamination bias part of the decomposition in eq. (23), which appears minimal for both treatment arms.

[28] Students in regular-sized classes were randomly reassigned between classrooms with and without a teaching aide after kindergarten, complicating the interpretation of the aide effect in later grades. The randomization of students entering the sample after kindergarten was also complicated by the uneven availability of slots in small and regular-sized classes (Krueger, 1999).

[29] Our sample and estimates are very similar to—but not exactly the same as—those in Krueger (1999). We use heteroskedasticity-robust (non-clustered) standard errors throughout this analysis, since the randomization of students to classrooms is at the individual leve.

	A. Treatment effect estimates				
	$\hat{\beta}$	Own	ATE	EW	CW
	(1)	(2)	(3)	(4)	(5)
Small	5.357	5.202	5.561	5.295	5.577
	(0.778)	(0.778)	(0.763)	(0.775)	(0.764)
			[0.744]	[0.743]	[0.742]
Aide	0.177	0.360	0.070	0.263	0.011
	(0.720)	(0.714)	(0.708)	(0.715)	(0.712)
			[0.694]	[0.691]	[0.695]
Number of controls	77				
Sample size	5,868				

	B. Contamination bias estimates		
		Worst-Case Bias	
	Bias	Negative	Positive
	(1)	(2)	(3)
Small class size	0.155	−1.654	1.670
	(0.160)	(0.185)	(0.187)
Teaching aide	−0.183	−1.529	1.530
	(0.149)	(0.176)	(0.177)

Notes: Panel A gives estimates of small class and teaching aide treatment effects for the Project STAR kindergarten analysis. Col. 1 reports estimates from a partially linear model in eq. (21), col. 2 reports the own-treatment component of the decomposition in eq. (23), col. 3 reports the interacted regression estimates based on eq. (17), col. 4 reports estimates based on the EW scheme using one-treatment-at-a-time regressions in eq. (24), and col 5 uses the CW scheme based on eq. (25). Panel B gives the contamination bias component of the decomposition in eq. (23) in col. 1, while cols. 2 and 3 reports the smallest (largest) possible contamination bias from reordering the conditional ATEs to be as negatively (positively) correlated with the cross-treatment weights as possible. Robust standard errors are reported in parentheses. Robust standard errors that assume the propensity scores are known are reported in square brackets.

Table 1: Project STAR contamination bias and treatment effect estimates

It is useful to decompose the contamination bias further into the standard deviation of the school-specific treatment effect $\tau_\ell(W_i)$, standard deviation of the contamination weights, and their correlation, as discussed in Remark 1. Figure D.2 in Appendix D does this graphically, plotting estimates of the school-specific treatment effects $\tau_\ell(W_i)$ against the contamination weights $\lambda_{k\ell}(W_i)$ for $\ell \neq k$. As can be seen from Figure D.2, the variability of school-specific treatment effects is substantial: Adjusting for estimation error, we estimate the standard deviation of $\tau_k(W_i)$ to be 11.0 for the small class treatment and of 9.1 for the aide treatment.[30] Both standard deviations are an order of magnitude larger than the standard errors in Table 1. On the other hand, the standard deviations for the contamination weights for the small class and aide treatment are only moderate: 0.14 and 0.11, respectively. Moreover, the correlation between the conditional treatment effects and the contamination weights is weak: 0.10 for the small class effect estimate and -0.13 for the aide effect estimate. The moderate variation in the contamination weights coupled with weak correlation between the weights and the treatment effects explains why the contamination bias is small, even though the treatment effects vary substantially across schools.

Had the experimental design been such that the contamination weights strongly correlate with the treatment effects, sizable contamination bias could have resulted. To illustrate this, we compute worst-case (positive and negative) weighted averages of the estimated $\tau_\ell(W_i)$ by re-ordering them across the computed cross-treatment weights $\lambda_{k\ell}(W_i)$. This exercise highlights potential scenarios in which the randomization strata happened to have been highly correlated with the effect heterogeneity. Columns 2 and 3 in panel B of Table 1 show that both bounds on possible contamination bias are an order of magnitude larger than the actual contamination bias: $[-1.65, 1.67]$ for the small class size treatment and $[-1.53, 1.53]$ for the teaching aide treatment.[31] Overall, for both treatments, the underlying heterogeneity in this setting makes substantial contamination bias possible even though actual contamination bias turns out to be relatively small.

Columns 2–5 of panel A report four treatment effect estimates that are free of contamination bias. Column 2 gives the own-treatment effect component of the decomposition in eq. (23), netting out the contamination bias estimate from column 1. This doubles the teaching aide effect estimate, from 0.18 to 0.36, but the estimate remains statistically insignificant with standard errors of around 0.71; the small classroom estimate moves very little. The remaining columns report the three solutions to contamination bias discussed in Section 4.

[30] We adjust for estimation error by subtracting the average squared standard error from the empirical variance of the treatment effect estimates and taking the square root.

[31] The point estimates and standard errors in Columns 4 and 5 in Table 1 do not account for the fact that the re-ordering is based on estimates of $\tau_k(W_i)$ rather than the true treatment effects. This biases the reported estimates away from zero. The reported estimates and associated confidence intervals can be interpreted as giving an upper bound for the worst-case contamination bias.

Column 3 estimates the unweighted ATEs of the small class size and teaching aide treatment, by estimating the interacted regression specification in eq. (17). Column 4 estimates the one-treatment-at-a-time regressions in eq. (24) for $k = 1, 2$. Finally, column 5 runs a weighted regression of Y_i onto X_i using the CW scheme in eq. (25).

There turns out to be little difference between these alternative estimates. The small class size effect varies between 5.2 and 5.6, which is close to the original estimate. The teaching aide effect varies between 0.01 and 0.26. To understand this lack of variation, recall from Remark 7 that the difference between the unweighted ATE and an estimand that uses weights $\lambda(W_i)$ is given by the covariance between $\lambda(W_i)$ and the conditional ATEs $\tau_k(W_i)$. Given the sizable variability in the treatment effect estimates, the covariance will be small only if the correlation between the weights and the treatment effects is small and if the weights display limited variability. This turns out to be the case here, as depicted graphically in Figure D.3 in Appendix D. The figure shows that the correlations fall below 0.25 in absolute value for all weighting schemes, and that the weights only vary between 0.7 and 1.2.

As a consequence of strong overlap, the standard errors are similar across the columns. Indeed, the efficiency gain of the EW scheme relative to the ATE based on an efficiency bound comparison using eq. (18) with $\lambda = \lambda^k$ vs $\lambda = 1$ is less than 1.6% for both treatments under homoskedasticity; the gain is even smaller under the CW scheme. The reported standard errors, which allow for heteroskedasticity and don't assume known propensity scores, align with this prediction.[32] As discussed in Remark 8 in Appendix A.3, these standard errors are affected by the assumption of known propensity scores, used to derive the weighting schemes underlying the estimates in columns 2 and 3. To gauge the impact of this assumption, we also report a version of the standard errors computed under the assumption that the sample treatment probabilities in each school match the true propensity scores. This changes the standard errors little, showing that there is minimal cost to estimating the weights.

5.2 Further Applications

We next study the broader relevance of contamination bias using data from eight additional studies with multiple-treatment regressions. These studies were identified by a systematic search of papers in the AEA Data and Code Repository from 2013–2022 (see Appendix C.1 for details). Five studies are experiments like Project STAR; the remaining three use observational regressions to estimate racial disparities across multiple race groups (which we interpret as descriptive, following Remark 4). We replicate a single representative specification for each

[32]The standard errors reported in parentheses in Panel B are valid for the population analogs β_k and $\beta_{\lambda^{\text{CW}}}$, i.e. $E[\lambda^k(W_i)\tau_k(W_i)]/E[\lambda^k(W_i)]$ and $E[\lambda^{\text{CW}}(W_i)\tau_k(W_i)]/E[\lambda^{\text{CW}}(W_i)]$. Since these standard errors are potentially conservative when viewed as standard errors for β_k and $\beta_{\lambda^{\text{CW}}}$, the standard error comparison gives an upper bound on the cost to estimating the weights.

Paper	Journal (1)	Type (2)	Spec. (3)	Sample size Original (4)	Sample size Overlap (5)	var($p(W)$) > 0? (6)
Benhassine et al. (2015)	AEJ:AE	Exp.	5(1)	11,074	6,996	Yes
Cole et al. (2013)	AEJ:AE	Exp.	7(6)	132	73	Yes
de Mel et al. (2013)	AEJ:AE	Exp.	2(2)	520	520	No
Drexler et al. (2014)	AEJ:AE	Exp.	2(2)	796	796	No
Duflo et al. (2015)	AER	Exp.	2A(1)	9,116	8,664	No
Fryer and Levitt (2013)	AER	Obs.	3(4)	8,806	6,623	Yes
Rim et al. (2020)	AER:P&P	Obs.	2(3)	4,037	620	Yes
Weisburst (2019)	AER:P&P	Obs.	2A	7,488	7,488	Yes

Notes: This table summarizes the five experimental studies and three observational studies of racial disparities collected from a search of the AEA Data and Code Repository from 2013–2022 (See Appendix C.1 for details of this search). Column 3 reports the table and panel of the replicated specification with the column or row of the specification in parentheses. See Appendix C.2 for details on the overlap sample and tests for propensity score variation, summarized in columns 6 and 7.

Table 2: Further Applications

paper, corresponding to the first relevant regression discussed in the paper's introduction.[33] Table 2 lists the papers and specifications.

We conduct two preliminary analyses of each study before assessing contamination bias and comparing alternative estimators. First, we ensure that the estimation sample satisfies overlap, since otherwise the decomposition in Proposition 1 is typically not identified. If overlap fails, we identify a large subset of each analysis sample where it is satisfied. Columns 4 and 5 of Table 2 list the number of observations in the full and overlap samples (the sample sizes are equal if the original estimation sample satisfies overlap). Second, we check for propensity score variation in each of the studies. In principle, protocol descriptions can reveal whether some regression controls are necessary (and hence generate propensity score variation) or whether the controls are just added to improve precision. In practice, however, this is not always clear from paper descriptions.[34] In Column 6 of Table 2, we conduct simple statistical tests for whether there is statistically significant propensity score variation. Appendix C.2 details the overlap sample construction and these tests. We replicate the analyses from Table 1

[33] "Relevant" here means a multiple-treatment regression specification with controls, where at least one treatment coefficient was statistically significant. The introduction in Cole et al. (2013) did not discuss any relevant specifications; we instead pick the first specification with variation in treatment probabilities across strata where our results would be most relevant.

[34] Moreover, some regression specifications are run on a non-random subsample of the full experimental population (due to, e.g., attrition, or in a susample analysis). This could generate propensity score variation even in simple experimental protocols.

for each of the eight papers in Appendix C.3; we summarize the takeaways here.

Figure 1 summarizes the statistical and practical significance of contamination bias in the estimated effect of each treatment for each specification (as estimated in the overlap sample). Column A shows the absolute value of the contamination bias t-statistics for each regression coefficient, obtained from the decomposition in eq. (23). In both columns, we sort treatments within papers by this absolute t-statistic and sort papers by the maximum absolute t-statistic across treatments. Column B shows a normalized version of the decomposition that divides each term by the standard error of the regression coefficient. The darker bar shows the own-treatment effect component of the decomposition, while the lighter bar denotes the contamination bias component (which can be of the same or opposite sign).

The figure shows economically and statistically meaningful contamination bias in several regression specifications. Notably, the largest contamination bias is in the observational samples and the smallest is in the experimental samples. This is consistent with the intuition that the variability in the contamination weights is larger in observational studies as a consequence of weaker overlap. Specifications from both de Mel et al. (2013) and Drexler et al. (2014) have some of the smallest contamination bias and also have no statistically significant propensity score variation, consistent with the theoretical results that contamination bias requires variation in the contamination weights which in turn requires variation in the propensity scores. On the other hand, the specification from Cole et al. (2013) shows the most economically meaningful contamination bias among the experimental studies while also exhibiting meaningful propensity score variation.

In Figure 2, we plot the normalized estimates of the treatment effects for each estimator from Table 1. We also plot a line between the estimates from OLS regression and from the common-weights (CW) estimator we propose. Among observational studies, we see substantial variation across the different estimates, and a much larger difference between the OLS estimator and the CW estimator. In the experimental papers, the difference is much smaller.[35] This is consistent with the observational papers including more observations with extreme propensity scores, magnifying the impact of the choice of weighting scheme.

6 Conclusion

Regressions with multiple treatments and flexible controls are common across a wide range of empirical settings in economics. We show that such regressions generally fail to estimate a convex weighted average of treatment effects, with coefficients on each treatment generally contaminated by the effects of other treatments. We provide intuition for why the influen-

[35]The same pattern arises when comparing the estimates in the full sample; see Appendix C.3.

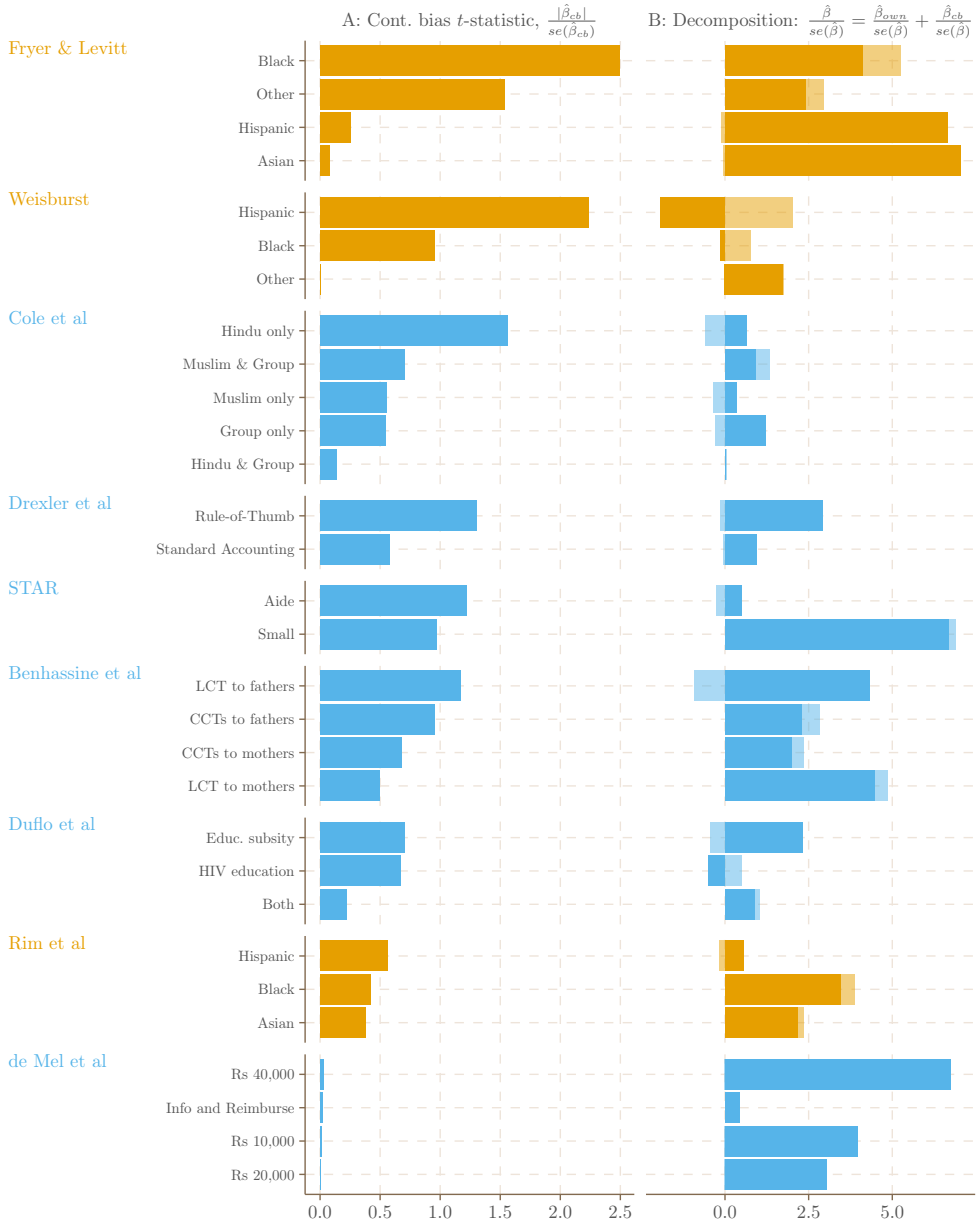

Figure 1: Contamination bias across all applications

Notes: This figure summarizes the analysis of contamination bias in the STAR application and the additional applications in Table 2. The six experimental studies are shown in blue; the three observational studies of racial disparities are shown in orange. Column A shows the absolute value of contamination bias t-statistics for each regression coefficient, given by eq. (23). Column B shows a normalized version of this decomposition that divides each term by the standard error of the regression coefficient. The darker bar shows the own-treatment effect component, while the lighter bar shows the contamination bias component.

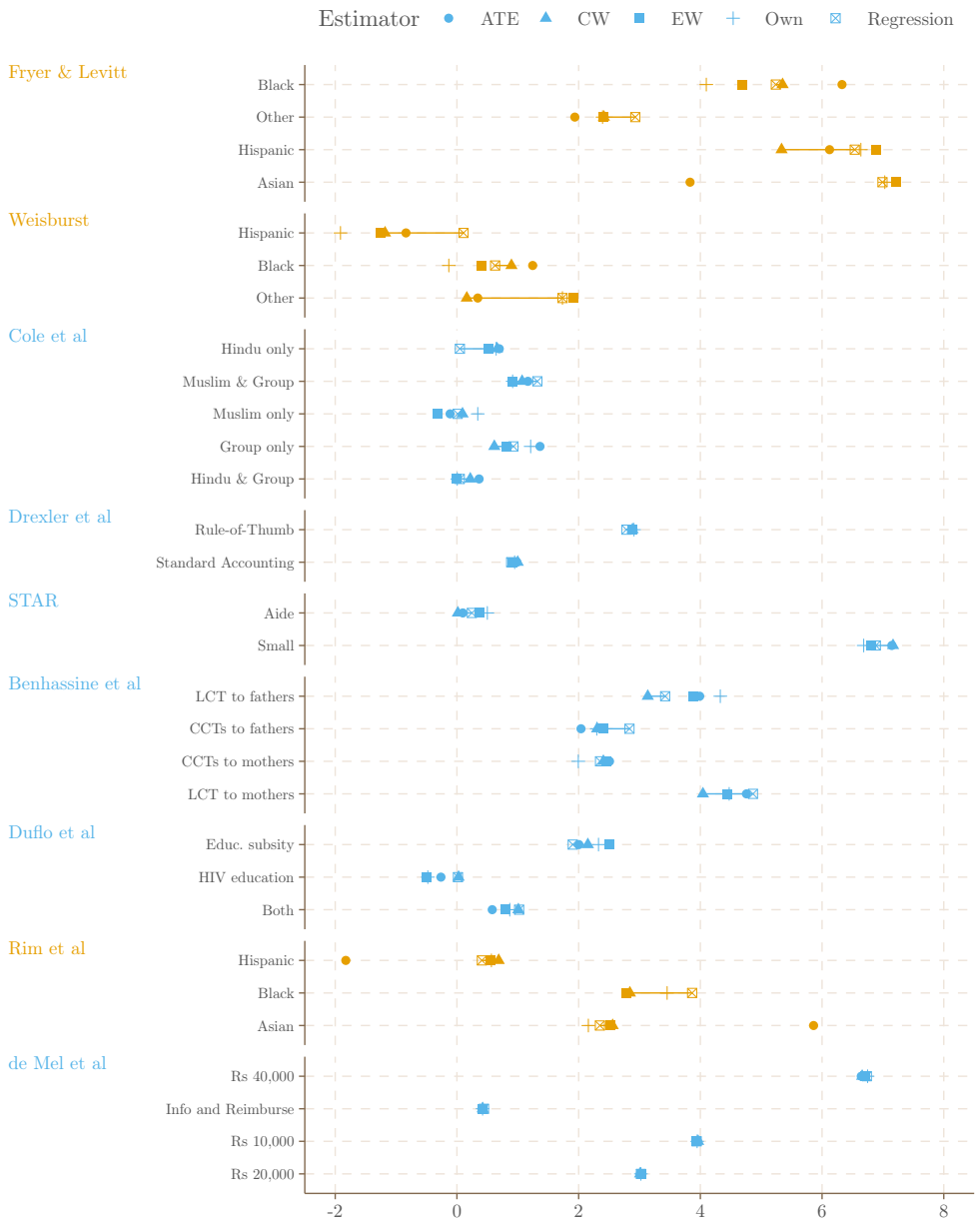

Figure 2: Treatment effect estimates with using different estimators

Notes: This figure plots normalized estimates of treatment effects for each estimator from of Table 1, applied to the STAR application and additional applications in Table 2. The six experimental studies are shown in blue; the three observational studies of racial disparities are shown in orange. Each specification includes a line connecting the estimate from OLS regression and the common-weights (CW) estimator we propose. EW stands for the easiest-to-estimate weighting. For the Rim et al. application the ATE estimate for the "Asian" coefficient equals −8.4, and it is not displayed as it falls outside the axis limits.

tial result of Angrist (1998) fails to generalize to multiple treatments, and show how the contamination bias problem connects to a recent literature studying DiD regressions. We then discuss three alternative estimators that are free of this bias, including a new approach targeting easiest-to-estimate weighted average effects.

Our analysis of nine empirical applications finds economically and statistically meaningful contamination bias in observational studies. Contamination bias in experimental studies is more limited, even in papers that display statistically significant variation in the propensity scores. We also find that the choice among alternative estimators that are free of contamination bias matters more in the observational studies. Overall, our analysis highlights the importance of testing the empirical relevance of theoretical concerns with how regression combines heterogeneous effects.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*(1), 465–503.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), 265–296.

Abaluck, J., Bravo, M. C., Hull, P. D., & Starc, A. (2021). Mortality effects and choice across private health insurance plans. *The Quarterly Journal of Economics*, *136*(3), 1557–1610.

Angrist, J., Hull, P., Pathak, P. A., & Walters, C. (2021). Credible school value-added with undersubscribed school lotteries. *The Review of Economics and Statistics*.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, *66*(2), 249–288.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, *14*(1), 57–67.

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1277–1366, Vol. 3A). Elsevier.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, *132*(2), 871–919.

Aronow, P. M., & Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, *60*(1), 250–267.

Athey, S., & Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, *226*(1), 62–79.

Behaghel, L., Crépon, B., & Gurgand, M. (2013). *Robustness of the encouragement design in a two-treatment randomized control trial* (Discussion Paper No. 7447). Institute for the Study of Labor (IZA). Bonn, Germany.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2015). Turning a shove into a nudge? a "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, *7*(3), 86–125.

Berman, A., & Plemmons, R. J. (1994). *Nonnegative matrices in the mathematical sciences*. Society for Industrial; Applied Mathematics.

Bhuller, M., & Sigstad, H. (2022). *2SLS with multiple treatments*. arXiv: 2205.07836.

Blandhol, C., Bonney, J., Mogstad, M., & Torgovitsky, A. (2022). *When is TSLS actually LATE?* (Working Paper). SSRN.

Borusyak, K., Jaravel, X., & Spiess, J. (2022). *Revisiting event study designs: Robust and efficient estimation*. arXiv: 2108.12419.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772–793.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*(2), 138–154.

Chattopadhyay, A., & Zubizarreta, J. R. (2021). *On the implied weights of linear regression for causal inference*. arXiv: 2104.06581.

Chen, X., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, *36*(2), 808–843.

Chernozhukov, V., Newey, W. K., & Singh, R. (2021). *Automatic debiased machine learning of causal and structural effects*. arXiv: 1809.05224.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, *90*(4), 1501–1535.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Cole, S., Giné, X., Tobacman, J., Topalova, P., Townsend, R., & Vickery, J. (2013). Barriers to household risk management: Evidence from india. *American Economic Journal: Applied Economics*, *5*(1), 104–135.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2006). *Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand* (Working Paper No. 330). National Bureau of Economic Research. Cambridge, MA.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996.

de Chaisemartin, C., & D'Haultfœuille, X. (2022). *Two-way fixed effects regressions with several treatments*. arXiv: 2012.10077.

de Mel, S., McKenzie, D., & Woodruff, C. (2013). The demand for, and consequences of, formalization among informal firms in Sri Lanka. *American Economic Journal: Applied Economics*, *5*(2), 122–150.

Dobbie, W., & Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, *105*(3), 1272–1311.

Drexler, A., Fischer, G., & Schoar, A. (2014). Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics*, *6*(2), 1–31.

Duflo, E., Dupas, P., & Kremer, M. (2015). Education, hiv, and early fertility: Experimental evidence from kenya. *American Economic Review*, *105*(9), 2757–2797.

Frandsen, B., Lefgren, L., & Leslie, E. (2019). *Judging judge fixed effects* (Working Paper No. 25528). National Bureau of Economic Research. Cambridge, MA.

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193.

Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, *2*(1), 176–196.

Fryer, R. G., & Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, *103*(2), 981–1005.

Geruso, M., Layton, T., & Wallace, J. (2020). *Are all managed care plans created equal? evidence from random plan assignment in Medicaid* (Working Paper No. 27762). National Bureau of Economic Research. Cambridge, MA.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Graham, B. S., & Pinto, C. C. d. X. (2022). Semiparametrically efficient estimation of the average linear regression function. *Journal of Econometrics*, *226*(1), 115–138.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*(2), 315–331.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Hull, P. D. (2018a). *IsoLATEing: Identifying counterfactual-specific treatment effects with cross-stratum comparisons* (Working Paper). SSRN.

Hull, P. D. (2018b). *Estimating treatment effects in mover designs*. arXiv: 1804.06721.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Imbens, G. W., Newey, W. K., & Ridder, G. (2007). *Mean-squared-error calculations for average treatment effects* [Unpublished manuscript, MIT].

Imbens, G. W., & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. 14607). National Bureau of Economic Research. Cambridge, MA.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, *114*(526), 645–656.

Keogh-Brown, M. R., Bachmann, M. O., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J. N. V., Torgerson, D. J., Miles, S., Elbourne, D., Harvey, I., & Campbell, M. J. (2007). Contamination in trials of educational interventions. *Health Technology Assessment*, *11*(43).

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*(6), 2021–2042.

Kirkeboen, L. J., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, *131*(3), 1057–1111.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *Quarterly Journal of Economics*, *131*(4), 1795–1848.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, *96*(3), 863–876.

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

Lee, S., & Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, *86*(6), 1939–1963.

Li, F., & Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, *13*(4), 2389–2415.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propenstiy score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400.

Maestas, N., Mullen, K. J., & Strand, A. (2013). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review*, *103*(5), 1797–1829.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review*, *77*(3), 251–277.

McNeney, B., & Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *Journal of Statistical Planning and Inference*, *91*(2), 441–480.

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619.

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, *111*(11), 3663–3698.

Mountjoy, J., & Hickman, B. (2020). *The returns to college(s): Estimating value-added and match effects in higher education* (Working Paper No. 2020-08). University of Chicago, Becker Friedman Institute for Economics. Cambridge, MA.

Mueller-Smith, M. (2015). *The criminal and labor market impacts of incarceration* [Unpubished manuscript, University of Michigan].

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*(6), 1349–1382.

Norris, S. (2019). *Examiner inconsistency: Evidence from refugee appeals* (Working Paper No. 2018-75). University of Chicago, Becker Friedman Institute for Economics.

Rim, N., Ba, B., & Rivera, R. (2020). Disparities in police award nominations: Evidence from chicago. *AEA Papers and Proceedings*, *110*, 447–451.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, *48*(2), 479.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, *56*(4), 931.

Słoczyński, T., & Wooldridge, J. M. (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory*, *34*(1), 112–133.

Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A., & Brookhart, M. A. (2017). Instrumental variable estimation with a stochastic monotonicity assumption. *Statistical Science*, *32*(4), 561–579.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

van der Vaart, A., & Wellner, J. A. (1989). *Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory, with applications to convolution and asymptotic minimax theorems* [Unpublished manuscript, Unversity of Seattle].

van der Vaart, A., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Weisburst, E. K. (2019). Police use of force as an extension of arrests: Examining disparities across civilian and officer race. *AEA Papers and Proceedings*, *109*, 152–156.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, *141*(2), 1281–1301.

Wooldridge, J. M. (2021). *Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators* (Working Paper). SSRN.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, *72*(4), 1055–1065.

Zhou, X., & Opacic, A. (2022). *Marginal interventional effects*. arXiv: 2206.10717.

Appendix A Proofs and additional results

A.1 Proof of Proposition 1

We prove a generalization of the Proposition 1 which allows any vector of treatments X_i (which may not be binary or mutually exclusive). We continue to consider the partially linear model in eq. (8), and maintain Assumption 2, as well as conditional mean-independence of the potential outcomes $E[Y_i(x) \mid X_i, W_i] = E[Y_i(x) \mid W_i]$, which extends Assumption 1. We also assume that the potential outcomes $Y_i(x)$ are linear in x, conditional on W_i:

$$E[Y_i(x) \mid W_i = w] = E[Y_i(0) \mid W_i = w] + x'\tau(w),$$

for some function τ. This condition holds trivially in the main-text discussion of mutually exclusive binary treatments. More generally, $\tau_k(w)$ corresponds to the conditional average effect of increasing X_{ik} by one unit among observations with $W_i = w$. Although this assumption is not essential, it considerably simplifies the derivations. We continue to define $\tau = E[\tau(W_i)]$ as the average vector of per-unit effects.

We now prove that under these assumptions β_k is given by the expression in eq. (15). We further prove that $E[\lambda_{kk}(W_i)] = 1$ and $E[\lambda_{k\ell}(W_i)] = 0$ for $\ell \neq k$ in general, and give a more detailed characterization of the weights in the case of mutually exclusive treatment indicators.

First note that by iterated expectations and conditional mean-independence, $E[\tilde{\tilde{X}}_{ik} Y_i] = E[E[\tilde{\tilde{X}}_{ik} Y_i \mid X_i, W_i]] = E[\tilde{\tilde{X}}_{ik} E[Y_i(0) \mid W_i]] + E[\tilde{\tilde{X}}_{ik} X_i' \tau(W_i)]$. By definition of projection, $E[\tilde{X}_i g(W_i)] = 0$ for all $g \in \mathcal{G}$ (van der Vaart, 1998, Theorem 11.1); thus if eq. (13) holds $E[\tilde{\tilde{X}}_{ik} E[Y_i(0) \mid W_i]] = 0$. Similarly, under eq. (12), $E[\tilde{\tilde{X}}_{ik} \mid W_i] = 0$, so by iterated expectations, $E[\tilde{\tilde{X}}_{ik} E[Y_i(0) \mid W_i]] = E[E[\tilde{\tilde{X}}_{ik} \mid W_i] E[Y_i(0) \mid W_i]] = 0$. Thus,

$$\beta_k = \frac{E[\tilde{\tilde{X}}_{ik} X_i' \tau(W_i)]}{E[\tilde{\tilde{X}}_{ik}^2]} = \frac{E[\tilde{\tilde{X}}_{ik} X_{ik} \tau_k(W_i)]}{E[\tilde{\tilde{X}}_{ik}^2]} + \frac{\sum_{\ell \neq k} E[\tilde{\tilde{X}}_{ik} X_{i\ell} \tau_\ell(W_i)]}{E[\tilde{\tilde{X}}_{ik}^2]}.$$

This proves eq. (15).

To show that $E[\lambda_{kk}(W_i)] = 1$ and $E[\lambda_{k\ell}(W_i)] = 0$ for $\ell \neq k$ in general, note that

$$E[\lambda_{kk}(W_i)] = \frac{E[\tilde{\tilde{X}}_{ik} X_{ik}]}{E[\tilde{\tilde{X}}_{ik}^2]} = 1,$$

since $\tilde{\tilde{X}}_{i,k}$ is a residual from projecting X_{ik} onto the space spanned by functions of the form $\tilde{g}(W_i) + X_{i,-k}' \tilde{\beta}_{-k}$, so that $E[\tilde{\tilde{X}}_{ik} X_{ik}] = E[\tilde{\tilde{X}}_{ik}^2]$. Furthermore, $\tilde{\tilde{X}}_{i,k}$ must also be orthogonal to $X_{i,-k}$ by definition of projection, so that $E[\lambda_{k\ell}(W_i)] = E[\tilde{\tilde{X}}_{ik} X_{i\ell}] / E[\tilde{\tilde{X}}_{ik}^2] = 0$.

Finally, if X_i are mutually exclusive treatment indicators, write $E^*[X_{ik} \mid X_{i,-k}, W_i] =$

$X'_{i,-k}\tilde{\delta}_k + \tilde{g}_k(W_i)$. Since $X_{ik}X_{i,-k} = 0$, we may write

$$\lambda_{kk}(W_i) = \frac{p_k(W_i)(1-\tilde{g}_k(W_i))}{E[\tilde{\tilde{X}}_{ik}^2]} = \frac{p_k(W_i)(1-E^*[X_{ik} \mid X_{i,-k}=0, W_i])}{E[\tilde{\tilde{X}}_{ik}^2]},$$

and, by similar arguments, $\lambda_{k\ell}(W_i) = -p_\ell(W_i)E^*[X_{ik} \mid X_{i\ell}=1, W_i]/E[\tilde{\tilde{X}}_{ik}^2]$, which yields the second expression for the weights. It remains to show that $\lambda_{kk}(W_i) \geq 0$ if eq. (12) holds and X_i consists of mutually exclusive indicators. To that end, observe that $\lambda_{k\ell}(W_i)$ is given by the (k,ℓ) element of

$$\Lambda(W_i) = E[\tilde{X}_i\tilde{X}_i']^{-1}E[\tilde{X}_iX_i' \mid W_i]$$

If eq. (12) holds, then we can write this as $\Lambda(W_i) = E[v(W_i)]^{-1}v(W_i)$ where $v(W_i) = E[\tilde{X}_i\tilde{X}_i' \mid W_i]$. If X is a vector of mutually exclusive indicators, then $v(W_i) = \text{diag}(p(W_i)) - p(W_i)p(W_i)'$. Let $v_{-k}(W_i)$ denote the submatrix with the kth row and column removed, and let $p_{-k}(W_i)$ denote subvector with the kth row removed. Then by the block matrix inverse formula,

$$\lambda_{kk}(W_i) = \frac{p_k(W_i)(1-p_k(W_i)) - E[p_k(W_i)p_{-k}(W_i)']E[v_{-k}(W_i)]^{-1}p_{-k}(W_i)p_k(W_i)}{E[p_k(W_i)(1-p_k(W_i))] - E[p_k(W_i)p_{-k}(W_i)']E[v_{-k}(W_i)]^{-1}E[p_k(W_i)p_{-k}(W_i)]}$$

Note $p_0(W_i) = 1 - \sum_{k=1}^K p_k(W_i)$ and $p_k(W_i)p_{-k}(W_i) = v_{-k}(W_i)\iota - p_0(W_i)p_{-k}(W_i)$, where ι denotes a $(K-1)$-vector of ones. Thus, the numerator can be written as

$$\begin{aligned} p_k(W_i)(1-p_k(W_i)) &- \iota' p_{-k}(W_i)p_k(W_i) \\ &+ E[p_0(W_i)p_{-k}(W_i)']E[v_{-k}(W_i)]^{-1}p_{-k}(W_i)p_k(W_i) \\ &= p_k(W_i)p_0(W_i) + E[p_0(W_i)p_{-k}(W_i)']E[v_{-k}(W_i)]^{-1}p_{-k}(W_i)p_k(W_i). \end{aligned}$$

The eigenvalues of $E[v_{-k}(W_i)]$ are positive because it is a covariance matrix. Furthermore, the off-diagonal elements of $E[v(W_i)]$ are negative, and hence the off-diagonal elements of $E[v_{-k}(W_i)]$ are also negative. It therefore follows that $E[v_{-k}(W_i)]$ is an M-matrix (Berman & Plemmons, 1994, property D_{16}, p. 135). Hence, all elements of $E[v_{-k}(W_i)]^{-1}$ are positive (Berman & Plemmons, 1994, property N_{38}, p. 137). Thus, both summands in the above expression are positive, so that $\lambda_{kk}(W_i) \geq 0$.

A.2 Proof of Proposition 2

The parameter of interest $\theta_{\lambda,c}$ depends on the realizations of the controls. We therefore derive the semiparametric efficiency bound conditional on the controls; i.e. we show that eq. (18) is almost-surely the variance bound for estimators that are regular conditional on the controls. Relative to the earlier results in Hahn (1998) and Hirano et al. (2003), we need to account

for the fact that the data are no longer i.i.d. once we condition on the controls.

To that end, we use the notion of semiparametric efficiency based on the convolution theorem of van der Vaart and Wellner (1989, Theorem 2.1) (see also van der Vaart & Wellner, 1996, Chapter 3.11). We first review the result for convenience. Consider a model $\{P_{n,\theta} \colon \theta \in \Theta\}$ parametrized by (a possibly infinite-dimensional) parameter θ. Let $\dot{\mathcal{P}}$ denote a tangent space, a linear subspace of some Hilbert space with an inner product $\langle \cdot, \cdot \rangle$. Suppose that the model is locally asymptotically normal (LAN) at θ relative to a tangent space $\dot{\mathcal{P}}$: for each $g \in \dot{\mathcal{P}}$, there exists a sequence $\theta_n(g)$ such that the likelihood ratios are asymptotically quadratic, $dP_{n,\theta_n(g)}/dP_{n,\theta} = \Delta_{n,g} - \langle g, g \rangle/2 + o_{P_{n,\theta}}(1)$, where $(\Delta_{n,g})_{g \in \dot{\mathcal{P}}}$ converges under $P_{n,\theta}$ to a Gaussian process with covariance kernel $\langle g_1, g_2 \rangle$. Suppose also that the parameter $\beta_n(P_{n,\theta})$ is differentiable: for each g, $\sqrt{n}(\beta_n(P_{n,\theta_n(g)}) - \beta_n(P_{n,\theta})) \to \langle \psi, g \rangle$ for some ψ that lies in the completion of $\dot{\mathcal{P}}$. Then the semiparametric efficiency bound is given by $\langle \psi, \psi \rangle$: the asymptotic distribution of any regular estimator of this parameter, based on a sample $\mathbf{S}_n \sim P_{n,\theta}$, is given by the convolution of a random variable $Z \sim \mathcal{N}(0, \langle \psi, \psi \rangle)$ and some other random variable U that is independent of Z.

To apply this result in our setting, we proceed in three steps. First, we define the tangent space and the probability-one set over which we will prove the efficiency bound. Next, we verify that the model is LAN. Finally, we verify differentiability and derive the efficient influence function ψ.

Step 1 By the conditional independence assumption in eq. (11), we can write the density of the vector $(Y_i(0), \ldots, Y_i(K), D_i)$ (with respect to some σ-finite measure) conditional on $W_i = w$ as $f(y_0, \ldots, y_K \mid w) \cdot \prod_{k=0}^{K} p_k(w)^{\mathbb{1}\{d=k\}}$, where f denotes the conditional density of the potential outcomes, conditional on the controls. The density of the observed data $\mathbf{S}_N = \{(Y_i, D_i)\}_{i=1}^N$ conditional on $(W_1, \ldots, W_N) = (w_1, \ldots, w_N)$ is given by $\prod_{i=1}^N \prod_{k=0}^K (f_k(y_i \mid w_i) p_k(w_i))^{\mathbb{1}\{d_i=k\}}$, where $f_k(y \mid w) = \int f(y_k, y_{-k} \mid w) dy_{-k}$.

Since the propensity scores are known, the model is parametrized by $\theta = f$. Consider one-dimensional submodels of the form $f_k(y \mid w; t) = f_k(y \mid w)(1 + t \times s_k(y \mid w))$, where the function s_k is bounded and satisfies $\int s_k(y \mid w) f_k(y \mid w) dy = 0$ for all $w \in \mathcal{W}$ with $\mathcal{W}$ denoting the support of W_i. For small enough t, we have $f_k(y \mid w; t) \geq 0$ by boundedness of s_k; hence $f_k(y \mid w; t)$ is a well-defined density for t small enough. The joint log-likelihood, conditional on the controls, is given by

$$\sum_{i=1}^{N} \sum_{k=0}^{K} \mathbb{1}\{D_i = k\} (\log f_k(Y_i \mid w_i; t) + \log p_k(w_i)).$$

The score at $t = 0$ is $\sum_{i=1}^N s(Y_i, D_i \mid w_i)$, with $s(Y_i, D_i \mid w_i) = \sum_{k=0}^K \mathbb{1}\{D_i = k\} s_k(Y_i \mid w_i)$.

This result suggests defining the tangent space to consist of functions $s(y, d \mid w) = \sum_{k=0}^{K} \mathbb{1}\{d = k\} s_k(y \mid W_i = w)$, such that s_k is bounded and satisfies $\int s_k(y \mid w) f_k(y \mid w) dy = 0$ for all $w \in \mathcal{W}$. Define the inner product on this space by $\langle s_1, s_2 \rangle = E[s_1(Y_i, D_i \mid W_i) s_2(Y_i, D_i \mid W_i)]$. Note this is a marginal (rather than a conditional) expectation, over the unconditional distribution (Y_i, D_i, W_i) of the observed data.

We will prove the efficiency bound on the event $\mathcal{E}$ that (i) $\frac{1}{N}\sum_{i=1}^{N} E[s(Y_i, D_i \mid W_i)^2 \mid W_i] \to E[s(Y, D_i \mid W_i)^2]$, (ii) $\frac{1}{N}\sum_{i=1}^{N} \lambda(W_i) \to E[\lambda(W_i)]$, and (iii) $\frac{1}{N}\sum_{i=1}^{N} \lambda(W_i) \sum_{k=0}^{K} c_k \cdot E[Y_i(k) s_k(Y_i(k) \mid W_i) \mid W_i] \to \sum_{k=0}^{K} c_k E[\lambda(W_i) Y_i(k) s_k(Y_i(k) \mid W_i)]$. By assumptions of the proposition, these are all averages of functions of W_i with finite absolute moments. Hence, by the law of large numbers, $\mathcal{E}$ is a probability one set.

Step 2 We verify that the conditions (3.7–12) of Theorem 3.1 in McNeney and Wellner (2000) hold on the set $\mathcal{E}$ conditional on the controls, with $\theta_N(s) = f(\cdot \mid \cdot; 1/\sqrt{N})$. Let $\alpha_{Ni} = \prod_{k=0}^{K} (f_k(Y_i \mid w_i; 1/\sqrt{N}) / f_k(Y_i \mid w_i))^{\mathbb{1}\{D_i = k\}} = \prod_{k=0}^{K} (1 + s_k(Y_i \mid w)/\sqrt{N})^{\mathbb{1}\{D_i = k\}}$ denote the likelihood ratio associated with the ith observation. Since this is bounded by the boundedness of s_k, condition (3.7) holds. Also since $(1 + t s_k)^{1/2}$ is continuously differentiable for t small enough, with derivative $s_k / 2\sqrt{1 + t s_k}$, it follows from Lemma 7.6 in van der Vaart (1998) that $N^{-1} \sum_{i=1}^{N} E[\sqrt{N}(\alpha_{Ni}^{1/2} - 1) - s(Y_i, D_i \mid w_i)/2 \mid W_i = w_i]^2 \to 0$ such that the quadratic mean differentiability condition (3.8) holds. Since s_k is bounded, the Lindeberg condition (3.9) also holds. Next, $\frac{1}{N}\sum_{i=1}^{N} E[s(Y_i, D_i \mid W_i)^2 \mid W_i]$ converges to $E[s(Y, D_i \mid W_i)^2] = \langle s, s \rangle$ on $\mathcal{E}$ by assumption. Hence, conditions (3.10) and (3.11) also hold. Since the scores $\Delta_{N,s} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} s(Y_i, D_i \mid w_i)$ are exactly linear in s, condition (3.12) also holds. It follows that the model is LAN on $\mathcal{E}$.

Step 3 Write the parameter of interest $\theta_{\lambda,c}$ as $\beta_N(f) = \sum_{i=1}^{N} \lambda(w_i) \int y \sum_{k=0}^{K} c_k f_k(y \mid w_i) dy / \sum_{i=1}^{N} \lambda(w_i)$. It follows that

$$
\begin{aligned}
&\sqrt{N}(\beta_N(f(\cdot \mid \cdot; 1/\sqrt{N})) - \beta_N(f)) \\
&\quad = \frac{1}{N^{-1}\sum_{i=1}^{N} \lambda(w_i)} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda(w_i) \int y \sum_{k=0}^{K} c_k (f_k(y \mid w_i; 1/\sqrt{N}) - f_k(y \mid w_i)) dy \\
&\quad = \frac{1}{N^{-1}\sum_{i=1}^{N} \lambda(w_i)} \frac{1}{N} \sum_{i=1}^{N} \lambda(w_i) \sum_{k=0}^{K} c_k \int y s_k(y \mid w_i) f_k(y \mid w_i) dy,
\end{aligned}
$$

which converges to $\sum_{k=0}^{K} c_k E[\lambda(W_i)Y_i(k)s_k(Y_i(k) \mid W_i)]/E[\lambda(W_i)]$ on $\mathcal{E}$ by assumption. We can write this as $\langle \psi, s \rangle$, where

$$\psi(Y_i, D_i, W_i) = \sum_{k=0}^{K} \mathbb{1}\{D_i = k\}\lambda(W_i)c_k \frac{(Y_i - \mu_k(W_i)).}{p_k(W_i)E[\lambda(W_i)]}.$$

Observe that ψ is in the model tangent space, with the summands playing the role of $s_k(y \mid w)$ (more precisely, since ψ is unbounded, it lies in the completion of the tangent space). Hence, the semiparametric efficiency bound is given by $E[\psi^2]$.

A.3 Efficiency of the CW estimator

The next result shows that the estimator in eq. (26) is efficient. We defer its proof to Appendix A.4.

Proposition 3. *Suppose eq. (11) holds in an i.i.d. sample of size N, with known non-degenerate propensity scores $p_k(W_i)$. Let $\beta^*_{\lambda^{\mathrm{CW}},k} = E[\lambda^{\mathrm{CW}}(W_i)\tau_k(W_i)]/E[\lambda^{\mathrm{CW}}(W_i)]$, and $\alpha^*_k = \beta^*_{\lambda^{\mathrm{CW}},k} + E[\lambda^{\mathrm{CW}}(W_i)\mu_0(W_i)]/E[\lambda^{\mathrm{CW}}(W_i)]$. Suppose that the fourth moments of $\lambda^{\mathrm{CW}}(W_i)$ and $\mu(W_i)$ are bounded, and that $p_k \in \mathcal{G}$, $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^{\mathrm{CW}}(W_i)^2}{p_{k'}(W_i)^2} \in \mathcal{G}$, and $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^{\mathrm{CW}}(W_i)}{p_k(W_i)} \in \mathcal{G}$ for all k, k'. Then, provided it is asymptotically linear and regular, $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ achieves the semiparametric efficiency bound for estimating $\beta_{\lambda^{\mathrm{CW}}}$, with diagonal elements of its asymptotic variance of:*

$$\frac{1}{E[\lambda^{\mathrm{CW}}(W_i)]^2} E\left[\frac{\lambda^{\mathrm{CW}}(W_i)^2\sigma_0^2(W_i)}{p_0(W_i)} + \frac{\lambda^{\mathrm{CW}}(W_i)^2\sigma_k^2(W_i)}{p_k(W_i)} + \lambda^{\mathrm{CW}}(W_i)^2(\tau_k(W_i) - \beta^*_{\lambda^{\mathrm{CW}},k})^2\left(\sum_{k'=0}^{K}\frac{\lambda^{\mathrm{CW}}(W_i)^2}{p_k(W_i)^3} - 1\right)\right].$$

This efficiency result doesn't rely on homoskedasticity: under heteroskedasticity, the estimator $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ is still efficient for $\beta_{\lambda^{\mathrm{CW}}}$ (although the weighting $\lambda^{\mathrm{CW}}(W_i)$ need not be optimal under heteroskedasticity). It is stated under the high-level condition that $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ is regular; the proof uses calculations from Newey (1994) to verify the estimator achieves the efficiency bound. Primitive regularity conditions will depend on the form of $\mathcal{G}$ and are omitted for brevity.

Remark 8. The asymptotic variance of the estimator $\hat{\beta}_{\lambda^{\mathrm{CW}}}$ is larger than the asymptotic variance of the infeasible estimator that replaces the estimated weights $\hat{\lambda}^{\mathrm{CW}}(W_i)/\hat{p}_{D_i}(W_i)$ in eq. (26) with the infeasible weights $\lambda^{\mathrm{CW}}(W_i)/p_{D_i}(W_i)$. The latter achieves the asymptotic

variance implied by Corollary 2,

$$\frac{1}{E[\lambda^{\mathrm{CW}}(W_i)]^2} E\left[\frac{\lambda^{\mathrm{CW}}(W_i)^2\sigma_0^2(W_i)}{p_0(W_i)} + \frac{\lambda^{\mathrm{CW}}(W_i)^2\sigma_k^2(W_i)}{p_k(W_i)}\right]. \tag{27}$$

The extra term of the asymptotic variance in Proposition 3 relative to eq. (27) reflects the cost of having to estimate the weights.[36] Analogous term is present in the expression for the asymptotic variance of the one-treatment-at-a-time estimator implementing the weights from Corollary 1.

A.4 Proof of Proposition 3

We first derive the semiparametric efficiency bound for estimating $\beta_{\lambda^{\mathrm{CW}}}$ when the propensity scores are not known, using the same steps, notation, and setup as in the proof of Proposition 1. We then verify that the estimator $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ achieves this bound.

Step 1 Since the propensity scores are not known, the model is now parametrized by $\theta = (f, p)$. Consider one-dimensional submodels of the form $f_k(y \mid w; t) = f_k(y \mid w)(1 + ts_{y,k}(y \mid w))$, and $p_k(w; t) = p_k(w)(1 + ts_{p,k}(x))$, where the functions $s_{y,k}, s_{p,k}$ are bounded and satisfy $\int s_{y,k}(y \mid w) f_k(y \mid w) dy = 0$ and $\sum_{k=0}^K p_k(w) s_{p,k}(w) = 0$ for all $w \in \mathcal{W}$. These conditions ensure that $f_k(y \mid w; t)$ and $p_k(w; t)$ are positive for t small enough and that $\sum_{k=0}^K p_k(w; t) = \sum_{k=0}^K p_k(w) = 1$, so that the submodel is well-defined. The joint log-likelihood, conditional on the controls, is given by

$$\sum_{i=1}^N \sum_{k=0}^K \mathbb{1}\{D_i = k\}(\log f_k(Y_i \mid w_i; t) + \log p_k(w_i; t)).$$

The score at $t = 0$ is given by $\sum_{i=1}^N s(Y_i, D_i \mid w_i)$, with $s(Y_i, D_i \mid w_i) = \sum_{k=0}^K \mathbb{1}\{D_i = k\}(s_{y,k}(Y_i \mid w_i) + s_{p,k}(w_i))$.

In line with this result, we define the tangent space to consist of all functions $s(y, d \mid w) = \sum_{k=0}^K \mathbb{1}\{d = k\}(s_{y,k}(y \mid w) + s_{p,k}(w))$ such that $s_{y,k}$ and $s_{p,k}$ satisfy the above restrictions. Define the inner product on this space by the marginal expectation $\langle s_1, s_2 \rangle = E[s_1(Y_i, D_i \mid W_i) s_2(Y_i, D_i \mid W_i)]$. We will prove the efficiency bound on the event $\mathcal{E}$ that (i) $\frac{1}{N}\sum_{i=1}^N E[s(Y_i, D_i \mid W_i)^2 \mid W_i] \to E[s(Y, D_i \mid W_i)^2]$; (ii) $N^{-1}\sum_i \lambda^{\mathrm{CW}}(W_i) \to E[\lambda^{\mathrm{CW}}(W_i)]$; (iii) $N^{-1}\sum_i \lambda^{\mathrm{CW}}(W_i) \sum_{k=0}^K c_k E[Y_i(k) \cdot s_{y,k}(Y_i \mid W_i) \mid W_i] \to \sum_{k=0}^K c_k E[\lambda^{\mathrm{CW}}(W_i) Y_i(k) \cdot$

[36] The extra term shows this cost is zero if either there is no treatment effect heterogeneity, so that $\tau_k(W_i) = \beta^*_{\lambda^{\mathrm{CW}},k}$, or if the treatment assignment is completely randomized so that $p_k(W_i) = 1/(K+1)$. In the latter case $\lambda^*(W_i) = 1/(K+1)^2$ so $\sum_{k=0}^K \lambda^{\mathrm{CW}}(W_i)^2/p(W_i)^3 = 1$. The extra term can be avoided altogether if we interpret $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ as an estimator of $\beta_{\hat{\lambda}^{\mathrm{CW}}}$. This follows from arguments in Crump et al. (2006, Lemma B.6).

$s_{y,k}(Y_i(k) \mid W_i)]$; (iv) $N^{-1}\sum_{i=1}^N \lambda^{\text{CW}}(W_i)^2 \sum_{k,k'} c_{k'}\mu_{k'}(W_i)\frac{s_{p,k}(W_i)}{p_k(W_i)} \to E[\lambda^{\text{CW}}(W_i)^2 \cdot \sum_{k,k'} c_{k'} \cdot \mu_{k'}(W_i)\frac{s_{p,k}(W_i)}{p_k(W_i)}]$; (v) $N^{-1}\sum_{i=1}^N \lambda^{\text{CW}}(W_i)^2 \sum_{k=0}^K \frac{s_{p,k}(W_i)}{p_k(W_i)} \to E[\lambda^{\text{CW}}(W_i)^2 \sum_{k=0}^K \frac{s_{p,k}(W_i)}{p_k(W_i)}]$; and (vi) $\beta_{\lambda^{\text{CW}}} \to \beta^*_{\lambda^{\text{CW}}}$. Under the proposition assumptions and the law of large numbers, $\mathcal{E}$ is a probability-one set.

Step 2 We verify that the conditions (3.7–3.12) of Theorem 3.1 in McNeney and Wellner (2000) hold on the set $\mathcal{E}$ conditional on the controls, with $\theta_N(s) = (f(\cdot \mid \cdot; 1/\sqrt{N}), p(\cdot; 1/\sqrt{N}))$. Let $\alpha_{Ni} = \prod_{k=0}^K (f_k(Y_i \mid w_i; 1/\sqrt{N})p_k(w_i; 1/\sqrt{N})/f_k(Y_i \mid w_i)p_k(w_i))^{\mathbb{1}\{D_i=k\}} = \prod_{k=0}^K ((1 + N^{-1/2}s_{y,k}(Y_i \mid W_i; N^{-1/2}))(1 + N^{-1/2}s_{p,k}(w_i; 1/\sqrt{N})))^{\mathbb{1}\{D_i=k\}}$ denote the likelihood ratio associated with the ith observation. Since this is bounded by the boundedness of $s_{y,k}, s_{p,k}$, condition (3.7) holds. Also, since $(1+ts_{p,k})^{1/2}$ and $(1+ts_{y,k})^{1/2}$ are continuously differentiable for t small enough, it follows from Lemma 7.6 in van der Vaart (1998) that the quadratic mean differentiability condition (3.8) holds. Since s_k is bounded, the Lindeberg condition (3.9) also holds. Next, $\frac{1}{N}\sum_{i=1}^N E[s(Y_i, D_i \mid W_i)^2 \mid W_i]$ converges to $E[s(Y, D_i \mid W_i)^2] = \langle s, s\rangle$ on $\mathcal{E}$ by assumption. Hence, conditions (3.10) and (3.11) also hold. Since the scores $\Delta_{N,s} = \frac{1}{\sqrt{N}}\sum_{i=1}^N s(Y_i, D_i \mid w_i)$ are exactly linear in s, condition (3.12) also holds. It follows that the model is LAN on $\mathcal{E}$.

Step 3 Write the parameter of interest, $\beta_{\lambda^{\text{CW}}}$, as $\beta_N(\theta) = \sum_{i=1}^N \lambda^{\text{CW}}(w_i) \int y \sum_{k=0}^K c_k f_k(y \mid w_i)dy / \sum_{i=1}^N \lambda^{\text{CW}}(w_i)$, where $\lambda^{\text{CW}}(w_i) = 1/\sum_{k=0}^K p_k(w_i)^{-1}$. Letting $\dot\beta_N(\theta)$ denote the derivative of $\beta_N(\theta(\cdot \mid \cdot; t))$ at $t = 0$, we have

$$\sqrt{N}(\beta_N(\theta(\cdot \mid \cdot; 1/\sqrt{N})) - \beta_N(\theta)) = \dot\beta_N(\theta) + o(1).$$

Let $h(w) = \lambda^{\text{CW}}(w)\sum_{k=0}^K c_k \int y s_{y,k}(y \mid w) f_k(y \mid w)dy$, and $\tilde h(W_i) = \sum_{k'=0}^K c_{k'}\mu_{k'}(W_i) - \beta^*_{\lambda^{\text{CW}}}$. The derivative may then be written as

$$\begin{aligned}
\dot\beta_N(\theta) &= \frac{1}{\sum_{i=1}^N \lambda^{\text{CW}}(w_i)} \sum_{i=1}^N \left(h(w_i) + \lambda^{\text{CW}}(w_i)^2 \sum_{k=0}^K \frac{s_{p,k}(w_i)}{p_k(w_i)} \left(\sum_{k'=0}^K c_{k'}\mu_{k'}(w_i) - \beta_N(\theta) \right) \right) \\
&\to \frac{1}{E[\lambda_i^{\text{CW}}]} E\left[h(W_i) + (\lambda_i^{\text{CW}})^2 \sum_{k=0}^K \frac{s_{p,k}(W_i)}{p_k(W_i)} \left(\sum_{k'=0}^K c_{k'}\mu_{k'}(W_i) - \beta^*_{\lambda^{\text{CW}}} \right) \right] \\
&= \frac{1}{E[\lambda_i^{\text{CW}}]} E\left[\lambda_i^{\text{CW}} \sum_{k=0}^K X_{ki} \left(c_k \frac{Y_i - \mu_k(W_i)}{p_k(W_i)} + \frac{\lambda_i^{\text{CW}} \tilde h(W_i)}{p_k(W_i)^2} \right) s(Y_i, D_i \mid W_i) \right],
\end{aligned}$$

where $\lambda_i^{\text{CW}} = \lambda^{\text{CW}}(W_i)$, the limit on the second line holds on the event $\mathcal{E}$, and the third line uses $E[X_{ki}(Y_i - \mu_k(W_i))s(Y_i, D_i \mid W_i) \mid W_i] = p_k(W_i)E[Y_i(k)s_{y,k}(Y_i(k) \mid W_i) \mid W_i]$ and $E[X_{ki}s(Y_i, D_i \mid W_i) \mid W_i] = p_k(W_i)s_{p,k}(W_i)$. Since for any function $a(W_i)$, $E[a(W_i)s(Y_i, D_i \mid$

$W_i)] = 0$, subtracting $\frac{1}{E[\lambda_i^{\text{CW}}]} \sum_{k=0}^{K} E[(\lambda_i^{\text{CW}})^2 \frac{\tilde{h}(W_i)}{p_k(W_i)} s(Y_i, D_i \mid W_i)] = 0$ from the preceding display implies $\sqrt{N}(\beta_N(\theta(\cdot \mid \cdot; 1/\sqrt{N})) - \beta_N(\theta)) = E[\psi(Y_i, D_i, W_i) s(Y_i, D_i \mid W_i)] + o(1)$, where

$$\psi(Y_i, D_i, W_i) = \sum_{k=0}^{K} X_{ki} \cdot \left(\frac{\lambda_i^{\text{CW}}}{E[\lambda_i^{\text{CW}}]} c_k \frac{Y_i - \mu_k(W_i)}{p_k(W_i)} + \frac{\lambda_i^{\text{CW}}}{E[\lambda_i^{\text{CW}}]} \tilde{h}(W_i) \left(\frac{\lambda_i^{\text{CW}}}{p_k^2} - 1 \right) \right).$$

Observe that ψ lies in the completion of the tangent space, with the expression in parentheses playing the role of $s_{y,k}(Y_i \mid W_i) + s_{p,k}(W_i)$. Hence, the semiparametric efficiency bound is given by $E[\psi^2]$, which yields the expression in the statement of the Proposition.

Attainment of the bound We derive the result in two steps. First, we show that

$$\sqrt{N}(\beta_{\lambda^{\text{CW}}} - \beta^*_{\lambda^{\text{CW}}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi^*(W_i) + o_p(1) \text{ where } \psi^*(W_i) = \frac{\lambda_i^{\text{CW}}}{E[\lambda_i^{\text{CW}}]} (\tau(W_i) - \beta^*_{\lambda^{\text{CW}}}). \quad (28)$$

Second, we show that

$$\sqrt{N}(\hat{\beta}_{\hat{\lambda}^{\text{CW}}} - \beta^*_{\lambda^{\text{CW}}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, D_i, W_i) + o_p(1), \quad (29)$$

where, letting $\epsilon_{ki} = Y_i - \mu_k(W_i)$,

$$\psi_k(Y_i, D_i, W_i) = \frac{\lambda_i^{\text{CW}}}{E[\lambda_i^{\text{CW}}]} \left(\frac{X_{ki} \epsilon_{ki}}{p_k(W_i)} - \frac{X_{0i} \epsilon_{0i}}{p_k(W_i)} + (\tau_k(W_i) - \beta^*_{\lambda^{\text{CW}},k}) \lambda_i^{\text{CW}} \sum_{k'} \frac{X_{k'i}}{p_{k'}(W_i)^2} \right).$$

Together, these results imply that the asymptotic variance of $\hat{\beta}_{\hat{\lambda}^{\text{CW}}}$ as an estimator of $\beta_{\lambda^{\text{CW}}}$ is given by $\text{var}(\psi - \psi^*)$, which coincides with the semiparametric efficiency bound.

Equation (28) follows directly under the assumptions of the proposition by the law of large numbers and the fact that the variance of $\lambda_i^{\text{CW}}(\tau(W_i) - \beta^*_{\lambda^{\text{CW}}})$ is bounded. To show eq. (29), write $\hat{\beta}_{\hat{\lambda}^{\text{CW}},k} = \hat{\alpha}_k - \hat{\alpha}_0$, where $\hat{\alpha}$ is a two-step method of moments estimator based on the $(K+1)$ dimensional moment condition $E[m(Y_i, D_i, W_i, \alpha^*, p)] = 0$ with elements $m_k(Y_i, D_i, W_i, \alpha^*, p) = \lambda_i^{\text{CW}} \frac{X_{ki}}{p_k(W_i)} (Y_i - \alpha_k^*)$, and α^* is a $(K+1)$ dimensional vector with elements $\alpha_k^* = E[\lambda_i^{\text{CW}} \mu_k(W_i)] / E[\lambda_i^{\text{CW}}]$.

Consider a one-dimensional path F_t such that the distribution of the data is given by F_0. Let $p_{k,t}(W_i) = E_{F_t}[X_{ki} \mid W_i]$ denote the propensity score along this path. The derivative of

$E[m_k(Y_i, D_i, W_i, \alpha^*, p_t)]$ with respect to t evaluated at $t = 0$ is

$$E\left[\frac{\lambda_i^{\mathrm{CW}} X_{ki}}{p_k(W_i)}(Y_i - \alpha_k^*)\left(\lambda_i^{\mathrm{CW}} \sum_{k'=0}^{K} \frac{\dot{p}_{k'}(W_i)}{p_{k'}(W_i)^2} - \frac{\dot{p}_k(W_i)}{p_k(W_i)}\right)\right] = \sum_{k'=0}^{K} E[\delta_{kk'}(W_i)'\dot{p}_{k'}(W_i)],$$

where $\dot{p}_k$ denotes the derivative of $p_{k,t}$ at $t = 0$, and

$$\delta_{k,k'}(W_i) = \lambda_i^{\mathrm{CW}}(\mu_k(W_i) - \alpha_k^*)\left(\frac{\lambda_i^{\mathrm{CW}}}{p_{k'}(W_i)^2} - \frac{\mathbb{1}\{k = k'\}}{p_k(W_i)}\right).$$

Under the assumptions of the proposition, $\delta_{k,k'} \in \mathcal{G}$. It therefore follows by Proposition 4 in Newey (1994) that the influence function for $\hat{\alpha}_k$ is given by

$$\frac{1}{E[\lambda_i^{\mathrm{CW}}]}\left(\frac{\lambda_i^{\mathrm{CW}} X_{ki}}{p_k(W_i)}(Y_i - \alpha_k^*) + \sum_{k'} \delta_{kk'}(W_i)(X_{k'i} - p_{k'}(W_i))\right)$$
$$= \frac{\lambda_i^{\mathrm{CW}}}{E[\lambda_i^{\mathrm{CW}}]}\left(\frac{X_{ki}\epsilon_{ki}}{p_k(W_i)} + (\mu_k(W_i) - \alpha_k^*)\lambda_i^{\mathrm{CW}} \sum_{k'} \frac{X_{k'i}}{p_{k'}(W_i)^2}\right),$$

which yields eq. (29).

Appendix B Connections to the DiD Literature

In this appendix we elaborate on the connections between Proposition 1 and the recent literature studying potential biases from heterogeneous treatment effects in DiD regressions and related specifications (e.g. Goodman-Bacon, 2021; Sun & Abraham, 2021; Hull, 2018b; de Chaisemartin & D'Haultfœuille, 2020, 2022; Callaway & Sant'Anna, 2021; Borusyak et al., 2022; Wooldridge, 2021). We first show how our framework fits a TWFE regression with a general treatment specification. We then show how Proposition 1 applies to three particular specifications: a static binary treatment, a dynamic "event study" treatment, and a static multivalued treatment (or "movers regression"). In each case we discuss whether there is a potential for bias—either contamination bias or own-treatment negative weighting—and give a numerical illustration.

Consider a panel of units indexed by $j = 1, \ldots, n$ which are observed over time periods $t = 1, \ldots, T$. For simplicity, we assume the panel is balanced such that the sample size is $N = nT$. For an observation $i = (j, t)$, let $J_i = j$ and $T_i = t$ denote the corresponding unit and time period, respectively. In a TWFE specification, the controls only comprise these two variables, $W_i = (J_i, T_i)$, and they enter the control function as dummies, $g(W_i) = \alpha + (\mathbb{1}\{J_i = 2\}, \ldots, \mathbb{1}\{J_i = n\}, \mathbb{1}\{T_i = 2\}, \ldots, \mathbb{1}\{T_i = T\})'\gamma$, with the indicators $\mathbb{1}\{J_i = 1\}$ and $\mathbb{1}\{T_i = 1\}$

omitted to avoid perfect collinearity.

To study these specifications, we follow de Chaisemartin and D'Haultfœuille (2020) and Borusyak et al. (2022) in considering the n observed units as fixed, and we condition on their treatment status (results when the units are sampled from a large population are analogous). For each unit j, we observe a random T-vector of outcomes $Y_j = (Y_{j1}, \ldots, Y_{jT})$ and a fixed T-vector of $(\mathcal{K}+1)$-valued treatments $\mathcal{D}_j = (\mathcal{D}_{j1}, \ldots, \mathcal{D}_{jT})$. These treatments are used to construct a vector of $(K+1)$-valued "treatments states" $D_j = (D_{j1}, \ldots, D_{jT})$, with $D_{jt} \in \{0, \ldots, K\}$. Setting $D_j = \mathcal{D}_j$ covers scenarios with static treatments; as we show below, other choices of D_j allows us to cover scenarios with dynamic treatment effects. As in the main text, X_{jt} denotes a K-vector of treatment status indicators derived from D_{jt}.

We make two assumptions. First, we assume that potential outcomes $Y_{jt}(d_t)$ depend on the T-vector of treatments only through the current value d_t of the treatment state, such that $Y_{jt} = Y_{jt}(D_{jt})$.[37] Second, we make a parallel trends assumption by writing the untreated potential outcomes as

$$Y_{jt}(0) = \alpha_j + \lambda_t + \eta_{jt},$$

for fixed α_j and λ_t, and assuming

$$E[\eta_{jt}] = 0. \tag{30}$$

Together these expressions imply $E[Y_{jt}(0)] = \alpha_j + \lambda_t$, which is how parallel trends is sometimes formalized (c.f. Assumption 1 in Borusyak et al. (2022); weaker versions of the parallel trends assumption yield analogous results). We do not restrict the dependence of η_{jt} across units or time, nor do we make restrictions on the potentially random treatment effects $\tau_{jt,k} = Y_{jt}(k) - Y_{jt}(0)$. Collecting these effects in a vector τ_{jt}, we have

$$Y_{jt} = X_{jt}'\tau_{jt} + \alpha_j + \lambda_t + \eta_{jt}. \tag{31}$$

This outcome model reduces to a conventional TWFE model under the assumption of constant treatment effects: $\tau_{jt} = \beta$ for all (j, t).

Since the only source of randomness are the shocks η_{jt} and the treatment effects τ_{jt}, this setup fits into the framework of Section 3 if we interpret the expectation in eq. (8) as averaging over the joint distribution of $\{\tau_{jt}, \eta_{jt}\}_{j=1,t=1}^{n,T}$. Specifically, (β, g) are the minimizers of $N^{-1}\sum_{j=1}^{n}\sum_{t=1}^{T} E_{\tau,\eta}[(Y_{jt} - X_{jt}'\tilde{\beta} - \tilde{g}(W_{jt}))^2]$, where the subscript on the expectation makes explicit that we only integrate over the joint distribution of $\{\tau_{jt}, \eta_{jt}\}_{j=1,t=1}^{n,T}$. The parallel trends assumption implies $\mu_0(W_i) = \alpha_{J_i} + \lambda_{T_i}$, so that eq. (13) in Assumption 2 holds. In other words, the parallel trend assumption implies that our controls $g(W_i)$ correctly specify

[37] This assumption rules out misspecification of the treatment states, such as when there are dynamic effects but $D_{jt} = \mathcal{D}_{jt}$ only indexes contemporaneous treatment status, as noted in footnote 10.

the untreated potential outcome mean. Additionally, Assumption 1 holds trivially because the treatment vector is non-random.

To make the link to Proposition 1, note that $\tilde{X}_{jt} = X_{jt} - \bar{X}_j - \bar{X}_t + \bar{X}$ coincides with the sample residual from regressing X_i onto unit and time effects. Here $\bar{X}_j = \frac{1}{T}\sum_{t=1}^{T} X_{jt}$, $\bar{X}_t = \frac{1}{n}\sum_{j=1}^{n} X_{jt}$, and $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} \bar{X}_j$. We may then write eq. (10) as

$$\begin{aligned}\beta &= \left(\sum_{j=1}^{n}\sum_{t=1}^{T} E_{\tau,\eta}[\tilde{X}_{jt}\tilde{X}_{jt}']\right)^{-1} \sum_{j=1}^{n}\sum_{t=1}^{T} E_{\tau,\eta}[\tilde{X}_{jt}Y_{jt}] \\ &= \left(\sum_{j=1}^{n}\sum_{t=1}^{T} \tilde{X}_{jt}\tilde{X}_{jt}'\right)^{-1} \sum_{j=1}^{n}\sum_{t=1}^{T} \tilde{X}_{jt}X_{jt}'E[\tau_{jt}],\end{aligned} \tag{32}$$

where the second equality uses eqs. (30) and (31), and the fact that only η_{jt} and τ_{jt} are stochastic. Proposition 1 implies that the coefficient on the kth element on X_{jt} is given by

$$\beta_k = \sum_{j,t} \lambda_{kk}(j,t)E[\tau_{jt,k}] + \sum_{\ell\neq k}\sum_{j,t} \lambda_{k\ell}(j,t)E[\tau_{jt,\ell}] \tag{33}$$

where

$$\lambda_{kk}(j,t) = \frac{\ddot{\tilde{X}}_{jt,k}X_{jt,k}}{\sum_{j,t}\ddot{\tilde{X}}_{jt,k}^2}, \qquad \text{and} \qquad \lambda_{k\ell}(j,t) = \frac{\ddot{\tilde{X}}_{jt,k}X_{jt,\ell}}{\sum_{j,t}\ddot{\tilde{X}}_{jt,k}^2},$$

and $\ddot{\tilde{X}}_{jt,k}$ is the sample residual from regressing $\tilde{X}_{jt,k}$ onto the remaining elements of $\tilde{X}_{jt}$. Recall that since we do not assume that eq. (12) holds, it is not guaranteed that $\lambda_{kk}(j,t) \geq 0$.

To unpack this result, we now consider four special cases from the literature.

Static binary treatment Consider a DiD setting where units adopt (and potentially drop) a binary treatment at different time periods—as studied by de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021). For example, different states j may choose to roll out a policy in different years and a researcher wishes to estimate the average effect of this policy using this staggered adoption. We assume that the treatment is static, setting $D_{jt} = \mathcal{D}_{jt}$, with $K = \mathcal{K} = 1$. Since the treatment is binary, $X_{jt} = D_{jt}$ is a scalar with $\ddot{\tilde{X}}_{jt,1} = \tilde{X}_{jt}$, and the second term in eq. (33) drops; the weights on the first term simplify to

$$\lambda_{11}(j,t) = \frac{\tilde{X}_{jt}X_{jt}}{\sum_{j',t'}\tilde{X}_{j't'}^2} = \frac{(1-\overline{X}_j-\overline{X}_t+\overline{X})X_{jt}}{\sum_{j',t'}\tilde{X}_{j't'}^2},$$

which coincides with the expression in Theorem 1 of de Chaisemartin and D'Haultfœuille (2020). These treatment weights are not guaranteed to be convex since eq. (12) does not

hold.[38] In contrast, Athey and Imbens (2022) consider staggered DiD regressions where eq. (12) holds because intervention timing is assumed to be random (in place of the parallel trends assumption). Under this design-based assumption, Proposition 1 shows the treatment weights (corresponding to those in Theorem 1(iv) of Athey and Imbens (2022)) are convex.

The above expression for λ_{11} yields a simple necessary and sufficient condition for convex weights, which is that for units j that are treated in period t, $1 - \overline{X}_j - \overline{X}_t + \overline{X} \geq 0$. In staggered adoption designs, $\overline{X}_t$ is increasing with t. Thus, in staggered adoption designs, it suffices to check this condition for $t = T$, and for unit j that adopts the treatment first—that is, to check whether

$$1 - \max_j \overline{X}_j - \overline{X}_T + \overline{X} \geq 0. \tag{34}$$

Condition (34) holds in the canonical DiD case with a single intervention date, where the first $n_1 < n$ units treated in the last $T_1 < T$ periods and untreated in the earlier periods $1, \ldots, T - T_1$. The remaining units are never treated, so that $D_{jt} = \mathcal{D}_{jt} = \mathbb{1}\{j \leq n_1, t \geq T - T_1\}$. This nests the simplest DiD specification where $T = 2$ and $T_1 = 1$ (e.g. Card & Krueger, 1994). In this case, when units in the treatment group are treated, $1 - \overline{X}_j - \overline{X}_t + \overline{X} = (1 - n_1/n)(1 - T_1/T)$ so that the weights $\lambda_{11}(j,t)$ are non-negative, and eq. (33) simplifies to:

$$\beta_1 = \sum_{j,t} \lambda_{11}(j,t) E[\tau_{jt,1}], \qquad \lambda_{11}(j,t) = \frac{(1 - \frac{n_1}{n})(1 - \frac{T_1}{T}) X_{jt}}{(1 - \frac{n_1}{n})(1 - \frac{T_1}{T}) \frac{n_1 T_1}{nT}} = \frac{X_{jt}}{n_1 T_1 / N},$$

which is simply the average treatment effect for the $n_1 T_1$ treated observations.

However, in presence of multiple treatment adoption dates, eq. (34) may fail. To illustrate, consider a case with three time periods ($T = 3$) and three groups of units: $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{N}$, with respective sizes n_E, n_L, and n_N. Units $j \in \mathcal{E}$ are "early adopters", and are treated beginning in period 2. Units $j \in \mathcal{L}$ are "late adopters", and are treated only in period 3. Units in the last group are never treated.[39] In this case, eq. (34) simplifies to $1 - 2/3 - (n_E + n_L)/n + (2/3 n_E + 1/3 n_L)/n = (n_N - n_L)/3n$, which is negative if there are more late adopters than never adopters; otherwise, if $n_L < n_N$, all weights are positive. Indeed, some algebra shows

$$\begin{aligned}
\lambda_{11}(j,3) &= \frac{n_E + 2n_N}{\kappa} \qquad j \in \mathcal{L}, \\
\lambda_{11}(j,2) &= \frac{n_N + 2n_L}{\kappa} \qquad j \in \mathcal{E}, \\
\lambda_{11}(j,3) &= \frac{n_N - n_L}{\kappa} \qquad j \in \mathcal{E},
\end{aligned}$$

[38] Since $E[X_{jt} \mid W_{jt}] = X_{jt} \in \{0,1\}$, if eq. (12) held, then the residual $\tilde{X}_{jt}$ must be zero (this is true if, e.g., all units have the same treatment adoption date). But that would generate a multicollinearity issue, precluding the researcher from including unit and time effects in the regression.

[39] This example is a special case of the example discussed in Figure 2 of Goodman-Bacon (2021).

where $\kappa = 2(n_E n_L + n_E n_N + n_N n_L)$ and $\lambda_{11}(j,t) = 0$ otherwise.

Condition (34) is generally quite restrictive. Consider, for instance, a setting in which no units are treated in the first period and a fraction $1/T$ of observations adopts the treatment in period $t = 2, \ldots, T$. Then for the group adopting treatment in period 2, eq. (34) becomes $(3 - T)/2T$, which is negative if $T \geq 4$. Similarly, condition (34) fails if there exists an always-treated group, or if everyone is treated in the last period.

Dynamic binary treatment with staggered adoption Next, consider an "event study" setting in which each unit j starts being treated in period $A(j) \in \{1, 2, \ldots, T\} \cup \infty$ and remains treated thereafter, with $A(j) = \infty$ denoting a unit that is never treated. Thus, $\mathcal{D}_{jt} = \mathbb{1}\{t > A(j)\}$, with $\mathcal{K} = 1$. Unlike in previous cases, we allow for dynamic effects by letting $D_{jt} = t - A(j)$ index the number of periods since the treatment adoption date (breaking with our usual indexing convention of $D_{jt} \geq 0$), assuming no anticipation effect one period before adoption, and correspondingly normalizing $D_{jt} = -1$ for the never-treated group. X_{jt} then consists of indicators for all leads and lags relative to the adoption date: $X_{jt} = (\mathbb{1}\{D_{jt} = -(T-1)\}, \ldots, \mathbb{1}\{D_{jt} = -2\}, \mathbb{1}\{D_{jt} = 0\}, \ldots, \mathbb{1}\{D_{jt} = T-1\})'$, with the indicator for the period just prior to adoption ($D_{jt} = -1$) excluded. This specification avoids perfect collinearity when all treatment adoption dates are represented in the data (including the never-treated group). Sun and Abraham (2021) and Borusyak et al. (2022) study such "fully-dynamic" event study specifications.

Since X_{jt} is now a vector with $K = 2(T-1)$, the second contamination bias term in eq. (33) will generally be present. As such, Sun and Abraham (2021) and Borusyak et al. (2022) study the potential for contamination across estimates of post- and pre-treatment effects (with the latter used in conventional pre-trend specification tests). Furthermore, like in the previous case with static treatment, the own-treatment weights in the first term are potentially negative. While random treatment timing assumptions may solve the issue of negative own treatment weights, contamination bias remains a concern even under such assumptions.

To illustrate the potential for contamination bias, consider again the example with early, late, and never adopters and $T = 3$, except we now allow the treatment effect to be dynamic. Let $\tau_{jts} = Y_{jt}(s) - Y_{jt}(-\infty)$, $s \in \{-2, 1, 0, 1\}$ denote the effect on unit j in time period t of adopting the treatment s periods ago. If s is negative, we interpret this as the anticipation effect of adopting the treatment $-s$ periods from now. Under our assumptions $\tau_{jt,-1} = 0$, such that there is no anticipation effect immediately before treatment adoption. To test whether the two-period-ahead anticipation effect is zero, and whether the effect of the treatment fades out over time, we let $X_{jt} = (\mathbb{1}\{D_{jt} = -2\}, \mathbb{1}\{D_{jt} = 0\}, \mathbb{1}\{D_{jt} = T-1\})'$. Thus, for instance, $X_{j1} = (1, 0, 0)'$ for late adopters while $X_{j2} = (0, 1, 0)'$ for early adopters. Let $\tau_{E,ts} = n_E^{-1} \sum_{j \in \mathcal{E}} E[\tau_{jts}]$ denote the average effect among early adopters, and define $\tau_{L,ts}$

similarly. Then some rather tedious algebra shows that

$$\beta = \begin{pmatrix} \tau_{L,1,-2} \\ 0 \\ \tau_{E,3,1} \end{pmatrix} + \lambda_{E,0}\tau_{E,2,0} + \lambda_{L,0}\tau_{L,3,0},$$

where

$$\lambda_{E,0} = \frac{1}{\zeta}\begin{pmatrix} 3n_L n_E + n_N n_E \\ 3n_L n_E + 2n_N n_E \\ -n_L n_N \end{pmatrix}, \qquad \lambda_{L,0} = \frac{1}{\zeta}\begin{pmatrix} -3n_L n_E - n_N n_E \\ 3n_E n_L + 2n_N n_L \\ n_N n_L \end{pmatrix},$$

and $\zeta = 2(3n_L n_E + n_E n_N + n_L n_N)$. In other words, the estimand for the two-period-ahead anticipation effect β_1 equals the anticipation effect for late adopters in period 1 (this is the only group we ever observe two periods before treatment) plus a contamination bias term coming from the effect of the treatment on impact. Similarly, the estimand for the effect of the treatment one period since adoption, β_3, equals the effect for early adopters in period 3 (this is the only group we ever observe one period after treatment) plus a contamination bias term coming from the effect of the treatment on impact. The estimand for the effect of the treatment upon adoption, β_0, has no contamination bias, and equals a weighted average of the effect for early and late adopters. In this example, the own treatment weights are always positive, but the contamination weights can be large. For instance, with equal-sized groups, $\lambda_{E,0} = (2/5, 1/2, -1/10)'$ and $\lambda_{L,0} = (-2/5, 1/2, 1/10)'$, so the contamination weights in the estimand β_1 are almost as large as the own treatment weights for β_2.

It is worth noting that if all treated units share a single adoption date then contamination bias disappears and a TWFE regression recovers a vector of average dynamic treatment effects for the treated, in analogy to the static case discussed above. To show this result, let us set $A(j) = T_1$ for the first n_1 units, with $A(j) = \infty$ for the remaining $n_0 = n - n_1$ units. Excluding the indicator just prior to the adoption date, as well as leads and lags that are always zero for all units, the treatment vector has length $T-1$: $X_{jt} = (\mathbb{1}\{D_{jt} = -(T_1 - 1)\}, \ldots, \mathbb{1}\{D_{jt} = -2\}, \mathbb{1}\{D_{jt} = 0\}, \ldots, \mathbb{1}\{D_{jt} = T - T_1\})$. For the control units, this vector is always zero. For the adopters, $X_{jt} = e_t$ (the tth unit vector) if $t \leq T_1 - 2$, $X_{j,T-1}$ is zero, and $X_{jt} = e_{t-1}$ for $t \geq T_1$. We may write this compactly as $X_{jt} = e_t\,\mathbb{1}\{t < T_1 - 1\} + e_{t-1}\,\mathbb{1}\{t \geq T_1\}$ for $j \leq n_1$. Partialling out the unit and time effects therefore yields

$$\tilde{X}_{jt} = (\mathbb{1}\{j \leq n_1\} - n_1/n)(e_t\,\mathbb{1}\{t < T_1 - 1\} + e_{t-1}\,\mathbb{1}\{t \geq T_1\} - \iota_{T-1}/T),$$

where ι_{T-1} is a $T-1$ vector of ones. Hence, $\sum_{j=1}^{n}\sum_{t=1}^{n}\tilde{X}_{jt}\tilde{X}_{jt}' = \frac{n_1 n_0}{n}\left(I_{T-1} - \iota_{T-1}\iota_{T-1}'/T\right)$.

By the Woodbury identity, we therefore obtain

$$\Lambda(j,t) = \Big(\sum_{j=1}^{n}\sum_{t=1}^{n} \tilde{X}_{jt}\tilde{X}_{jt}'\Big)^{-1}\tilde{X}_{jt}X_{jt}' = \frac{n}{n_1 n_0}(I_{T-1} + \iota_{T-1}\iota_{T-1}')\tilde{X}_{jt}X_{jt}'$$
$$= \frac{1}{n_1}(I_{T-1} + \iota_{T-1}\iota_{T-1}')(X_{jt} - \iota_{T-1}/T)X_{jt}' = \frac{1}{n_1}X_{jt}X_{jt}'.$$

Hence, by eq. (32), TWFE regression identifies the average treatment for the treated, $\beta = \frac{1}{n_1}\sum_{j=1}^{n_1}(\tau_{j1,-(T-1)}, \ldots, \tau_{j,T_1-2,-2}, \tau_{jT_1,1}, \ldots, \tau_{jT,T-T_1})$. Intuitively, since the contamination weights sum to zero and there is only one group of adopters, the contamination weights must be identically zero.

Mover regressions: multiple treatments with multiple transitions. Finally, consider a "mover regression" in a setting with a static multivalued treatment $\mathcal{D}_{jt} \in \{0, \ldots, K\}$ with multiple transitions of units between treatment states, leading to multiple treatment paths. We focus on the static treatment case, setting $D_{jt} = \mathcal{D}_{jt}$. This setting has been studied by Hull (2018b) and de Chaisemartin and D'Haultfœuille (2022). Our Proposition 1 shows that such specifications can suffer from two distinct sources of bias: own-treatment negative weighting from multiple transitions and contamination bias from the multiple treatments. As before the former bias disappears under random treatment timing (as in Athey and Imbens (2022)), or other assumptions which make eq. (12) hold.

To illustrate this case, consider a setting with $T = 3$ periods, $K = 3$ treatments, and three groups of units, $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{N}$. Units in the first group start out untreated, move to treatment 2 in period 1, and move to treatment 3 in period 3. Units in the second group start in treatment 1, move to being untreated in period 2, and move to treatment 2 in period 3. Units in group $\mathcal{N}$ are never treated. This example is isomorphic to the previous event study example, in that it leads to the same regression specification and the same eq. (33) characterization of regression coefficients. Thus, there are no negative own-treatment weights in this example, but there are potentially large contamination weights depending on the relative group sizes.

Appendix C Details on the Further Applications

This appendix details our procedure for selecting the additional empirical examples in Section 5.2. We also discuss the implementation details and provide the full set of results.

C.1 Article Search Protocol

We scraped the American Economic Association (AEA) website for a list of all published articles across all AEA journals over 2013–2022. This search included all articles from the following journals: *American Economic Review*, *American Economic Review: Insights*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Microeconomics*, *Journal of Economic Literature* (excluding articles with "review" in the title and articles labeled as Front Matter, Doctoral Dissertations, and Annotated Listings), *Journal of Economic Perspectives*, and *AER/AEA Papers and Proceedings* (excluding articles with "report" or "minutes" in the title). We limited this search to articles with online replication packages which include at least one data file.[40]

We next filtered articles by two keyword searches of titles, abstracts, and main texts:

- Experiments (keywords: stratified, random, RCT, experiment).
- Racial disparities (keywords: racial/ethnic differences, discrimination, disparities, gaps).

These searches yielded a total of 1,848 experiments and 67 observational studies on race. To further narrow down experiments, we restricted attention to papers where one of the keywords appears in the paper's title, abstract, or associated tweet.

For each search, we then manually reviewed papers in reverse citation order (as measured by Google Scholar) keeping those which include in the main text a linear regression of some outcome on multiple treatments or race indicators and controls. We ignored specifications where a single treatment or race indicator is interacted with some set of fixed effects or controls, such as event study specifications. We stopped the review when five papers were identified with such a specification, or when we exhausted all papers in the search.

C.2 Overlap Sample and Propensity Score Variation

For each main specification, we identify a subset of the analysis sample with full treatment overlap using the following procedure. First, we define a primary strata variable (when not otherwise obvious from the paper) as the discrete variable with the greatest number of unique levels. In the experimental applications this is always the randomization strata; in the observational applications this is the "finest" fixed effect. We then drop observations for the levels of this variable which do not exhibit all levels of the treatment. Finally, in the remaining sample, we drop any additional controls which have no within-treatment variation.

[40] Here "data files" refers to those with any of the following extensions: Stata (`'dta'`), Excel (`'xls'` or `'xlsx'`), Matlab (`'mat'`), R (`'rdata'`, `'rda'`, `'rds'`), HDFS (`'h5'`, `'hdf5'`), Apache (`'parquet'`, `'arrow'`), SAS (`'sas7bdat'`), and delimited files (`'csv'`, `'tsv'`).

	Wald			LM		
	Statistic	(d.f.)	p-value	Statistic	(d.f.)	p-value
Project STAR	308.9	(154)	0.000	302.9	(154)	0.000
Benhassine et al.	207.2	(159)	0.006	217.2	(194)	0.121
Cole et al.	22.7	(39)	0.983	70.3	(54)	0.067
de Mel et al.	0.9	(392)	1.000	1.1	(392)	1.000
Drexler et al.	12.4	(14)	0.574	12.6	(14)	0.555
Duflo et al.	109.6	(254)	1.000	94.5	(258)	1.000
Fryer and Levitt	3947.6	(630)	0.000	4164.0	(681)	0.000
Rim et al.	1403.5	(88)	0.000	233.0	(234)	0.506
Weisburst	2350.0	(69)	0.000	223.2	(48)	0.000

Notes: This table summarizes Wald and Lagrange multiplier tests of the null hypothesis that the coefficients on the controls in a multinomial logit regression of the treatment on the controls all equal zero. The tests allow for clustering in Benhassine et al., Duflo et al., Rim et al., and Weisburst, and for heteroskedasticity in the remaining applications.

Table C.1: Tests of Propensity Score Variation

We check for meaningful propensity score variation in each specification with two tests, summarized in Table C.1. Specifically, we compute the Wald and LM tests of the null hypothesis that, in a multinomial logit regression of the treatment on the controls, all coefficients on the controls equal zero. The table gives evidence for statistically significant propensity score variation (at 10% level) in the Project STAR application, two of the additional experimental applications (Cole et al. and Benhassine et al.), and all three observational studies.

C.3 Full Results

In Table C.2-C.9, we report the estimated effects for each additional application. Panel A of each table first reports the $\hat{\beta}$ estimates from the multiple-treatment regression as reported in the original paper and corresponding standard errors. We also report the own-treatment effect component from the decomposition in eq. (23) along with three alternative estimators: the ATE estimator, the easiest-to-estimate weighted ATE estimator (EW) and the common-weight (CW) estimator. Panel B reports the difference between $\hat{\beta}$ and these 4 alternative estimators. The $\hat{\beta}$, EW and CW estimators are consistent even without overlap. However, if overlap fails in the full sample, the own-treatment effect component from the decomposition in eq. (23) may not be identified for all treatments, and the ATE is not identified. If identification of the decomposition fails for the full treatment vector, we subset to the overlap sample, as described in Appendix C.2 above, and report the full set of estimates from the different estimators.

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
LCT to fathers	0.074			0.089	0.056	0.067	0.084	0.078	0.076	0.061
	(0.016)			(0.017)	(0.018)	(0.019)	(0.024)	(0.015)	(0.020)	(0.020)
				[0.012]	[0.011]			[0.014]	[0.014]	[0.012]
LCT to mothers	0.078			0.067	0.071	0.081	0.075	0.079	0.074	0.068
	(0.014)			(0.013)	(0.017)	(0.017)	(0.017)	(0.014)	(0.015)	(0.017)
				[0.009]	[0.011]			[0.012]	[0.011]	[0.012]
CCTs to fathers	0.055			0.062	0.041	0.047	0.038	0.033	0.039	0.038
	(0.014)			(0.013)	(0.018)	(0.016)	(0.015)	(0.014)	(0.016)	(0.017)
				[0.009]	[0.012]			[0.012]	[0.012]	[0.012]
CCTs to mothers	0.053			0.045	0.040	0.039	0.033	0.042	0.041	0.040
	(0.013)			(0.013)	(0.018)	(0.017)	(0.016)	(0.015)	(0.017)	(0.018)
				[0.011]	[0.013]			[0.014]	[0.013]	[0.013]
Number of controls	57					26				
Sample size	11,074					6,996				
B. Bias										
LCT to fathers				−0.016	0.018		−0.018	−0.011	−0.009	0.006
				(0.010)	(0.018)		(0.015)	(0.016)	(0.010)	(0.019)
LCT to mothers				0.012	0.007		0.007	0.002	0.007	0.014
				(0.009)	(0.016)		(0.013)	(0.011)	(0.010)	(0.015)
CCTs to fathers				−0.007	0.014		0.009	0.013	0.007	0.009
				(0.005)	(0.015)		(0.009)	(0.010)	(0.006)	(0.015)
CCTs to mothers				0.008	0.013		0.006	−0.003	−0.002	−0.001
				(0.007)	(0.015)		(0.009)	(0.009)	(0.006)	(0.015)

Notes: This table reports estimates from the Benhassine et al. application, as described in Appendix C.3. The regression specification comes from column 1 of Table 5 in Benhassine et al. (2015). Standard errors clustered by school sector are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.2: Full results: Benhassine et al. (2015)

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
Muslim only	0.160			0.095	0.033	0.001	0.038	−0.012	−0.036	0.010
	(0.086)			(0.086)	(0.094)	(0.111)	(0.138)	(0.109)	(0.120)	(0.093)
				[0.079]	[0.098]			[0.109]	[0.121]	[0.104]
Hindu only	0.121			0.058	0.062	0.006	0.075	0.080	0.060	0.076
	(0.089)			(0.088)	(0.101)	(0.116)	(0.123)	(0.106)	(0.116)	(0.096)
				[0.062]	[0.100]			[0.097]	[0.080]	[0.092]
Group only	0.239			0.229	0.103	0.107	0.140	0.158	0.093	0.071
	(0.097)			(0.098)	(0.112)	(0.115)	(0.130)	(0.086)	(0.106)	(0.108)
				[0.076]	[0.097]			[0.082]	[0.099]	[0.091]
Muslim & Group	0.169			0.092	−0.094	−0.109	−0.075	−0.096	−0.075	−0.088
	(0.087)			(0.083)	(0.079)	(0.082)	(0.074)	(0.080)	(0.070)	(0.075)
				[0.038]	[0.076]			[0.078]	[0.062]	[0.072]
Hindu & Group	0.018			−0.052	−0.027	−0.004	0.000	−0.034	0.000	−0.021
	(0.080)			(0.075)	(0.096)	(0.094)	(0.093)	(0.094)	(0.087)	(0.094)
				[0.056]	[0.089]			[0.090]	[0.075]	[0.086]
Number of controls	13					3				
Sample size	132					73				
B. Bias										
Muslim only				0.065	0.127		−0.037	0.014	0.038	−0.009
				(0.044)	(0.073)		(0.066)	(0.060)	(0.061)	(0.061)
Hindu only				0.063	0.059		−0.069	−0.075	−0.054	−0.071
				(0.050)	(0.083)		(0.044)	(0.085)	(0.041)	(0.081)
Group only				0.010	0.136		−0.033	−0.050	0.014	0.036
				(0.060)	(0.103)		(0.060)	(0.081)	(0.064)	(0.102)
Muslim & Group				0.077	0.263		−0.033	−0.013	−0.033	−0.021
				(0.056)	(0.091)		(0.048)	(0.063)	(0.047)	(0.060)
Hindu & Group				0.071	0.046		−0.004	0.030	−0.004	0.016
				(0.048)	(0.080)		(0.028)	(0.056)	(0.036)	(0.061)

Notes: This table reports estimates from the Cole et a. application, as described in Appendix C.3. The regression specification comes from column 6 of Table 7 in Cole et al. (2013). Robust standard errors are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.3: Full results: Cole et al. (2013)

	Full sample				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW
Info and Reimburse	−0.010	−0.010	−0.010	−0.010	−0.010
	(0.023)	(0.014)	(0.007)	(0.012)	(0.007)
			[0.000]	[0.000]	[0.000]
Rs 10,000	0.134	0.134	0.135	0.134	0.135
	(0.034)	(0.032)	(0.017)	(0.027)	(0.017)
			[0.000]	[0.000]	[0.000]
Rs 20,000	0.105	0.105	0.104	0.105	0.104
	(0.035)	(0.030)	(0.017)	(0.026)	(0.017)
			[0.008]	[0.009]	[0.007]
Rs 40,000	0.273	0.273	0.269	0.272	0.270
	(0.041)	(0.038)	(0.020)	(0.033)	(0.020)
			[0.000]	[0.000]	[0.000]
Number of controls	98				
Sample size	520				
B. Bias					
Info and Reimburse		−0.001	−0.001	−0.001	0.000
		(0.022)	(0.022)	(0.020)	(0.022)
Rs 10,000		0.000	−0.001	0.000	−0.001
		(0.019)	(0.029)	(0.020)	(0.029)
Rs 20,000		0.000	0.000	0.000	0.000
		(0.021)	(0.030)	(0.023)	(0.030)
Rs 40,000		0.000	0.004	0.001	0.003
		(0.019)	(0.035)	(0.024)	(0.035)

Notes: This table reports all results from the de Mel et al. (2013) application, as described in Appendix C.3. The regression specification comes from column 2 of Table 2 in de Mel et al. (2013). Robust standard errors are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.4: Full results: de Mel et al. (2013)

	Full sample				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW
Standard Accounting	0.036	0.038	0.040	0.037	0.040
	(0.041)	(0.041)	(0.040)	(0.041)	(0.040)
			[0.040]	[0.040]	[0.040]
Rule-of-Thumb	0.109	0.114	0.113	0.112	0.113
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
			[0.039]	[0.039]	[0.039]
Number of controls	7				
Sample size	796				
B. Bias					
Standard Accounting		−0.002	−0.004	−0.001	−0.004
		(0.004)	(0.005)	(0.003)	(0.005)
Rule-of-Thumb		−0.005	−0.004	−0.004	−0.004
		(0.004)	(0.005)	(0.003)	(0.005)

Notes: This table reports estimates from the Drexler et al. (2014) application, as described in Appendix C.3. The regression specification comes from row 2 of Table 2 in Drexler et al. (2014). Robust standard errors are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.5: Full results: Drexler et al. (2014)

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
Educ. subsity	−0.031			−0.036	−0.029	−0.024	−0.029	−0.025	−0.032	−0.027
	(0.012)			(0.011)	(0.011)	(0.013)	(0.012)	(0.007)	(0.011)	(0.010)
				[0.000]	[0.000]			[0.001]	[0.001]	[0.001]
HIV education	0.003			0.009	0.002	0.000	0.005	0.003	0.005	0.000
	(0.011)			(0.009)	(0.012)	(0.011)	(0.010)	(0.007)	(0.010)	(0.011)
				[0.000]	[0.001]			[0.001]	[0.001]	[0.001]
Both	−0.016			−0.019	−0.020	−0.012	−0.010	−0.007	−0.009	−0.012
	(0.012)			(0.010)	(0.011)	(0.012)	(0.010)	(0.007)	(0.010)	(0.010)
				[0.000]	[0.000]			[0.001]	[0.001]	[0.001]
Number of controls	86					79				
Sample size	9,116					8,664				
B. Bias										
Educ. subsity				0.005	−0.002		0.005	0.001	0.008	0.003
				(0.008)	(0.012)		(0.008)	(0.011)	(0.007)	(0.011)
HIV education				−0.006	0.001		−0.005	−0.003	−0.006	0.000
				(0.007)	(0.011)		(0.008)	(0.010)	(0.007)	(0.011)
Both				0.003	0.004		−0.002	−0.005	−0.003	0.000
				(0.008)	(0.013)		(0.008)	(0.011)	(0.008)	(0.012)

Notes: This table reports estimates from the Duflo et al. (2015) application, as described in Appendix C.3. The regression specification comes from column 1 of Table 2, panel A in Duflo et al. (2015). Standard errors clustered by school reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.6: Full results: Duflo et al. (2015)

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
Black	−0.213	−0.182		−0.193	−0.202	−0.191	−0.150	−0.231	−0.171	−0.195
	(0.032)	(0.035)		(0.034)	(0.065)	(0.037)	(0.041)	(0.038)	(0.040)	(0.059)
				[0.031]	[0.045]			[0.037]	[0.035]	[0.043]
Hispanic	−0.249			−0.257	−0.171	−0.209	−0.212	−0.196	−0.220	−0.171
	(0.028)			(0.030)	(0.046)	(0.032)	(0.035)	(0.033)	(0.034)	(0.045)
				[0.028]	[0.039]			[0.033]	[0.031]	[0.039]
Asian	−0.294			−0.324	−0.330	−0.275	−0.276	−0.150	−0.283	−0.317
	(0.035)			(0.038)	(0.085)	(0.039)	(0.043)	(0.058)	(0.043)	(0.082)
				[0.033]	[0.057]			[0.056]	[0.036]	[0.055]
Other	−0.132			−0.116	−0.127	−0.127	−0.104	−0.084	−0.105	−0.105
	(0.038)			(0.039)	(0.046)	(0.043)	(0.045)	(0.035)	(0.044)	(0.047)
				[0.029]	[0.035]			[0.034]	[0.031]	[0.035]
Number of controls	176					127				
Sample size	8,806					6,623				
B. Bias										
Black		−0.031		−0.020	−0.011		−0.042	0.040	−0.020	0.004
		(0.016)		(0.013)	(0.056)		(0.017)	(0.028)	(0.014)	(0.048)
Hispanic				0.008	−0.077		0.003	−0.013	0.011	−0.038
				(0.009)	(0.038)		(0.013)	(0.021)	(0.011)	(0.035)
Asian				0.030	0.036		0.001	−0.124	0.009	0.043
				(0.018)	(0.074)		(0.018)	(0.057)	(0.016)	(0.068)
Other				−0.015	−0.005		−0.023	−0.043	−0.023	−0.022
				(0.013)	(0.048)		(0.015)	(0.038)	(0.014)	(0.049)

Notes: This table reports estimates from the Fryer and Levitt (2013) application, as described in Appendix C.3. The regression specification comes from column 4 of Table 3 in Fryer and Levitt (2013). Robust standard errors are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.7: Full results: Fryer and Levitt (2013)

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
Black	−4.059			−3.907	−3.786	−4.441	−3.969	8.071	−3.199	−3.266
	(1.107)			(1.210)	(1.597)	(1.149)	(1.059)	(11.922)	(1.039)	(1.403)
				[0.393]	[0.747]			[3.991]	[0.537]	[0.628]
Hispanic	−1.119			−0.837	1.290	−0.658	−0.908	2.927	−0.879	−1.099
	(0.731)			(0.698)	(3.949)	(1.603)	(1.461)	(3.403)	(1.446)	(2.460)
				[0.142]	[0.637]			[2.150]	[0.305]	[0.620]
Asian	−2.536			−2.117	−4.375	−3.383	−3.110	−8.439	−3.633	−3.685
	(0.978)			(1.206)	(2.896)	(1.440)	(1.114)	(3.606)	(0.930)	(1.824)
				[0.314]	[0.384]			[1.685]	[0.351]	[0.638]
Number of controls	268					35				
Sample size	4,037					620				
B. Bias										
Black				−0.152	−0.274		−0.472	−12.513	−1.243	−1.175
				(0.406)	(1.902)		(1.117)	(12.089)	(1.277)	(1.210)
Hispanic				−0.282	−2.409		0.250	−3.584	0.222	0.442
				(0.212)	(3.813)		(0.446)	(3.269)	(0.344)	(1.154)
Asian				−0.418	1.839		−0.273	5.056	0.249	0.302
				(0.632)	(2.804)		(0.713)	(3.259)	(0.842)	(1.445)

Notes: This table reports estimates from the Rim et al. (2020) application, as described in Appendix C.3. The regression specification comes from column 3 of Table 2 in Rim et al. (2020). Standard errors clustered by cohort are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.8: Full results: Rim et al. (2020)

	Full sample				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW
Black	0.172	−0.037	0.342	0.109	0.246
	(0.274)	(0.305)	(0.396)	(0.267)	(0.292)
			[0.323]	[0.152]	[0.178]
Hispanic	0.043	−0.754	−0.330	−0.496	−0.466
	(0.394)	(0.404)	(0.395)	(0.341)	(0.289)
			[0.312]	[0.221]	[0.169]
Other	1.130	1.130	0.223	1.244	0.106
	(0.652)	(0.654)	(0.622)	(0.679)	(0.712)
			[0.394]	[0.347]	[0.566]
Number of controls	256				
Sample size	7,488				
B. Bias					
Black		0.209	−0.169	0.063	−0.074
		(0.218)	(0.337)	(0.190)	(0.264)
Hispanic		0.797	0.373	0.539	0.508
		(0.356)	(0.390)	(0.310)	(0.330)
Other		0.001	0.907	−0.113	1.025
		(0.125)	(0.340)	(0.120)	(0.578)

Notes: This table reports all results from the Weisburst (2019) application, as described in Appendix C.3. The regression specification comes from Table 2, panel A in Weisburst (2019). Standard errors clustered by police beat are reported in parentheses. Standard errors that assume the propensity scores are known are reported in square brackets.

Table C.9: Full results: Weisburst (2019)

Appendix D Additional Figures

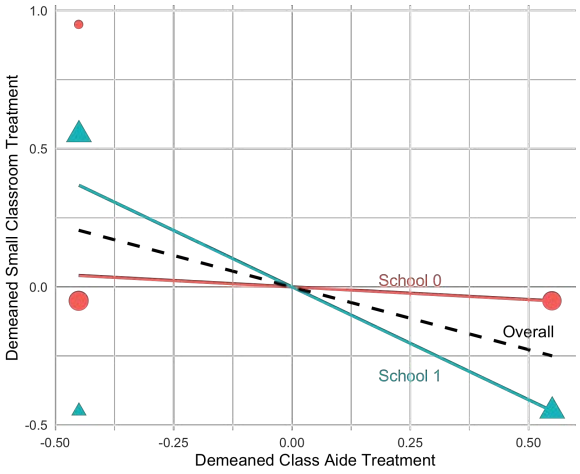

Figure D.1: Regression of Small Classroom Treatment on Class Aide Treatment

Note: This figure plots values of the demeaned class aide treatment ($\tilde{X}_{2i}$, the x-axis) against values of the demeaned small classroom treatment ($\tilde{X}_{1i}$, the y-axis) in our numerical example from Section 2.3. The size of the points corresponds to the density of observations. The solid red and blue lines mark the within-school regression of the two residualized treatments, while the dashed black line is the overall regression line. The residuals from this line give $\ddot{\tilde{X}}_{i1}$.

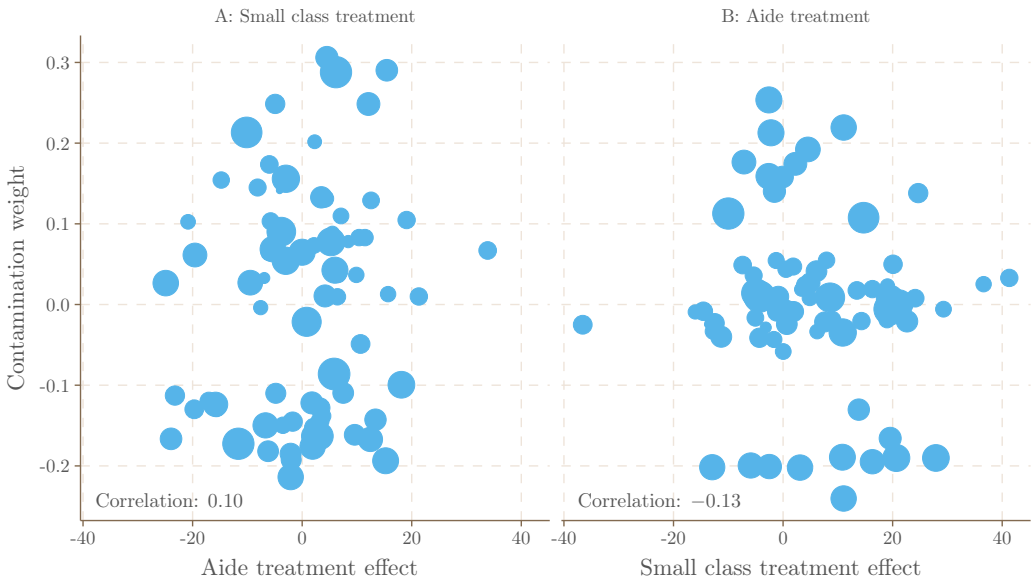

Figure D.2: Project STAR contamination weights.

Notes: This figure shows correlations between estimated school-specific treatment effects and contamination weights. Panel A depicts the correlation between the estimated teaching aide treatment effects by school against the estimated contamination weight for the small class estimate. Panel B gives the correlation between the estimated small class treatment effects by school against the estimated contamination weight for the teaching aide estimate. Correlations are reported on each panel. The size of the points is proportional to the number of students enrolled in each school.

Figure D.3: Project STAR treatment weights

Notes: This figure shows correlations between estimated school-specific treatment effects and the weights used by different estimators. Panel A gives the correlations for the small class treatment, and Panel B gives them for the teaching aide treatment. The first row plots the own treatment weights from the contamination bias decomposition in eq. (23). The second row gives plots the EW scheme from Corollary 1, and the third row gives the CW scheme from Corollary 2. Correlations are reported on each panel. The size of the points is proportional to the number of students enrolled in each school.