CONTAMINATION BIAS IN LINEAR REGRESSIONS

Paul Goldsmith-Pinkham
Peter Hull
Michal Kolesár

WORKING PAPER 30108

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 August 2022

We thank Jason Abaluck, Isaiah Andrews, Josh Angrist, Tim Armstrong, Kirill Borusyak, Kyle Butts, Clément de Chaisemartin, Peng Ding, Jin Hahn, Xavier D'Haultfoeuille, Simon Lee, Bernard Salanié, Pedro Sant'Anna, Tymon Słoczyński, Jonathan Roth, Jacob Wallace, 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, 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 August 2022
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 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 a new estimator of efficiently weighted average effects. We find minimal bias in a re-analysis of Project STAR, due to idiosyncratic effect heterogeneity. But sizeable contamination bias arises when effect heterogeneity becomes correlated with treatment propensity scores.

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-multe/

1 Introduction

Consider a linear regression of an outcome Y_i on a vector of mutually exclusive treatment indicators 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, lagged achievement, or other controls which yield a credible selection-on-observables assumption. The regression might also be the first stage of an instrumental variables (IV) regression, perhaps leveraging the as-good-as-random assignment of multiple decision-makers (e.g. bail judges or benefit administrators) indicated in X_i, 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 identify convex weighted averages of heterogeneous treatment 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 weighted 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. Despite a set of treatments being completely randomly assigned within groups, as in a stratified multi-armed RCT, a regression on treatment and strata indicators generally fails to yield causally interpretable regression coefficients. Instead, regression estimates of each treatment's effect are generally contaminated by a non-convex average of the effects of other treatments: the regression coefficient for a given RCT treatment arm generally incorporates the effects of *all* arms.

We first derive a general characterization of this "contamination bias" in multiple-treatment regressions. To separate the problem from the well-known 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. We show that the regression coefficient on each treatment identifies a convex weighted average of its causal effects, plus a contamination

[1]Prominent RCT examples 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 (Hull, 2018a; Abaluck et al., 2021; Geruso et al., 2020). Prominent "judge IV" examples include Kling (2006), Maestas et al. (2013), and Dobbie and Song (2015).

bias term that is generally non-zero. The bias term is given by a linear combination of the causal effects of other treatments, with weights that sum to zero. As a result, each treatment effect estimate will generally incorporate the effects of other treatments, unless the effects are uncorrelated with the contamination weights. Since the contamination 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, consider the most flexible specification of controls as a set of 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.

Contamination bias also arises under an alternative "model-based" identifying assumption that, instead of making assumptions on the treatment's "design" (i.e. propensity scores), assumes the regression parametrization of covariates is flexible enough to include the conditional mean of the potential outcome under no treatment, $Y_i(0)$. In a linear model with two-way unit and time fixed effects, this reduces to the parallel trends restriction 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.[2] We show that replacing the restriction on propensity scores with an assumption on $Y_i(0)$ generates an additional issue: the own-treatment effect weights are no longer guaranteed to be positive. 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.[3] Negative weighting arises because regressions leveraging model-based restrictions on $Y_i(0)$ are generally not robust to treatment effect heterogeneity. Contamination bias arises because linear regression fails to account for

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

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

the non-linear dependence across multiple dependent treatments and controls.

Even more broadly, contamination bias can arise in descriptive regressions which seek to estimate averages of certain conditional group contrasts without assuming a causal framework—as in studies of treatment or outcome disparities across multiple racial or ethnic groups, studies of regional variation in healthcare utilization or outcomes, or studies of industry wage gaps.[4] Our analysis shows that in such regressions the coefficient on a given group or region averages the conditional contrasts across all other groups or regions, with non-convex weights.

Our bias characterization also has implications for IV regressions leveraging multiple correlated instruments, such as indicators for as good as randomly assigned judges. Contamination bias in the first-stage regression of treatment on multiple instruments and flexible controls (e.g. courtroom fixed effects) can generate violations of the effective first-stage monotonicity restriction, even when conventional first-stage monotonicity is satisfied unconditionally. We show how this problem is distinct from previous concerns over the monotonicity assumption in judge IV designs (Mueller-Smith, 2015; Frandsen et al., 2019; Norris, 2019; Mogstad et al., 2021) and over insufficient flexibility in the control parametrization (Blandhol et al., 2022).[5]

We then discuss three solutions to the contamination bias problem, and their trade-offs, in the baseline case of conditionally ignorable treatments. One conceptually principled solution is to adapt approaches to estimating the average treatment effect (ATE) of a conditionally ignorable binary treatment (see Imbens & Wooldridge, 2009, for a review) to the multiple treatment case (e.g. Cattaneo, 2010; Chernozhukov et al., 2021; Graham & Pinto, 2022). For example, one could run an expanded regression that includes interactions between the treatments and demeaned controls.[6] Such ATE estimators achieve the semiparametric efficiency bound under an assumption of strong overlap of the covariate distribution for units in each treatment arm. But this approach may be infeasible or yield imprecise estimates under limited overlap—a common scenario in practice (Crump et al., 2009).

This practical consideration motivates an alternative solution: 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 that they minimize a semiparametric efficiency bound under homoskedasticity. These optimal weights are convex and 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 optimal weights for a given treatment-control contrast are similarly convex and given by a linear regression which restricts estimation to

[4]Prominent examples of such analyses respectively include Fryer and Levitt (2013), Skinner (2011), and Krueger and Summers (1988).

[5]The contamination bias issue is also distinct from the Freedman (2008a, 2008b) critique of regression to analyze randomized trials, which concerns estimation, not identification.

[6]In the judge IV case, the analogous solution interacts judge indicators with courtroom fixed effects.

the individuals who are either in the control group or the treatment group of interest. For estimating effects that are directly comparable across multiple treatments, our optimality characterization leads to a new estimator of convex average effects. We give guidance for how applied researchers can gauge the extent of contamination bias in practice and apply the different solutions with help from a new Stata package, `multe`.[7]

We illustrate the contamination bias problem and solutions in an application to the Project STAR trial, which randomized students within schools to either a small classroom treatment, a teaching aide treatment, or control conditions. We find the potential for sizeable bias in estimates of both treatment effects, from regressions with school fixed effects, due to significant treatment effect heterogeneity. Nevertheless, we show that the actual contamination is likely to be minimal because the effect heterogeneity turns out to be largely uncorrelated with the contamination weights. The application thus highlights the importance of testing the empirical relevance of theoretical concerns with how regression combines heterogeneous effects.

We structure the rest of the paper as follows. Section 2 illustrates contamination bias in a simple example with two mutually exclusive treatment indicators and one binary control. Section 3 characterizes the general problem in regressions with multiple treatments and flexible controls, and discusses connections to previous analyses. Section 4 discusses the robustness and efficiency properties of three solutions. Section 5 gives guidance for measuring and avoiding contamination bias in practice, and illustrates these tools in the Project STAR experiment. Section 6 concludes. All proofs and extensions are given in Appendix A.

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 in our application in Section 5.2.

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 a constant:

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

[7] The Stata package is available at https://github.com/gphk-metrics/stata-multe.

By definition, U_i is a mean-zero regression residual that is uncorrelated with D_i and W_i. Krueger (1999), for example, primarily studied the effect of small class size D_i on the test scores Y_i of middle school students indexed by i. Project STAR randomized students to classes within schools with at least three classes per grade. The number of students assigned to each intervention thus varied both by the number of students in a school and the relative classroom size. To account for this non-random treatment variation, Krueger (1999) followed earlier analyses of Project STAR in estimating regressions with school (and sometimes school-by-period) 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 regression 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 $\tau_i = Y_i(1) - Y_i(0)$, and we can write realized achievement as $Y_i = Y_i(0) + \tau_i D_i$. To formalize the random assignment of treatment within schools, we assume that D_i is conditionally independent of potential outcomes given the control W_i:

$$(Y_i(0), Y_i(1)) \perp D_i \mid W_i. \tag{2}$$

Angrist (1998) showed that regression coefficients like β identify a weighted average of within-strata ATEs, with convex weights.[8] In our stylized Project STAR regression, this result shows that:

$$\beta = \phi\tau(0) + (1-\phi)\tau(1), \quad \text{where} \quad \phi = \frac{\text{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1}\text{var}(D_i \mid W_i = w)\Pr(W_i = w)} \in [0,1] \tag{3}$$

gives a convex weighting scheme, and $\tau(w) = E[Y_i(1) - Y_i(0) \mid W_i = w]$ 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 $\tau(w)$ across the two schools.

Equation (3) 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:

$$\beta = \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{E[\tilde{D}_i Y_i(0)]}{E[\tilde{D}_i^2]} + \frac{E[\tilde{D}_i D_i \tau_i]}{E[\tilde{D}_i^2]}, \tag{4}$$

[8] See Słoczyński (2022) for an alternative representation of this estimand, in terms of conditional average effects on the treated and untreated, under slightly different assumptions.

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,

$$E[\tilde{D}_i Y_i(0)] = E[E[\tilde{D}_i Y_i(0) \mid W_i]] = E[E[\tilde{D}_i \mid W_i] E[Y_i(0) \mid W_i]] = 0. \tag{5}$$

The first equality in eq. (5) 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. (4) is zero. Analogous arguments show that

$$E[\tilde{D}_i D_i \tau_i] = E[E[\tilde{D}_i D_i \tau_i \mid W_i]] = E[E[\tilde{D}_i D_i \mid W_i] E[\tau_i \mid W_i]] = E[\text{var}(D_i \mid W_i)\tau(W_i)],$$

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. (4) as

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

proving the representation of β in eq. (3).

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. (4) even without this property. We next show, however, that the remaining steps in the derivation of eq. (3) 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 bias term in the expression for the regression coefficient.

2.2 Contamination Bias with Two Randomized Treatments

In reality, as noted above, Project STAR randomized two mutually exclusive interventions within schools: a reduction in class size ($D_i = 1$) and the introduction of 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 the first element $X_{i1} = \mathbb{1}\{D_i = 1\}$ indicates assignment to a small class and the second element $X_{i2} = \mathbb{1}\{D_i = 2\}$ indicates assignment to a class with a full-time aide. We continue

to include a constant and the school indicator W_i as controls, yielding the regression

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma W_i + U_i. \tag{6}$$

To account for the second treatment, 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

$$\beta_1 = \frac{E[\tilde{\tilde{X}}_{i1} Y_i]}{E[\tilde{\tilde{X}}_{i1}^2]} = \frac{E[\tilde{\tilde{X}}_{i1} Y_i(0)]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i1} \tau_{i1}]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}]}{E[\tilde{\tilde{X}}_{i1}^2]}, \tag{7}$$

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 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. (7) 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. (5), since the auxilliary regression residuals are still uncorrelated with any individual characteristic like $Y_i(0)$.[9] In this sense, the regression does not suffer from OVB. However, we do not generally have $E[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}] = 0$. Instead, simplifying eq. (7) by the same steps as before leads to the expression

$$\beta_1 = E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)] \tag{8}$$

as a generalization of eq. (3). 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. (3). Thus, if not for the second term in eq. (8), β_1 would similarly identify a convex average of the conditional

[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. (5), are both uncorrelated with $Y_i(0)$. It follows that $E[\tilde{\tilde{X}}_{i1} Y_i(0)] = 0$.

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. (8) 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 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, 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 a classroom with a full-time aide 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, such that 45 percent of students were assigned to a small classroom, 45 percent were assigned to a classroom with a full-time aide, and only 10 percent were assigned to the control group. Therefore, $p_1(0) = 0.05$, $p_2(0) = 0.45$, 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 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, while in school 1 it is -0.8.[10] The overall regression of $\tilde{X}_{i1}$ on $\tilde{X}_{i2}$ averages over these two correlations, leading to a misspecified residual $\tilde{\tilde{X}}_{i1}$ that is correlated with X_{i2} within each school. Figure 1 illustrates this averaging by plotting the different potential pairs of the two demeaned treatments $(\tilde{X}_{i1}, \tilde{X}_{i2})$, with the two school strata in different colors and shapes. The figure shows how within the first school, the value of the demeaned class

[10] Here the conditional correlation is $\text{corr}(\tilde{X}_{i1}, \tilde{X}_{i12} \mid W_i) = -\sqrt{p_1(W_i)/(1-p_1(W_i))}\sqrt{p_2(W_i)/(1-p_2(W_i))}$.

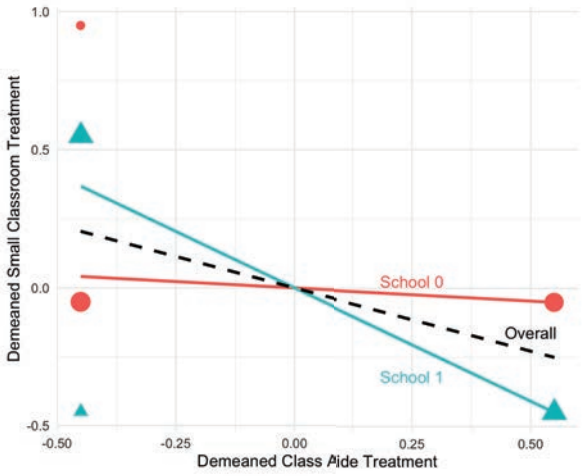


Figure 1: Regression of Small Classroom Treatment on Class Aide Treatment

Notes: 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. 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 $\tilde{\tilde{X}}_{i1}$.

aide treatment is only weakly predictive of the small classroom treatment, but it is highly predictive in the second school. The overall regression line in black averages over these two relationships, yielding residuals which 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 the 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 the school 0, the aide has no effect, $\tau_2(0) = 0$. Equation (8) then shows that 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 sizeable 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 made arbitrarily large or small (and positive or negative), depending on the heterogeneity of the teaching aide effects across schools.

To build further intuition for eq. (8), it is useful to consider two cases where the contamination bias term is zero. First, suppose the average effects of the teaching aide treatment are constant across the two schools: $\tau_2(0) = \tau_2(1) \equiv \tau_2$. Since regression residuals are by

construction uncorrelated with the included regressors, $E[\lambda_{12}(W_i)] = E[\tilde{\tilde{X}}_{i1}X_{i2}]/E[\tilde{\tilde{X}}_{i1}^2] = 0$. Thus, the contamination weights on the second treatment effects average to zero, and the contamination bias disappears: $E[\lambda_{12}(W_i)\tau_2(W_i)] = E[\lambda_{12}(W_i)]\tau_2 = 0$. More generally, the contamination bias will be small when the variation in average teacher's aide treatment effects across schools $\tau_2(W_i)$ is small, or when this treatment effect heterogeneity is only weakly correlated with the contamination weights across schools.

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. (6) can therefore be shown to be equivalent to the previous eq. (3), identifying the same convex average of $\tau_1(w)$. This case highlights that dependence across treatments is necessary for the contamination bias to arise.

Before proceeding to a general characterization of contamination bias, we note that the above intuition about the non-linear conditional expectation $E[X_{i1} \mid W_i, X_{i2}]$ also suggests a simple solution to the problem. By including interactions of W_i and X_{i2} in eq. (6), the auxiliary regression of X_{i1} on the other regressors will be saturated and thus will capture the inherent nonlinearity in $E[X_{i1} \mid W_i, X_{i2}]$. We show below how such interacted regressions can obviate contamination bias. In particular, we show how a particular interacted regression specification gives an efficient estimator of (unweighted) ATEs that is immune to the bias of the simpler specification. We then propose a new class of estimators which—as with the Angrist (1998) result for binary treatments—identify a convex average of conditional ATEs. These estimators may yield smaller standard errors, while still being free from bias.

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 treatment vector in Appendix A.1.

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

$$Y_i = X_i'\beta + g(W_i) + U_i, \tag{9}$$

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

$$(\beta, g) = \underset{\tilde{\beta}\in\mathbb{R}^K, \tilde{g}\in\mathcal{G}}{\operatorname{argmin}} \; E[(Y_i - X_i'\tilde{\beta} - \tilde{g}(W_i))^2] \tag{10}$$

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. (9) 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 and are useful for developing our characterization of contamination bias below:

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, $A(j) \in \{1, \ldots, T\}$.

Example 1 nests the motivating RCT example in Section 2, allowing for an arbitrary number of experimental strata in W_i and random 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 effect 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.[11]

As a first step towards characterizing the β treatment coefficient vector, we solve the minimization problem in eq. (10). 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

$$\beta = E[\tilde{X}_i\tilde{X}_i']^{-1}E[\tilde{X}_iY_i]. \tag{11}$$

[11] 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. (10). For event studies this means we assume some units are never treated, with $A(j) = \infty$. See Roth et al. (2022) for a recent review of the literature.

A further application of the FWL theorem allows us to write each treatment coefficient as

$$\beta_k = \frac{E[\tilde{\tilde{X}}_{ik} Y_i]}{E[\tilde{\tilde{X}}_{ik}^2]},$$

where $\tilde{\tilde{X}}_{ik}$ is the residual from regressing $\tilde{X}_{ik}$ on $\tilde{X}_{i,-k} = (\tilde{X}_{i1}, \ldots, \tilde{X}_{i,k-1}, \tilde{X}_{i,k+1}, \ldots, \tilde{X}_{iK})'$.

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:

$$(Y_i(0), \ldots, Y_i(K)) \perp D_i \mid W_i. \tag{12}$$

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*

$$p_k \in \mathcal{G} \tag{13}$$

for all k, or

$$\mu_0 \in \mathcal{G}. \tag{14}$$

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. (13) 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. (13) fails because the propensity scores are binary—they cannot be captured by a linear combination of the two-way fixed effects. However, eq. (14) can still be 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 and assumption in Appendix B.[12]

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. (11) 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. (13), $E[\tilde{X}_i \mid W_i] = 0$, 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. (14) 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).

[12] Identification based on eq. (13) can be seen as "design-based" in that it leverages only the conditional random assignment of D_i and specifies the treatment assignment process. Identification based on eq. (14) 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.

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 we still have $E[\tilde{X}_i Y_i(0)] = 0$ under Assumptions 1 and 2, it follows from eq. (11) that

$$\beta = \frac{E[\tilde{X}_i X_i \tau_i]}{E[\tilde{X}_i^2]} = E[\lambda_{11}(W_i)\tau(W_i)], \quad \text{with} \quad \lambda_{11}(W_i) = \frac{E[\tilde{X}_i X_i \mid W_i]}{E[\tilde{X}_i X_i]}, \tag{15}$$

where the second equality uses iterated expectations and the fact that $E[\tilde{X}_i^2] = E[\tilde{X}_i X_i]$. Under eq. (13), $E[\tilde{X}_i X_i \mid W_i] = E[\tilde{X}_i^2 \mid W_i] = \text{var}(X_i \mid W_i)$, so the weights further simplify 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). The result provides a rationale for estimating the effect of a scalar as good as randomly assigned treatment using a partially linear model: so long as the specification of $\mathcal{G}$ is rich enough so that eq. (13) holds, this model will identify a convex average of heterogeneous treatment effects. Moreover, as we will show in Section 4, the weights $\lambda_{11}(W_i)$ are efficient in that they minimize the semiparametric efficiency bound (conditional on the controls) for estimating some weighted-average treatment effect. This result makes the partially linear specification (9) especially appealing with a single binary treatment. On the other hand, when eq. (14) holds but eq. (13) does not, the weights $\lambda_{11}(W_i)$ need not be positive. We return to this point in Section 3.2.

The next 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 the partially linear model (9) identify*

$$\beta_k = E[\lambda_{kk}(W_i)\tau_k(W_i)] + \sum_{\ell \neq k} E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)], \tag{16}$$

where

$$\lambda_{kk}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik} X_{ik} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]} \qquad \text{and} \qquad \lambda_{k\ell}(W_i) = \frac{E[\tilde{\tilde{X}}_{ik} X_{i\ell} \mid W_i]}{E[\tilde{\tilde{X}}_{ik}^2]}.$$

These weights satisfy $E[\lambda_{kk}(W_i)] = 1$ and $E[\lambda_{k\ell}(W_i)] = 0$. Furthermore, if eq. (13) holds, $\lambda_{kk}(W_i) \geq 0$ for each k.

Proposition 1 shows that the coefficient on X_{ik} in eq. (9) is a sum of two terms. The first term is a weighted average of conditional ATEs $\tau_k(W_i)$, with weights $\lambda_{kk}(W_i)$ that average to one and are guaranteed to be convex when eq. (13) holds. This term generalizes the characterization of the single-treatment case, eq. (15). The second term is a weighted average of treatment effects for *other* treatments $\tau_\ell(W_i)$, with weights $\lambda_{k\ell}(W_i)$ that average to zero.

Because these contamination weights are zero on average, they must be negative for some values of the controls unless they are all identically zero.[13]

Each treatment coefficient β_k thus generally suffers from contamination bias. Two exceptions are when $\lambda_{k\ell}(W_i) = 0$ almost-surely for all $\ell \neq k$, and when the conditional effects of these other treatments are homogeneous such that $\tau_\ell(W_i) = \tau_\ell$. In the second case $E[\lambda_{k\ell}(W_i)\tau_\ell(W_i)] = \tau_\ell E[\lambda_{k\ell}(W_i)] = 0$, so there is no contamination bias term. By the law of iterated expectations the first case holds if $E[\tilde{\tilde{X}}_{ik} \mid X_{i,-k}, W_i] = 0$, or, equivalently, if the conditional expectation of X_{ik} given $X_{i,-k}$ and W_i is partially linear (i.e. $E[X_{ik} \mid X_{i,-k}, W_i] = X'_{i,-k}\alpha + g_k(W_i)$ for some vector α and $g_k \in \mathcal{G}$). In other words, it holds when the assignment of treatment k depends linearly on the other treatment indicators and a flexible function of the controls. This condition is the analog of condition (13) 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. However, with mutually exclusive treatments, it cannot hold 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, we have $\alpha_\ell = -g_k(W_i)$ for all elements α_ℓ of α. This in turn implies the assignment of treatment k doesn't depend on W_i, which can't be the case unless the propensity score $p_k(W_i)$ is constant.

A third weaker case of no contamination bias is when the weights $\lambda_{k\ell}(W_i)$ and conditional ATEs $\tau_\ell(W_i)$ vary, but are uncorrelated with each other. More generally, contamination bias will tend to be small when the contamination weights $\lambda_{k\ell}(W_i)$ and the conditional ATEs $\tau_\ell(W_i)$ are only weakly correlated: that is, when the factors influencing treatment effect heterogeneity are largely unrelated to the factors influencing the treatment assignment process. We return to this possibility in our empirical application (Section 5.2).

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

Remark 1. Since the weights in eq. (16) are functions of the variances $E[\tilde{\tilde{X}}^2_{ik}]$ 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 $\tau_\ell(W_i)$. We illustrate such an approach in our empirical application.

Remark 2. The results in Proposition 1 are stated for the case when X_i are mutually exclusive 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-

[13] 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 contrast, our analysis interprets regression *estimands* in terms of weighted averages of conditional ATEs under a broad class of identifying assumptions.

treatment weights $\lambda_{kk}(W_i)$ may be negative even if eq. (13) holds.

Remark 3. 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. (9) and replace condition (14) 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 potential for negative weighting of a treatment's *own* effects when condition (14) holds but condition (13) fails. This negative weighting issue is relevant in various DiD regressions and related estimators which rely on such models for untreated potential outcomes (via parallel trends assumptions) while conditioning on treatment assignment. Although the recent DiD literature focuses on two-way fixed effect 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 eq. (13) holds.[14]

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 thus relevant whenever one uses an additive covariate adjustment, regardless of 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).[15]

[14] More broadly, negative weighting issues arise whenever the covariate specification is not flexible enough for eq. (13) to hold. For example, consider a scalar covariate W_i with a uniform distribution on $[0, 1]$, and a binary treatment with non-linear propensity score $p(W_i) = \min(2W_i, 0.9)$. Suppose that $Y_i(0) = 0$, so that (14) holds. Then $\lambda_{11}(W_i)$ is negative for $W_i \geq 2911/3402 \approx 0.855$. If, say $Y_i(1) = W_i$, so that $\tau(W_i) = W_i \geq 0$, the regression coefficient on the treatment is negative despite uniformly non-negative treatment effects.

[15] 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

The characterization in Proposition 1 also relates to concerns in interpreting multiple-treatment IV estimates with heterogeneous treatment effects (Behaghel et al., 2013; Kirkeboen et al., 2016; Kline & Walters, 2016; Hull, 2018c; Lee & Salanié, 2018; Bhuller & Sigstad, 2022). This connection comes from viewing eq. (9) 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. The first stage of such IV regressions will tend to have the form of eq. (9), 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 the true effects are always positive. In other words, the cross-instrument contamination of causal effects may cause monotonicity violations, even when specifications with individual instruments would be appropriate. 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 some 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. (9). 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.[16]

time effects. If eq. (13) holds with $K = 1$, Proposition 1 shows β estimates a convex combination of treatment 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.

[16] As we discuss in Section 4, one solution to this problem is to interact the examiner instruments with the

4 Solutions

We now discuss solutions to the contamination bias problem raised by Proposition 1. We focus in this discussion on the case of conditionally ignorable treatment assignment (in the sense that eq. (12) holds, and the propensity scores are not degenerate) since solutions that allow for degenerate propensity scores are generally different and have been previously explored in the literature in the context of DiD regressions. 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.

We propose three solutions, each targeting a distinct causal parameter. First, in Section 4.1, we discuss estimation of ATEs. The other two solutions, discussed in Sections 4.2 and 4.3, estimate weighted averages of individual treatment effects using weights that are "easiest" to estimate in that they minimize the semiparametric efficiency bound for estimating weighted ATEs under homoskedasticity. If the weights are allowed to vary across treatments, it is optimal to estimate the effect of each k using the partially linear model in eq. (9), 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 new weighted regression estimator.

4.1 Estimating Average Treatment Effects

Many estimators exist for the ATE of binary treatments—see Imbens and Wooldridge (2009) for a review. A number of these approaches extend naturally to multiple treatments, including matching, inverse propensity score weighting, regressions with interactions, or doubly-robust combinations of these methods (see, among others, Cattaneo (2010), Chernozhukov et al. (2021), and Graham and Pinto (2022)).

Rather than reviewing all of these approaches, we briefly outline a simple implementation of one method which follows the intuition given at the end of Section 2. Namely, one may estimate the ATE vector τ by expanding the partially linear model in eq. (9) to include treatment interaction terms. This generalizes the implementation in the binary treatment case discussed in Imbens and Wooldridge (2009, Section 5.3). Consider the model

$$Y_i = X_i'\beta + q_0(W_i) + \sum_{k=1}^{K} X_{ik}\left(q_k(W_i) - E[q_k(W_i)]\right) + \dot{U}_i, \tag{17}$$

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

controls, which would amount to computing "leniency" separately within location and time cells. This may greatly increase the effective number of instruments, heightening concerns of many-instrument bias in finite samples as well as the importance of appropriate leave-one-out corrections (e.g., Kolesár, 2013).

$E[\dot{U}_i^2]$. When $\mathcal{G}$ consists of linear functions, eq. (17) 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 $\tau_k(w) = \mu_k(w) - \mu_0(w)$. When 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.

Following the intuition at the end of Section 2, the added interactions in eq. (17) ensure that each treatment coefficients β_k is determined only by the outcomes in treatment arms with $D_i = 0$ and $D_i = k$, avoiding the other-treatment 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. (17) by least squares is conceptually straightforward, with sample averages replacing expectations. Furthermore, it can be shown that the resulting estimator achieves the semiparametric efficiency bound under strong overlap (i.e. when the propensity score is bounded away from zero and one) when implemented as a series estimator: it is impossible to construct another regular estimator of the ATE with smaller asymptotic variance.

Nonetheless, under weak overlap, the estimator may be imprecise, with poor finite-sample behavior. This is not a shortcoming of the specific estimator: Khan and Tamer (2010) shows that identification of the ATE is irregular under weak overlap, and it is not possible to estimate it at a $\sqrt{N}$-rate. These results formalize the intuition that it is difficult to reliably estimate the counterfactual outcomes for observations with extreme propensity scores. Overlap concerns tend to be more severe with multiple treatments, because some propensity scores necessarily become closer to zero or one as more treatment arms are added. We thus next turn to the problem of estimating weighted averages of conditional ATEs that downweight these difficult-to-estimate counterfactuals.

4.2 Efficient Weighted 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.[17] We focus on two specifications for the contrast vector, leading to two alternatives to estimating the ATE based on eq. (17). 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 $\sum_{i=1}^N \lambda(W_i) \sum_{k=0}^K \tau_k(W_i) / \sum_{i=1}^N \lambda(W_i)$, the

[17] 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, $\sum_{i=1}^N \lambda(W_i)$.

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.

Given the contrast vector c, we consider the problem of finding the weighting scheme $\lambda(W_i)$ that is the "easiest" to estimate in that it leads to the smallest possible standard errors. This objective has three motivations. First is a robustness concern: a researcher would like to estimate a given contrast as efficiently as possible, at least under the benchmark of constant treatment effects, while being robust to the possibility that the effects are heterogeneous. Under constant effects the weighting $\lambda(W_i)$ is of course immaterial. But the robustness property ensures that the estimand retains a causal interpretation as a convex average of conditional contrasts under weak conditions, avoiding the contamination bias displayed by the regression estimator per Proposition 1. Such a motivation 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.

The second motivation is that the easiest-to-estimate weighting scheme gives a bound on the information available in the data: if these weights nonetheless yield overly large standard errors, inference on other treatment effects (such as the unweighted ATE) will be at least as uninformative. Computing standard errors for this efficient weighted average of treatment effects can be useful as it reveals whether informative conclusions (regardless of how one specifies the treatment effect of interes) are only possible under additional assumptions or with the aid of additional data. If the easiest-to-estimate weighting scheme yields small standard errors even though the standard errors for the unweighted ATE are large, it can be concluded 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 easiest-to-estimate weighting scheme is exactly the same as the weights used by regression. This special case illustrates the second motivation: the optimal binary 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 regression solves the efficient binary treatment weighting scheme, regression estimates establish the extent to which internally valid and informative inference for *any* causal effect are possible with the data at hand.

The third motivation for the easiest-to-estimate weighting scheme is that it offers an

intermediate point along a particular robustness-efficiency "possibility frontier." The ATE estimator based on the interacted specification in eq. (17) 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 cost of large standard errors and non-standard inference under weak overlap. The regression estimator based on eq. (9) 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. (9) is correct, such that treatment effects are constant). But this efficiency comes at the cost of contamination bias under heterogeneous treatment effects. The easiest-to-estimate weighting scheme lies in between these extremes, purging contamination bias and retaining good performance under weak overlap by giving up control over the weights it uses to aggregate the conditional treatment effects $\tau(w)$.[18]

We derive the easiest-to-estimate weighting scheme for multiple treatments in two steps. First, we establish an efficiency 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 weighted average minimizes the semiparametric efficiency bound over the choice of $\lambda(W_i)$. When the contrast vector is specified to allow simultaneous comparison of all treatments, estimation of this efficient weighted average leads to a new estimator; we discuss its implementation 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. (12) 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

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

[18] There are other approaches to resolving the robustness-efficiency tradeoff, such as seeking efficient estimates subject to the weights λ remaining "close" 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.

As formalized in the proof (see Appendix A.2), the efficiency bound $\mathcal{V}_{\lambda,c}$ establishes a lower bound on the asymptotic variance of any regular estimator of $\theta_{\lambda,c}$ under the idealized situation of known propensity scores.[19]

To establish the second step, we choose λ to minimize eq. (18). Simple algebra shows that this variance-minimizing weighting scheme uses weights that are given (up to an arbitrary constant) by

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

The asymptotic variance of this easiest-to-estimate weighting,

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

is the harmonic mean of $\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(W_i)}{p_k(W_i)}$. In contrast, the efficiency bound for the unweighted contrast is given by the arithmetic mean $E\left[\left(\sum_{k=0}^{K} \frac{c_k^2 \sigma_k^2(W_i)}{p_k(W_i)}\right)\right]$, which can be considerably bigger when the propensity scores are not bounded away from zero or one. An appealing feature of the variance-minimizing weighting scheme is that it yields weights that are non-negative, $\lambda_c^* \geq 0$, so that $\theta_{\lambda_c^*,c}$ represents a convex average of conditional contrasts (i.e. an average for some well-defined subpopulation).

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

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*

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

with the semiparametric efficiency bound given by

$$\mathcal{V}_{\lambda^k,c^k} = \sigma^2 E\left[\frac{p_0(W_i) p_k(W_i)}{p_0(W_i) + p_k(W_i)} \right]^{-1}, \tag{21}$$

[19] 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 optimal 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 to characterize the optimal weights in eq. (19) while remaining completely agnostic about heterogeneity in treatment effects. See Crump et al. (2006) for additional discussion in the context of a binary treatment.

where $p_0(W_i) = \Pr(D_i = 0 \mid W_i) = 1 - \sum_{k=1}^{K} p_k(W_i)$.

Per eq. (15), the optimal weighting λ^k coincides with the implicit weighting of conditional ATEs from the partially linear model (9) when it is fit only on observations with $D_i \in \{0, k\}$. This follows since the propensity score in the subsample is given by $\Pr(D_i = k \mid W_i, D_i \in \{0, k\}) = \frac{p_k(W_i)}{p_0(W_i)+p_k(W_i)}$, making $\lambda^k(W_i)$ in eq. (20) the conditional variance of the treatment indicator. Moreover, it follows by standard arguments that regressing Y_i onto X_{ik} and $g(W_i)$ in the subsample with $D_i \in \{0, k\}$ efficiently estimates this weighted average effect provided g is sufficiently flexible (such as when g is linear and W_i consists of group indicators).[20] 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 justification for estimating the effect of any given treatment k by a partially linear regression with an additive covariate adjustment in the subsample with $D_i \in \{0, k\}$. To estimate the effects of all treatments, one runs K such regressions, restricting the sample to one treatment arm and the control group. Such one-treatment-at-a-time regressions are simple to implement and do not require explicitly estimating the propensity score. The regression coefficients are causally interpretable as weighted averages of conditional treatment effects $\tau_k(W_i)$, so long as $p_k/(p_0 + p_k) \in \mathcal{G}$. Moreover, the weighted averages are locally efficient in the sense of Corollary 1.[21]

While the robustness property of the one-treatment-at-a-time regression is well-established, by Angrist (1998) and subsequent extensions, our efficiency characterization appears novel. It 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 that the weighting $p_1(W_i)(1 - p_1(W_i))$ is optimal for estimating the effect of a binary treatment in that it minimizes the asymptotic variance of a particular class of inverse propensity score weighted estimators. Our Corollary 1 extends the optimality of this weighting to all regular estimators, and to multiple treatments. Importantly, this result formalizes a common motivation for using regression to estimate the effects of a single treatment instead of more involved unconditional ATE estimators: when treatment effect heterogeneity is minimal or only weakly correlated with the $\lambda_k^*(W_i)$ weights, the regression's weighted-average effect will be close to the ATE while being more precisely estimated.

A shortcoming of the optimal weighting scheme in Corollary 1 is that it is treatment-

[20] As we discuss in the next subsection, when the propensity score is unknown, the semiparametric efficiency bound for estimating θ_{λ^k, c^k} has an additional term relative to eq. (21) arising from the estimation of the optimal weights eq. (20). Thus, while the regression estimator is semiparametrically efficient, it does not generally attain the efficiency bound derived in Proposition 2 that assumes a known propensity score.

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

specific, so comparisons of the "one-at-a-time" weighted-average effects across treatments are generally not causally interpretable.[22] 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 to make causal comparisons across all teachers.[23]

We thus turn to the question of how Proposition 2 can be used to select an efficient weighting scheme that allows for simultaneous comparisons across all treatment arms. We are interested in reporting estimates of a vector $\beta_{\lambda^{\mathrm{C}}}$ of K coefficients with elements $\beta_{\lambda^{\mathrm{C}},k} = \sum_{i=1}^{N} \lambda^{\mathrm{C}}(W_i)\tau_k(W_i) / \sum_{i=1}^{N} \lambda^{\mathrm{C}}(W_i)$, where the weights λ^{C} are common across treatment arms. If we are equally interested in all $K(K+1)$ contrasts (that is, weighted averages $\mu_j(W_i) - \mu_k(W_i)$, for all $j \neq k$, $j, k = 0, \ldots, K$), a natural approach is to choose the weighting scheme λ^C that minimizes the average variance across all contrasts:

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

where F gives the uniform distribution over the possible (now random) contrasts c, so that $c_j = 1$ with probability $1/(K+1)$ and -1 with probability $1/(K+1)$. Minimizing this expression over λ is equivalent to minimizing eq. (18) with $c_k^2 = 2/(K+1)$, which leads to the following result:

Corollary 2. *Let F denote the uniform distribution over the possible contrast vectors. Suppose that $\sigma_k^2(W_i) = \sigma^2$ for all k. Then the weighting scheme minimizing the average variance bound $\int \mathcal{V}_{\lambda,c} dF(c)$ is given by*

$$\lambda^C(W_i) = 1 \Big/ \sum_{k=0}^{K} p_k(W_i)^{-1}. \tag{22}$$

The weights λ^{C} generalize 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 weights reduce to the one-at-a-time weights in Corollary 1, $\lambda^C(W_i) = \lambda^1(W_i) = \lambda^0(W_i) = p_1(W_i)p_0(W_i)$. With multiple treatments, however, the weights λ^C remain the same for every treatment, allowing for simultaneous comparisons across all treatment pairs (k, ℓ). Again, an appealing feature of these weights is that they are non-negative, so that β_{λ^C} represents

[22] 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 $\tau_1(W_i) - \tau_2(W_i)$.

[23] Note that 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 unsatisfying.

a convex average of conditional contrasts. We next consider estimating this convex average using a weighted regression approach.

4.3 Estimating Efficiently Weighted Average Effects

If the propensity scores $p(W_i)$ were known, one could estimate $\beta_{\lambda^{\mathrm{C}}}$ by a weighted regression of Y_i onto X_i and a constant, with each observation weighted by $\lambda^{\mathrm{C}}(W_i)/p_{D_i}(W_i)$. When the treatment is binary, this estimator reduces to the estimator studied in Li et al. (2018). Since the propensity score is unknown, we replace the infeasible weights with feasible weights $\hat{\lambda}^{\mathrm{C}}(W_i)/\hat{p}_{D_i}(W_i)$, where $\hat{p}_k(W_i)$ is a feasible estimate of the propensity score and $\hat{\lambda}^{\mathrm{C}}(W_i) = 1/\sum_{k=0}^{K} 1/\hat{p}_k(W_i)$. When $\mathcal{G}$ is finite-dimensional, we may use the regression estimator that projects X_{ik} onto $g(W_i)$:

$$\hat{p}_k(W_i) = \arg\min_{\tilde{p}\in\mathcal{G}} \sum_{i=1}^{N} \left(X_{ik} - \tilde{p}(W_i)\right)^2.$$

With linear $\mathcal{G}$, for example, $\hat{p}_k(W_i)$ is simply the fitted value from a linear regression of X_{ik} on W_i and a constant. The resulting estimator can be written as

$$\hat{\beta}_{\hat{\lambda}^{\mathrm{C}},k} = \frac{1}{\sum_{i=1}^{N} \frac{\hat{\lambda}^{\mathrm{C}}(W_i)}{\hat{p}_k(W_i)} X_{ik}} \sum_{i=1}^{N} \frac{\hat{\lambda}^{\mathrm{C}}(W_i)}{\hat{p}_k(W_i)} X_{ik} Y_i - \frac{1}{\sum_{i=1}^{N} \frac{\hat{\lambda}^{\mathrm{C}}(W_i)}{\hat{p}_0(W_i)} X_{i0}} \sum_{i=1}^{N} \frac{\hat{\lambda}^{\mathrm{C}}(W_i)}{\hat{p}_0(W_i)} X_{i0} Y_i. \tag{23}$$

When the treatment is binary and $\mathcal{G}$ is linear, this weighted regression estimator coincides with the usual (unweighted) regression estimator that regresses Y_i onto D_i and W_i.[24]

We now show that the estimator $\hat{\beta}_{\hat{\lambda}^{\mathrm{C}}}$ is efficient in the sense that is achieves the semiparametric efficiency bound for estimating $\beta_{\lambda^{\mathrm{C}}}$:

Proposition 3. *Suppose eq. (12) holds in an i.i.d. sample of size N, with known non-degenerate propensity scores $p_k(W_i)$. Let $\beta^*_{\lambda^C,k} = E[\lambda^C(W_i)\tau_k(W_i)]/E[\lambda^C(W_i)]$, and $\alpha^*_k = \beta^*_{\lambda^C,k} + E[\lambda^C(W_i)\mu_0(W_i)]/E[\lambda^C(W_i)]$. Suppose that the fourth moments of $\lambda^C(W_i)$ and $\mu(W_i)$ are bounded, and that $p_k \in \mathcal{G}$, $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^C(W_i)^2}{p_{k'}(W_i)^2} \in \mathcal{G}$, and $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^C(W_i)}{p_k(W_i)} \in \mathcal{G}$ for all k, k'. Then, provided it is asymptotically linear and regular, $\hat{\beta}_{\hat{\lambda}^C}$ achieves the semiparametric*

[24] To see this, note that in this case $\hat{\lambda}(W_i) = \hat{p}_1(W_i)\hat{p}_0(W_i)$, so that $\hat{\beta}_{\hat{\lambda}^{\mathrm{C}},1} = \frac{\sum_{i=1}^{N}(1-\hat{p}_1(W_i))D_iY_i}{\sum_{i=1}^{N}(1-\hat{p}_1(W_i))D_i} - \frac{\sum_{i=1}^{N}\hat{p}_1(W_i)(1-D_i)Y_i}{\sum_{i=1}^{N}\hat{p}_1(W_i)(1-D_i)} = \frac{\sum_{i=1}^{N}(D_i-\hat{p}_1(W_i))Y_i}{\sum_{i=1}^{N}(D_i-\hat{p}_1(W_i))^2}$, 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 $\sum_i X_{i1}\hat{p}_1(W_i) = \sum_{i=1}^{N}\hat{p}_1(W_i)^2$.

efficiency bound for estimating β_{λ^C}, with diagonal elements of its asymptotic variance of:

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

This efficiency result doesn't rely on homoskedasticity: under heteroskedasticity, the estimator $\hat{\beta}_{\hat{\lambda}^C}$ is still efficient for β_{λ^C} (although the weighting $\lambda^C(W_i)$ need not be optimal under heteroskedasticity). It is stated under the high-level condition that $\hat{\beta}_{\hat{\lambda}^C}$ 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 4. The asymptotic variance of the estimator $\hat{\beta}_{\lambda^C}$ is larger than the asymptotic variance of the infeasible estimator that replaces the estimated weights $\hat{\lambda}^C(W_i)/\hat{p}_{D_i}(W_i)$ in eq. (23) with the infeasible weights $\lambda^C(W_i)/p_{D_i}(W_i)$. The latter achieves the asymptotic variance implied by Corollary 2,

$$\frac{1}{E[\lambda^C(W_i)]^2} E\left[\frac{\lambda^C(W_i)^2\sigma_0^2(W_i)}{p_0(W_i)} + \frac{\lambda^C(W_i)^2\sigma_k^2(W_i)}{p_k(W_i)}\right]. \tag{24}$$

The extra term of the asymptotic variance in Proposition 3 relative to eq. (24) reflects the cost of having to estimate the weights.[25]

5 Practical Guidance and Application

5.1 Measuring and Avoiding Contamination Bias

A researcher interested in estimating the effects of multiple dependent treatments with regression can use Proposition 1 to measure the extent of contamination bias in her estimates. When the treatment assignment is conditionally ignorable, she can further compute one of the three alternative estimators discussed in Section 4.[26] Here we provide practical guidance on both procedures, which we illustrate in an application in the next subsection.

For simplicity, we focus on the case where g is linear and eq. (9) is estimated by ordinary

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

[26] We again 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 solutions under a parallel trends assumption.

least squares (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.

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

$$Y_i = \alpha + \sum_{k=1}^{K} X_{ik}\beta_k + W_i'\gamma + U_i, \tag{25}$$

and 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

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

An estimate of this weight matrix is given by the sample analog:

$$\hat{\Lambda}_i = (\dot{X}'\dot{X})^{-1}\dot{X}_i X_i',$$

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 ATEs 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

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\gamma_{0,k} + W_i'\alpha_{W,0} + \sum_{k=1}^{K} X_{ik}W_i'\gamma_{W,k} + \dot{U}_i. \tag{26}$$

Estimating eq. (26) 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 OLS normal equations, we then obtain the exact decomposition

$$\hat{\beta} = \sum_{i=1}^{N} \text{diag}(\hat{\Lambda}_i)\hat{\tau}(W_i) + \sum_{i=1}^{N} [\hat{\Lambda}_i - \text{diag}(\hat{\Lambda}_i)]\hat{\tau}(W_i), \tag{27}$$

which is the sample analog of the population decomposition in Proposition 1. The first term in this decomposition 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. (27) to other weighted averages of own-treatment effects, including the ones discussed next, to gauge the impact of the regression weighting $\text{diag}(\hat{\Lambda}_i)$.[27]

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 or a low correlation between such heterogeneity and the contamination weights.

Implementing the alternative estimators from Section 4 is also straightforward under the linearity assumptions. For the first solution, estimating the interacted regression

$$Y_i = \alpha_0 + \sum_{k=1}^K X_{ik}\tau_k + W_i'\alpha_{W,0} + \sum_{k=1}^K X_{ik}(W_i - \overline{W})'\gamma_{W,k} + \dot{U}_i. \tag{28}$$

by OLS yields estimates of the unweighted ATEs $\tau_k = E[\tau_k(W_i)]$. Here $\overline{W} = \frac{1}{N}\sum_i W_i$ is the sample average of the covariate vector. 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. (26).

The second solution is to estimate the uninteracted regression,

$$Y_i = \ddot{\alpha}_k + X_{ik}\ddot{\beta}_k + W_i'\ddot{\gamma}_k + \ddot{U}_{ik} \tag{29}$$

among observations assigned either to treatment k or the control group, $D_i \in \{0, k\}$, for each of the treatments $k = 1, \ldots, K$. These one-treatment-at-a-time regressions estimate convex weighted averages of treatment effects, with weights that are efficient under homoskedasticity (in the sense of corollary 1) but which will generally vary across the different treatments. This can make comparisons across treatment arms difficult.

The third solution is to estimate an efficiently weighted average of the conditional ATEs, with weights that are constrained to be common across treatments. Under linearity, we can estimate the common weights λ^{C} as

$$\hat{\lambda}^{\text{C}}(W_i) = \left(\sum_{k=0}^K \hat{p}_k(W_i)^{-1}\right)^{-1}, \tag{30}$$

[27] 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 $\hat{p}_k(W_i) = X_{ik} - \dot{X}_{ik}$ denote estimated propensity scores. We then regress Y_i on X_i, weighting each observation by $\hat{\lambda}^C(W_i)/\hat{p}_{D_i}(W_i)$.

While the second and third solutions may yield more precise estimates than the equal-weighted ATE estimates, the 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.

On the other hand, if the conditional treatment effects $\tau(W_i)$ are approximately independent of the propensity scores $p(W_i)$, the weighting scheme may have little effect on the estimands, even if the treatment effect heterogeneity is substantial. In such cases, we also expect the contamination bias to be small, since the contamination weights are a function of the propensity scores. We next investigate this possibility in our application.

5.2 Application

We illustrate the empirical relevance of our analysis using data from Project STAR, as analyzed 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] All analyses in this section are conducted with our Stata package, `multe`, which researchers can use to gauge the extent of contamination bias in similar applications.

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 uninteracted regression in eq. (25), 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

[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).

Table 1: Project STAR Contamination Bias and Treatment Effect Estimates

	A. Contamination Bias Estimates				
	Regression Coefficient	Own Effect	Bias	Worst-Case Bias	
				Negative	Positive
	(1)	(2)	(3)	(4)	(5)
Small Class Size	5.357	5.202	0.155	-1.654	1.670
	(0.778)	(0.778)	(0.160)	(0.185)	(0.187)
Teaching Aide	0.177	0.360	-0.183	-1.529	1.530
	(0.720)	(0.714)	(0.149)	(0.176)	(0.177)

	B. Treatment Effect Estimates		
	Unweighted	Efficiently-Weighted	
	(ATE)	One-at-a-time	Common
	(1)	(2)	(3)
Small Class Size	5.561	5.295	5.563
	(0.763)	(0.775)	(0.764)
	[0.744]	[0.743]	[0.742]
Teaching Aide	0.070	0.263	-0.003
	(0.708)	(0.715)	(0.712)
	[0.694]	[0.691]	[0.695]

Notes: Panel A estimates the contamination bias and range of potential contamination bias in regression estimates of small class and teaching aide treatment effects for the Project STAR kindergarten analysis. The analysis sample includes 5,868 students. Column 1 reports estimates from a partially linear model in eq. (25). Columns 2 and 3 estimate the own- and cross-treatment decomposition of this estimate in eq. (27). Columns 4 and 5 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. Panel B summarizes estimates of small class and teaching aide treatment effects from different specifications in the kindergarten sample of Project STAR. Column 1 reports estimates of small class and teaching aide treatment effects from the interacted model in eq. (28). Column 2 reports estimates from the treatment-specific regressions in eq. (29). Column 3 reports estimates from the efficiently weighted specification, using the estimated weights in eq. (30). Robust standard errors are reported in parentheses. Standard errors that assume the propensity scores are known are reported in square brackets.

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]

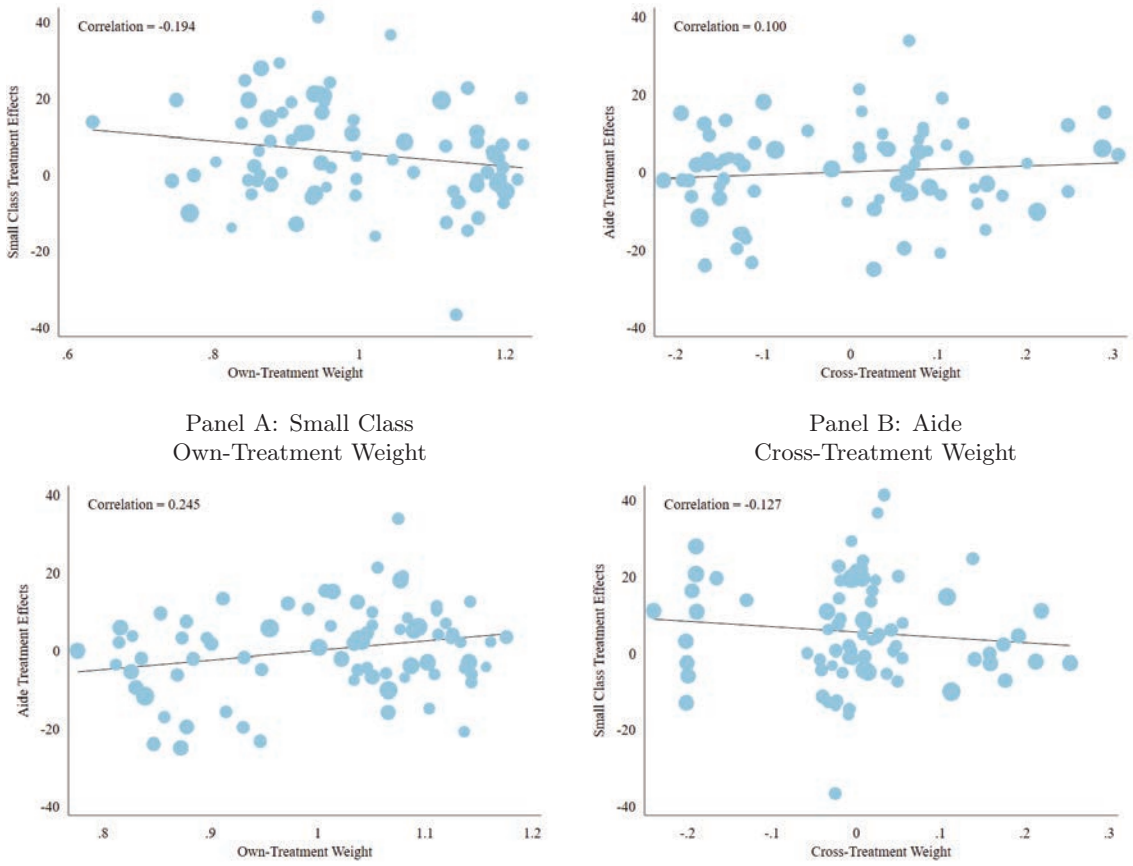
As discussed in Section 2, treatment assignment probabilities vary across the schools, indicated by the fixed effects in W_i. If treatment effects also vary across schools, and if this variation is correlated with the contamination weights $\lambda_{k\ell}(W_i)$, we expect the estimated effect of small class sizes to be partly contaminated by the effect of teaching aides (and vice versa). Net of any contamination bias, we expect each β_k to identify a weighted average of own treatment effects $\tau_k(W_i)$, with convex weights given by $\lambda_{kk}(W_i)$.

Columns 2 and 3 of Table 1 apply the decomposition in eq. (27) to the regression coefficients in column 1. The contamination bias appears to be minimal. The small class size regression estimate of 5.36 is composed of a weighted average of small class size treatment effects equalling 5.20 and a weighted average of teaching aide treatment effects equalling 0.16. Similarly, the teaching aide regression coefficient of 0.18 decomposes into a weighted average of teaching aide treatment effects equalling 0.36 and a weighted average of small class size effects equalling -0.18. Netting out the contamination bias estimate 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 lack of meaningful bias in the regression estimates of Project STAR effects is due to a weak correlation between the conditional treatment effects $\tau(W_i)$ and the contamination weights. These correlations are shown in Figure 2, which plots estimates of the school-specific treatment effects $\tau_k(W_i)$ against the own-treatment and contamination weights $\lambda_{kk}(W_i)$ and $\lambda_{k\ell}(W_i)$ for $\ell \neq k$. Panels A and C show that the own-treatment weight correlation is negative for the small class size treatment (-0.19) and positive for the aide treatment (0.25). The partially linear regression model's estimate of own-treatment effects (column 2 of Table 1) thus understates the average small class size effect and overstates the average aide effect, relative to the ATE. Panels B and D further show that correlation between estimated cross-treatment effects and weights is positive for the small class effect estimate (0.10) and negative for the aide effect estimate (-0.13). There is thus positive contamination bias in the partially linear regression model's estimate of small class size effects and negative contamination bias in the regression's aide effect estimate, as shown in column 2 of Table 1. But neither set of correlations is strong enough to meaningfully bias the estimates.

Importantly, Figure 2 shows that the lack of contamination bias is not due to a lack of treatment effect heterogeneity across schools. There is considerable variation along the y-axis

[29] Our sample and estimates are very similar to—but not exactly the same as—those in Krueger (1999). We use robust (non-clustered) standard errors throughout this analysis, since the randomization of students to classrooms is at the individual level (Abadie et al., 2017). Results are similar when we cluster by classroom.



Panel A: Small Class
Own-Treatment Weight

Panel B: Aide
Cross-Treatment Weight

Panel C: Aide
Own-Treatment Weight

Panel D: Small Class
Cross-Treatment Weight

Figure 2: Project STAR Treatment Effects and Regression Weights

Note: This figure shows correlations between estimated school-specific treatment effects and the implicit school-specific regression weights in column 1 of Table 1. Panels A and B show correlations for the decomposition of the small class treatment effect estimate in columns 2 and 3 of Table 1. Panel A plots the estimated small class treatment effects by school against the estimated own-treatment weights, while Panel B plots the estimated teaching aide treatment effects by school against the estimated cross-treatment weights. Panels C and D show analogous correlations for the decomposition of the teaching aide treatment effect estimate in columns 2 and 3 of Table 1. Panel C plots the estimated teaching aide treatment effects by school against the estimated own-treatment weights, while Panel D plots the estimated small class treatment effects by school against the estimated cross-treatment weights. Correlations and lines of best fit are reported on each panel. The size of the points is proportional to the number of students enrolled in each school.

of each plot. Adjusting for estimation error, we find a standard deviation of $\tau_k(W_i)$ across the schools indexed by W_i of 12.7 for the small class treatment and of 10.9 for the aide treatment.[30] Both standard deviations are an order of magnitude larger than the standard errors in Table 1. Thus, had the experimental design been such that the contamination weights strongly correlate with this variation, sizeable contamination bias could have resulted. In practice, the variation in school-specific propensity scores $p_k(W_i)$ appears to have been largely unrelated to school-specific treatment effects.[31]

To illustrate the potential for contamination bias in this setting, we compute worst-case (positive and negative) weighted averages of the estimated $\tau_k(W_i)$ 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 heterogeneity in treatment effects. Columns 4 and 5 in 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.[32] The worst-case contamination bias equals about 30% of the small class size treatment's magnitude. The relative magnitude is limited by the fact that the school-specific teaching aide treatment effects are all fairly small, so even if the contamination weights average them in the worst possible way we still end up with only a moderate bias. In contrast, the small class size effects are much bigger, so the potential contamination bias in the teaching aide treatment is large relative to the magnitude of the teaching aide treatment effect. Overall, for both treatments, the underlying heterogeneity in this setting makes severe contamination bias possible even though actual contamination bias turns out to be relatively small.

Panel B of Table 1 illustrates the three solutions to the contamination bias problem discussed in Section 4. Column 1 estimates the unweighted ATEs of the small class size and teaching aide treatment, by estimating the interacted regression specification in eq. (28). Column 2 estimates the one-treatment-at-a-time regressions in eq. (29) for $k = 1, 2$. Finally, column 3 estimates the efficiently-weighted ATEs of each treatment, by running a weighted regression of Y_i onto X_i, using the common weight estimates in eq. (30).

As discussed in Remark 4, the optimal weighting schemes underlying the estimates in columns 2 and 3 of Panel B are derived under the assumption that the propensity scores are known. To gauge the relative importance of this assumption, Panel B also reports a version of

[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 own-treatment weights in Figure 2 are highly correlated with the respective treatment propensity score. For the small class size (teaching aide) treatment this correlation is 0.92 (0.73).

[32] 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.

the standard errors computed under the assumption that the sample treatment probabilities in each school match the true propensity scores.[33] This changes the standard errors little, showing that there is minimal cost to estimating the optimal weights.[34]

There turns out to be little difference between the partially linear model estimates of Project STAR treatment effects and these alternative estimates. In columns 1 and 2 of Panel B we estimate a small class size effect of 5.56 and 5.30, which are close to the 5.36 estimate in column 1 of Panel A. Teaching aide effect estimates are also similar: 0.07 and 0.26 in columns 1 and 2 of Panel B, compared to 0.18 in column 1 of Panel A. The efficiently weighted estimates in column 3 of Panel B are again similar: 5.56 for the small class size treatment and 0.00 for the teaching aide treatment. Interestingly, the standard errors are roughly constant across the columns, regardless of whether the propensity score is treated as known.

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 heterogeneous effects, with coefficients on each treatment generally contaminated by the effects of other treatments. We provide intuition for why the influential 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 and related estimators. We discuss three alternative estimators that are free of this bias, including a new estimator that efficiently weights conditional average treatment effects. The analysis underling this estimator also formalizes a virtue of regression adjustment in the binary treatment case: the weighting that it implicitly uses to combine heterogeneous treatment effects minimizes the semiparametric efficiency bound for convex weighted averages of ATEs.

Our application to Project STAR shows that significant contamination bias could arise in RCTs when there is significant treatment effect heterogeneity. Whether the bias *does* arise, however, depends on the correlation between effect heterogeneity and the contamination weights we derive in our theoretical analysis. Researchers can estimate this correlation, and report it alongside the alternative estimates that are free of contamination bias. Such investigation reveals whether, as in our application, the results based on alternative estimators are more similar than the worst-case bounds implied by the theory. Broadly, our analysis

[33] This is the case under stratified block randomization, where a fixed proportion of students is assigned to the two treatments. In contrast, sample treatment proportions need not match the true propensity scores under a Bernoulli trial where each student is assigned to treatments according to a coin toss.

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

highlights the importance of testing the empirical relevance of theoretical concerns with how regression combines heterogeneous effects.

We expect the tools in this paper to be especially relevant in modern RCT designs that generate substantial variation in treatment propensity scores to maximize efficiency (e.g. Tabord-Meehan, 2021). Propensity scores are also likely to vary dramatically in quasi-experimental analyses, such as with teacher VAMs, where a large number of covariates are needed to make the conditionally ignorability of treatment plausible. Contamination bias diagnostics can be a useful tool for ensuring the reliability and robustness of regression estimates in such settings.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (Working Paper No. 24003). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w24003

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. https://doi.org/10.1093/qje/qjab017

Angrist, J., Hull, P., Pathak, P. A., & Walters, C. (2021). Credible school value-added with undersubscribed school lotteries. *The Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01149

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. https://doi.org/10.2307/2998558

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, *14*(1), 57–67. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G

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). Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

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. https://doi.org/10.1093/qje/qjx001

Aronow, P. M., & Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, *60*(1), 250–267. https://doi.org/10.1111/ajps.12185

Athey, S., & Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, *226*(1), 62–79. https://doi.org/10.1016/j.jeconom.2020.10.012

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.

Berman, A., & Plemmons, R. J. (1994). *Nonnegative matrices in the mathematical sciences*. Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611971262

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. https://doi.org/10.2139/ssrn.4014707

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. https://doi.org/10.1016/j.jeconom.2020.12.001

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. https://doi.org/10.1016/j.jeconom.2009.09.023

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

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

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. https://doi.org/10.1257/aer.104.9.2593

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. https://doi.org/10.3386/t0330

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. https://doi.org/10.1093/biomet/asn055

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996. https://doi.org/10.1257/aer.20181169

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

Dobbie, W., & Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, *105*(3), 1272–1311. https://doi.org/10.1257/aer.20130612

Frandsen, B., Lefgren, L., & Leslie, E. (2019). *Judging judge fixed effects* (Working Paper No. 25528). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w25528

Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, *2*(1), 176–196. https://doi.org/10.1214/07-AOAS143

Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180–193. https://doi.org/10.1016/j.aam.2006.12.003

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. https://doi.org/10.1257/aer.103.2.981

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. https://doi.org/10.3386/w27762

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

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. https://doi.org/10.1016/j.jeconom.2021.07.008

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*(2), 315–331. https://doi.org/10.2307/2998560

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Hull, P. D. (2018a). *Estimating hospital quality with quasi-experimental data* (Working Paper). SSRN. https://doi.org/10.2139/ssrn.3118358

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

Hull, P. D. (2018c). *IsoLATEing: Identifying counterfactual-specific treatment effects with cross-stratum comparisons* (Working Paper). SSRN. https://doi.org/10.2139/ssrn.2705108

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Imbens, G. W., & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

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. https://doi.org/10.3386/w14607

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*(6), 2021–2042. https://doi.org/10.3982/ECTA7372

Kirkeboen, L. J., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, *131*(3), 1057–1111. https://doi.org/10.1093/qje/qjw019

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. https://doi.org/10.1093/qje/qjw027

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, *96*(3), 863–876. https://doi.org/10.1257/aer.96.3.863

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. *Unpublished Working Paper*.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532. https://doi.org/10.1162/003355399556052

Krueger, A. B., & Summers, L. H. (1988). Efficiency wages and the inter-industry wage structure. *Econometrica*, *56*(2), 259–293. https://doi.org/10.2307/1911072

Lee, S., & Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, *86*(6), 1939–1963. https://doi.org/10.3982/ECTA14269

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. https://doi.org/10.1080/01621459.2016.1260466

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. https://doi.org/10.1257/aer.103.5.1797

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. https://doi.org/10.1016/S0378-3758(00)00193-2

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619. https://doi.org/10.3982/ECTA15463

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. https://doi.org/10.1257/aer.20190221

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. https://doi.org/10.2139/ssrn.3537773

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. https://doi.org/10.2307/2951752

Norris, S. (2019). *Examiner inconsistency: Evidence from refugee appeals* (Working Paper No. 2018-75). University of Chicago, Becker Friedman Institute for Economics. https://doi.org/10.2139/ssrn.3267611

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. https://doi.org/10.2307/2532304

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, *56*(4), 931. https://doi.org/10.2307/1912705

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2022). *What's trending in difference-in-differencees? A synthesis of the recent econometrics literature*. arXiv: 2201.01194.

Skinner, J. (2011). Causes and consequences of regional variations in health care. In M. V. Pauly, T. G. Mcguire, & P. P. Barros (Eds.), *Handbook of health economics* (pp. 45–93). Elsevier. https://doi.org/10.1016/B978-0-444-53592-4.00002-5

Słoczyński, T. (2022). Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics*, *104*(3), 1–9. https://doi.org/10.1162/rest_a_00953

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006

Tabord-Meehan, M. (2021). *Stratification trees for adaptive randomization in randomized controlled trials* (Working Paper). University of Chicago. Chicago, IL.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256

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. https://doi.org/10.1007/978-1-4757-2545-2

Wooldridge, J. M. (2021). *Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators* (Working Paper). SSRN. https://doi.org/10.2139/ssrn.3906345

Appendix A Proofs

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. (9), 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. (16). 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 that $\lambda_{kk}(W_i) \geq 0$ 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. (14) holds $E[\tilde{\tilde{X}}_{ik} E[Y_i(0) \mid W_i]] = 0$. Similarly, under eq. (13), $E[\tilde{\tilde{X}}_{ik} \mid W_i] = 0$, 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. (16).

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, we show that $\lambda_{kk}(W_i) \geq 0$ if eq. (13) 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 Equation (13) 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. (12), 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 $(1+ts_k)^{1/2}$ is continuously differentiable for t small enough, with derivative $s_k/2\sqrt{1+ts_k}$, 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 Proof of Proposition 3

We first derive the semiparametric efficiency bound for estimating β_{λ^C} 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 $\hat{\beta}_{\hat{\lambda}^C}$ 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) $N^{-1}\sum_i \lambda^C(W_i) \to E[\lambda^C(W_i)]$; (iii) $N^{-1}\sum_i \lambda^C(W_i)\sum_{k=0}^K c_k E[Y_i(k) s_{y,k}(Y_i \mid W_i) \mid W_i] \to \sum_{k=0}^K c_k E[\lambda^C(W_i) Y_i(k) s_{y,k}(Y_i(k) \mid W_i)]$; (iv) $N^{-1}\sum_{i=1}^N \lambda^C(W_i)^2 \sum_{k,k'} \frac{s_{p,k}(W_i)}{p_k(W_i)} c_{k'} \mu_{k'}(W_i) \to E[\lambda^C(W_i)^2 \sum_{k,k'} \frac{s_{p,k}(W_i)}{p_k(W_i)} c_{k'} \mu_{k'}(W_i)]$; (v) $N^{-1}\sum_{i=1}^N \lambda^C(W_i)^2 \sum_{k=0}^K \frac{s_{p,k}(W_i)}{p_k(W_i)} \to E[\lambda^C(W_i)^2 \sum_{k=0}^K \frac{s_{p,k}(W_i)}{p_k(W_i)}]$; and (vi) $\beta_{\lambda^C} \to \beta^*_{\lambda^C}$. 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, $\beta_{\lambda^{\mathrm{C}}}$, as $\beta_N(\theta) = \sum_{i=1}^{N}\lambda^{\mathrm{C}}(w_i)\int y\sum_{k=0}^{K} c_k f_k(y \mid w_i)dy / \sum_{i=1}^{N}\lambda^{\mathrm{C}}(w_i)$, where $\lambda^{\mathrm{C}}(w_i) = 1/\sum_{k=0}^{K} p_k(w_i)^{-1}$. 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 $h(w) = \lambda^{\mathrm{C}}(w)\sum_{k=0}^{K} c_k \int y s_{y,k}(y \mid w) f_k(y \mid w)dy$, and $\tilde{h}(W_i) = \sum_{k'=0}^{K} c_{k'}\mu_{k'}(W_i) - \beta^*_{\lambda^C}$. The derivative may then be written as

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

where 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 $\frac{1}{E[\lambda^{\mathrm{C}}(W_i)]}\sum_{k=0}^{K} E[\lambda^{\mathrm{C}}(W_i)^2\frac{\tilde{h}(W_i)}{p_k(W_i)}s(Y_i, D_i \mid W_i)] = 0$ 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

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

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

$$\sqrt{N}(\beta_{\lambda^{\mathrm{C}}} - \beta^*_{\lambda^C}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi^*(W_i) + o_p(1) \qquad \text{and} \qquad \psi^*(W_i) = \frac{\lambda^{\mathrm{C}}(W_i)}{E[\lambda^{\mathrm{C}}(W_i)]}(\tau(W_i) - \beta^*_{\lambda^{\mathrm{C}}}). \tag{31}$$

Second, we show that

$$\sqrt{N}(\hat{\beta}_{\hat{\lambda}^{\mathrm{C}}} - \beta^*_{\lambda^C}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, D_i, W_i) + o_p(1), \tag{32}$$

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

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

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

Equation (31) follows directly under the assumptions of the proposition by the law of large numbers and the fact that the variance of $\lambda^{\mathrm{C}}(W_i)(\tau(W_i) - \beta^*_{\lambda^{\mathrm{C}}})$ is bounded. To show eq. (32), write $\hat{\beta}_{\hat{\lambda}^{\mathrm{C}},k} = \hat{\alpha}_k - \hat{\alpha}_0$, 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 $m_k(Y_i, D_i, W_i, \alpha^*, p) = \lambda^{\mathrm{C}}(W_i)\frac{X_{ki}}{p_k(W_i)}(Y_i - \alpha^*_k)$, and α^* is a $(K+1)$ dimensional vector with elements $\alpha^*_k = E[\lambda^{\mathrm{C}}(W_i)\mu_k(W_i)]/E[\lambda^{\mathrm{C}}(W_i)]$.

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

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

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

$$\delta_{k,k'}(W_i) = \lambda^{\mathrm{C}}(W_i)(\mu_k(W_i) - \alpha^*_k) \left(\frac{\lambda^{\mathrm{C}}(W_i)}{p_{k'}(W_i)^2} - \frac{\mathbb{1}\{k = k'\}}{p_k(W_i)} \right).$$

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

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

which yields eq. (32).

Appendix B Connections to Difference-in-Differences 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 two-way fixed effects regression with a general treatment specification. We then show how Proposition 1 applies to four particular treatment specifications: a static binary treatment with a single intervention date, a static binary treatment with multiple intervention dates, a dynamic "event study" treatment specification, and a static multivalued treatment specification (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 $j = 1, \dots, n$ which are observed over time periods $t = 1, \dots T$. 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. A two-way fixed effects specification sets $W_i = (J_i, T_i)$ and $g(W_i) = \alpha + (\mathbb{1}\{J_i = 2\}, \dots, \mathbb{1}\{J_i = n\}, \mathbb{1}\{T_i = 2\}, \dots, \mathbb{1}\{T_i = T\})'\gamma$, 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 $Y_j = (Y_{j1}, \dots, Y_{jT})$ and a (fixed) T-vector of $(K+1)$-valued treatments $D_j = (D_{j1}, \dots, D_{jT})$, with $D_{jt} \in \{0, \dots, K\}$. As in the main text, X_{jt} denotes a K-vector of treatment indicators derived from D_{jt}. As we show below, X_{jt} will vary in complexity depending on whether the regression specification allows for dynamic treatment effects.

We make two assumptions on the potential outcomes. First we assume that potential outcomes $Y_{jt}(d)$ depend only on the current treatment status d, such that $Y_{jt} = Y_{jt}(D_{jt})$. As we show below, this assumption need not rule out dynamic treatment effects depending the specification of D_{jt} (e.g. D_{jt} can index each of the periods after an intervention). Second, we make a parallel trends assumption by writing untreated potential outcomes as

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

for fixed α_j and λ_t, and assuming

$$E[\eta_{jt}] = 0. \tag{33}$$

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

$$Y_{jt} = X_{jt}'\tau_{jt} + \alpha_j + \lambda_t + \eta_{jt}. \tag{34}$$

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

To fit this setup into the framework of Section 3, we interpret the expectation in eq. (9) as averaging over the unobserved shocks affecting potential outcomes for the observed units and time periods. That is $(\beta, g) = \text{argmin}_{\tilde{\beta}\in\mathbb{R}^K, \tilde{g}\in\mathcal{G}} N^{-1}\sum_{j=1}^n\sum_{t=1}^T E_{\tau,\eta}[(Y_{jt} - X_{jt}'\tilde{\beta} - \tilde{g}(W_{jt}))^2]$, where $E_{\tau,\eta}[\cdot]$ denotes expectation 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. (14) in Assumption 2 holds under the two-way fixed effects specification. In other words, the parallel trend assumption implies that our controls $g(W_i)$ correctly specify 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. (11) as

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

where the second equality uses eqs. (33) and (34), 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

$$\beta_k = \sum_{j,t}\lambda_{kk}(j,t)E[\tau_{jt,k}] + \sum_{\ell\neq k}\sum_{j,t}\lambda_{k\ell}(j,t)E[\tau_{jt,\ell}] \tag{35}$$

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. (13) 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 treatment for a single intervention date First, consider the canonical DiD case where the first $n_1 < n$ units are treated in the last $T_1 < T$ periods and are untreated in the earlier periods $1, \ldots, T - T_1$. The remaining units are never treated. Treatment effects are assumed to be static, in that outcomes only depend only on the current treatment status. This nests the simplest DiD specification where $T = 2$ and $T_1 = 1$ (e.g. Card & Krueger, 1994). Let $D_{jt} \in \{0, 1\}$ denote the indicator that unit j is treated in period t. In this setup there are only two unique treatment vectors D_j, either a vector of zeros or a vector of a series of zeros followed by a series of ones, so $X_{jt} = \mathbb{1}\{j \leq n_1\}\,\mathbb{1}\{t > T - T_1\}$ and $\tilde{X}_{jt} = \mathbb{1}\{j \leq n_1\}\,\mathbb{1}\{t > T - T_1\} - \frac{T_1}{T}\,\mathbb{1}\{j \leq n_1\} - \frac{n_1}{n}\,\mathbb{1}\{t > T - T_1\} + \frac{n_1 T_1}{N}$ from the above expressions. Since X_{jt} is scalar, $\tilde{\tilde{X}}_{jt,1} = \tilde{X}_{jt}$ and the second term in eq. (35) drops out; the remaining first term can be shown to simplify 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.

Thus, although the propensity score cannot be written as a linear combination of unit and time indicators ($E[X_{jt} \mid W_{jt}] = X_{jt} = \alpha_{J_i} \lambda_{T_i}$) and hence eq. (13) does not hold, the canonical DiD specification estimates a weighted average of treatment effects with positive and easily interpretable weights. Moreover, because the treatment is binary, there is no contamination bias from other treatments. These results are consistent with the literature, which finds no negative weighting issues with non-staggered and static DiD interventions.

Static treatment with multiple intervention dates Next, consider a DiD setting where units adopt (and potentially drop) a binary treatment at different time periods—a case that de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) study in detail. For example, different states j may choose to roll out a policy in different time periods and a researcher wishes to estimate the average effect of this policy using this staggered adoption. We continue to assume that the treatment is static, such that potential outcomes are still only indexed by the binary treatment D_{jt}. However, instead of two unique treatment paths as in the previous example, now the treatment vectors $D_j = (D_{j1}, \ldots, D_{jT})$ take on different values depending on the intervention date (i.e. T_1 now varies across the units j).

As before, since D_{jt} is binary, there is no scope for contamination bias in this setting. We continue to have $\tilde{X}_{jt} = \tilde{\tilde{X}}_{jt,1}$ and the second term in eq. (35) drops out. The remaining term coincides with the expression for the regression estimand that de Chaisemartin and D'Haultfœuille (2020) give in their Theorem 1. The treatment weights $\lambda_{11}(j,t)$ are not guaranteed to be convex since eq. (13) does not hold. In contrast, Athey and Imbens (2022) consider

staggered DiD regressions where eq. (13) 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.

To illustrate the negative weighting problem in our framework, 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.[35] Following the same steps as before, we obtain $\beta_1 = \sum_{j,t} \lambda_{11}(j,t) E[\tau_{jt,1}]$ with

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

where $\kappa = 2(n_E n_L + n_E n_N + n_N n_L)$ and $\lambda_{11}(j,t) = 0$ otherwise. The first two of these expressions are always non-negative, but the sign of $\lambda_{11}(j,3)$ for early adopters depends on the relative sizes of the other two groups. If there are more late adopters than never adopters, this weight is negative. Otherwise, all weights are positive.

Event study with staggered intervention dates 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. 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, the second contamination bias term in eq. (35) 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

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

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 $\tau_{jts} = Y_{jt}(s) - Y_{jt}(0),\ s \in \{-2, 1, 0, 1\}$ 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) matrix 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 periods 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.

Mover regressions: multiple treatments with multiple transitions. Finally, consider a "mover regression" in a setting with a static multivalued treatment $D_{jt} \in \{0, \ldots, K\}$ with multiple transitions of units between treatment states, leading to multiple treatment paths. 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. (13) 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. (35) 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.