American Economic Review 2024, 114(12): 4015–4051
https://doi.org/10.1257/aer.20221116

Contamination Bias in Linear Regressions[†]

By Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár*

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 nonconvex averages of the effects of other treatments. We discuss three estimation approaches that avoid such contamination bias, including the targeting of easiest-to-estimate weighted average effects. A reanalysis of nine empirical applications finds economically and statistically meaningful contamination bias in observational studies; contamination bias in experimental studies is more limited due to smaller variability in propensity scores. (*JEL* C21, C31, C51, H75, I21, I28)

Consider a linear regression of an outcome Y_i on a vector of treatments $\mathbf{X}_i$ and a vector of flexible controls $\mathbf{W}_i$. The treatments are assumed to be as good as randomly assigned conditional on the controls. For example, $\mathbf{X}_i$ 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 $\mathbf{W}_i$. Or, in an education value-added model (VAM), $\mathbf{X}_i$ might indicate the matching of students i to different teachers or schools with $\mathbf{W}_i$ including measures of student demographics and lagged achievement which yield a credible selection-on-observables assumption. The regression might also be the first stage of an instrumental variables (IV) regression leveraging the assignment of multiple decision-makers (e.g. bail judges) indicated in $\mathbf{X}_i$, which is as-good-as-random conditional on some controls $\mathbf{W}_i$. These sorts of regressions are widely used across many fields in economics.[1]

*Goldsmith-Pinkham: Yale University (email: paul.goldsmith-pinkham@yale.edu); Hull: Brown University (email: peter_hull@brown.edu); Kolesár: Princeton University (email: mkolesar@princeton.edu). Isaiah Andrews was the coeditor for this article. We thank Alberto Abadie, Jason Abaluck, Josh Angrist, Tim Armstrong, Kirill Borusyak, Kyle Butts, Clément de Chaisemartin, Peng Ding, Len Goff, Jin Hahn, Xavier D'Haultfœuille, Simon Lee, Bernard Salanié, Pedro Sant'Anna, Tymon Słoczyński, Isaac Sorkin, Jonathan Roth, Jacob Wallace, Stefan Wager, and numerous seminar participants for helpful comments. Hull acknowledges support from National Science Foundation grant SES-2049250. Kolesár acknowledges support by the Sloan Research Fellowship and by the National Science Foundation grant SES-22049356. Mauricio Cáceres Bravo, Jerray Chang, William Cox, and Dwaipayan Saha provided expert research assistance. An earlier draft of this paper circulated under the title "On Estimating Multiple Treatment Effects with Regression."

[†]Go to https://doi.org/10.1257/aer.20221116 to visit the article page for additional materials and author disclosure statement(s).

[1]Prominent RCTs where randomization probabilities vary across strata include Project STAR (Krueger 1999) and the RAND Health Insurance Experiment (Manning et al. 1987). Prominent VAM examples include studies of teachers (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2014), schools (Angrist et al. 2017; Angrist et al. 2024; Mountjoy and Hickman 2021), and health care institutions (Abaluck et al. 2021; Geruso, Layton, and Wallace 2020). Prominent "judge IV" examples include Kling (2006); Meestas, Mullen, and Strand (2013); and Dobbie and Song (2015).

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

We first derive a general characterization of such *contamination bias* in multiple-treatment regressions.[2] We show the core problem by focusing on the special case of a set of mutually exclusive treatment indicators, though our characterization applies even when the treatments are not restricted to be binary or mutually exclusive. To separate the problem from the typical challenge of omitted variables bias (OVB), we assume a best-case scenario where the covariate parametrization is flexible enough to include the treatment propensity scores (e.g., with a linear covariate adjustment, we assume that the propensity scores are linear in the covariates). This condition holds trivially if the only covariates are strata indicators. Under these conditions, we show that the regression coefficient on each treatment identifies a convex weighted average of its causal effects plus a contamination bias term given by a linear combination of the causal effects of other treatments, with weights that sum to zero. Thus, each treatment effect estimate will generally incorporate the effects of other treatments, unless the effects are uncorrelated with the contamination weights. Since these weights sum to zero some are necessarily negative, further complicating the interpretation of the coefficients.

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

Contamination bias also arises under an alternative "model-based" identifying assumption that, rather than making assumptions on the treatment's "design" (i.e. propensity scores), posits that the covariate specification spans the conditional mean of the potential outcome under no treatment, $Y_i(0)$. In a linear model with

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

unit and time fixed effects, this reduces to the parallel trends restriction often used in difference-in-differences (DiD) and event study regressions. It is common for $\mathbf{X}_i$ 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.[3] We show that replacing the restriction on propensity scores in our characterization with an assumption on $Y_i(0)$ generates an additional issue: the own-treatment weights are negative whenever the implicit propensity score model used by the regression to partial out the covariates and the other treatments fits probabilities greater than one. This result shows that the negative weighting and contamination bias issues documented previously in the context of two-way fixed effects regressions (e.g., Goodman-Bacon 2021; Sun and Abraham 2021; de Chaisemartin and D'Haultfoeuille 2020; De Chaisemartin and D'Haultfoeuille 2023; Callaway and Sant'Anna 2021; Borusyak, Jaravel, and Spiess 2024; Wooldridge 2021; Hull 2018b) are more general, and conceptually distinct, problems.[4] Negative weighting arises because regressions leveraging model-based restrictions on $Y_i(0)$ may fit treatment probabilities exceeding one. Contamination bias arises because additive covariate adjustments don't account for the nonlinear dependence of a given treatment on the other treatments and covariates. This generates a different form of propensity score misspecification: a nonzero fitted probability of a given treatment, even when one of the other treatments is known to be nonzero.[5]

We then discuss three solutions to the contamination bias problem, and their trade-offs. These solutions apply when the propensity scores are nondegenerate, such as in an RCT or other "design-based" regression specification.[6] First, a conceptually principled solution is to adapt approaches to estimating the average treatment effect (ATE) of a conditionally ignorable binary treatment to the multiple treatment case (e.g., Cattaneo 2010; Chernozhukov et al. 2018; Chernozhukov, Newey, and Singh 2022; de los Angeles Resa and Zubizarreta 2020; Graham and Campos de Xavier Pinto 2022). For example, one could run a regression that includes interactions between the treatments and demeaned controls, or combine such regression with inverse propensity score weighting for doubly robust estimation. Such ATE estimators work well under strong overlap of the covariate distribution for units in each treatment arm. But they may be imprecise under limited overlap or be outright infeasible with overlap failures—common scenarios in observational studies (Crump et al. 2009).

This practical consideration motivates an alternative approach: estimating a weighted average of treatment effects, as regression does in the binary treatment

[3] Alternatively $\mathbf{X}_i$ may indicate multiple contemporaneous treatments, as in certain "mover" regressions.

[4] Our analysis also relates to issues with interpreting multiple-treatment IV estimates (Behaghel, Crépon, and Gurgand 2013; Kirkeboen, Leuven, and Mogstad 2016; Kline and Walters 2016; Hull 2018a; Lee and Salanié 2018; Bhuller and Sigstad 2024).

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

[6] Solving the contamination bias problem under model-based identification approaches requires either targeting subpopulations of the treated or applying substantive restrictions on the conditional means of potential outcomes under treatment. We do not explore this case as it has already been studied extensively in the DiD context (e.g., De Chaisemartin and D'Haultfoeuille 2023; Sun and Abraham 2021; Callaway and Sant'Anna 2021; Borusyak et al. 2024; Wooldridge 2021).

case, while avoiding the contamination bias problem with multiple treatments. We derive the weights that are easiest to estimate, in the sense of minimizing a semiparametric efficiency bound under homoskedasticity. This easiest-to-estimate weighting (EW) scheme is always convex; it corresponds to weighting schemes previously proposed in Crump et al. (2006); Li, Morgan, and Zaslavsky (2018); and Li and Li (2019). The weights also coincide with the implicit linear regression weights when the treatment is binary (i.e. the Angrist 1998 case). In the multiple treatment case, the EW scheme that allows the weights to be treatment specific can be implemented by a simple second solution: a linear regression which restricts estimation to the individuals who are either in the control group or the treatment group of interest. Since the weights are treatment-specific, these one-treatment-at-a-time regressions make direct comparisons across treatment arms challenging. The third solution is to impose common weights across treatments in the EW scheme; these weights can be implemented using a weighted regression approach. We show how researchers can gauge the extent of contamination bias in practice and implement these tools with a new R and Stata package, multe.[7]

We study the empirical relevance of contamination bias in nine applications: six RCTs with stratified randomization and three observational studies of racial disparities. We find economically and statistically significant bias in two of the three observational studies with no evidence for bias in any of the experimental studies. In a detailed analysis of one experiment, the Project STAR trial, we show that the lack of contamination bias is driven by small variation in the contamination weights rather than limited effect heterogeneity. This analysis highlights the importance of conducting contamination bias diagnostics—particularly in observational studies where covariates are expected to generate high variability in propensity scores, and thus likely in contamination weights.

We structure the rest of the paper as follows. Section I illustrates contamination bias in a simple stylized setting. Section II characterizes the general problem, and discusses connections to previous analyses. Section III provides three solutions, and gives guidance for measuring and avoiding contamination bias in practice. Section IV illustrates these tools in nine applications. Section V concludes. Supplemental appendices collect all proofs and extensions, discuss the connection between our contamination bias characterization and that in the DiD literature, and provide details on the applications and additional exhibits.

I. Motivating Example

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

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

sections consider nonexperimental settings with richer control specifications, both theoretically and empirically.

A. *Convex Weights with One Randomized Treatment*

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

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

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

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

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

$$\beta = \phi\,\tau_1(0) + (1-\phi)\,\tau_1(1), \tag{2}$$

where

$$\phi = \frac{\text{var}\big[D_i \,|\, W_i = 0\big]\Pr(W_i = 0)}{\sum_{w=0}^{1}\text{var}\big[D_i \,|\, W_i = w\big]\Pr(W_i = w)} \in [0,1]$$

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

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

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

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

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

The first equality in equation (4) follows from the law of iterated expectations, the second equality follows by the conditional random assignment of D_i and the third equality uses $E[\tilde{D}_i \mid W_i] = 0$. Hence, the first summand in equation (3) is zero. Analogous arguments show that

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

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 $E\big[\text{var}[D_i \mid W_i]\big] = E\big[E[\tilde{D}_i^2 \mid W_i]\big] = E[\tilde{D}_i^2]$, it follows that we can write the second summand in equation (3) as

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

proving the representation of β in equation (2).

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

B. *Contamination Bias with Two Randomized Treatments*

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

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

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

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

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

where $\tilde{\tilde{X}}_{i1}$ again denotes a population residual, but now from regressing X_{i1} on W_i, a constant, *and* X_{i2}. Unlike before, this residual is uncorrelated with but *not* mean-independent of the remaining regressors (W_i, X_{i2}) because the dependence between X_{i1} and X_{i2} is nonlinear. 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} depends on W_i unless the treatment assignment is completely random. Thus, in general, $\tilde{\tilde{X}}_{i1} \neq X_{i1} - E[X_{i1} \,|\, W_i, X_{i2}]$.

Because $\tilde{\tilde{X}}_{i1}$ does not coincide with a conditionally de-meaned X_{i1}, we can not generally reduce equation (6) to an expression involving only the effects of the first treatment arm, τ_{i1}. It turns out that we nevertheless still have $E\big[\tilde{\tilde{X}}_{i1} Y_i(0)\big] = 0$, as in equation (4), since the auxiliary regression residuals are still uncorrelated with any individual characteristic like $Y_i(0)$.[8] The regression thus does not suffer from OVB. However, we do not generally have $E\big[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}\big] = 0$. Instead, simplifying equation (6) by the same steps as before leads to the expression

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

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

The second *contamination bias* term in equation (7) arises because the residualized small class treatment $\tilde{\tilde{X}}_{i1}$ is not conditionally independent of the second full-time aide treatment X_{i2} within schools, despite being uncorrelated with X_{i2} by construction. This can be seen by viewing $\tilde{\tilde{X}}_{i1}$ as the result of an equivalent two-step

[8] 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 equation (4), are both uncorrelated with $Y_i(0)$. It follows that $E\big[\tilde{\tilde{X}}_{i1} Y_i(0)\big] = 0$.

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* $\lambda_{12}(W_i)$ nonzero.

C. *Illustration and Intuition*

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

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

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

$$\begin{aligned}\beta_1 &= E\left[\lambda_{11}(W_i) \cdot 0\right] + E\left[\lambda_{12}(W_i)\,\tau_2(W_i)\right] \\ &= 0 + \left(-\frac{99}{106} \times 1 + \frac{99}{106} \times 0\right) \Big/ 2 \approx -0.47.\end{aligned}$$

Thus, in this example, a researcher would conclude that small classrooms have a sizable negative effect on student achievement—equal in magnitude to around one-half of the true teaching aide effect in school 1—despite the true small-classroom effect

being zero for all students. This treatment effect coefficient can be engineered to match an arbitrary magnitude and sign by varying the heterogeneity of the teaching aide effects across schools.

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

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

II. 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 $\mathbf{X}_i$ and flexible transformations of a control vector $\mathbf{W}_i$. We focus on the case of mutually exclusive indicators $X_{ik} = \mathbf{1}\{D_i = k\}$ for values of an underlying treatment $D_i \in \{0, \ldots, K\}$ (with the $\mathbf{1}\{D_i = 0\}$ indicator omitted). We extend the characterization to a general (i.e. potentially nonbinary) $\mathbf{X}_i$ in online Appendix A.1.

We suppose the effects of $\mathbf{X}_i$ on Y_i are estimated by a partially linear model:

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + g(\mathbf{W}_i) + U_i, \tag{8}$$

where $\boldsymbol{\beta}$ and g are defined as the minimizers of expected squared residuals $E\big[U_i^2\big]$:

$$(\boldsymbol{\beta}, g) = \underset{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^K, \tilde{g} \in \mathcal{G}}{\arg\min}\; E\Big[\big(Y_i - \mathbf{X}_i'\tilde{\boldsymbol{\beta}} - \tilde{g}(\mathbf{W}_i)\big)^2\Big] \tag{9}$$

for some linear space of functions $\mathcal{G}$. This setup nests linear covariate adjustment by setting $\mathcal{G} = \{\alpha + \mathbf{w}'\boldsymbol{\gamma} : [\alpha, \boldsymbol{\gamma}']' \in \mathbb{R}^{1+\dim(\mathbf{W}_i)}\}$, in which case equation (8) gives a linear regression of Y_i on $\mathbf{X}_i$, $\mathbf{W}_i$, 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 (Robinson 1988).

Two examples highlight the generality of this setup.

EXAMPLE 1 (Multi-armed RCT): $\mathbf{W}_i$ *is a vector of mutually exclusive indicators for experimental strata, within which* $\mathbf{X}_i$ *is randomly assigned to individuals i, and 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* $t = 1, \ldots, T$. $\mathbf{W}_i = (J_i, T_i)$ *where* $J_i = j$ *and* $T_i = t$ *denote the underlying unit and period, and* $g(\mathbf{W}_i) = \alpha + \left(\mathbf{1}\{J_i = 2\}, \ldots, \mathbf{1}\{J_i = n\}, \mathbf{1}\{T_i = 2\}, \ldots, \mathbf{1}\{T_i = T\}\right)'\gamma$ *includes unit and period indicators.* $\mathbf{X}_i$ *contains indicators for leads and lags relative to a deterministic treatment adoption date,* $A(j) \in \{1, \ldots, T, \infty\}$ *with at least one lead excluded to prevent collinearity.*

Example 1 nests the motivating RCT example in Section I, allowing for an arbitrary number of experimental strata in $\mathbf{W}_i$ and multiple treatment arms in $\mathbf{X}_i$. 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 and Abraham 2021; de Chaisemartin and D'Haultfoeuille 2020; De Chaisemartin and D'Haultfoeuille 2023; Callaway and Sant'Anna 2021; Borusyak, Jaravel, and Spiess 2024; Wooldridge 2021). We elaborate on the connections to this literature in online Appendix B by considering general two-way fixed effects (TWFE) specifications with nonrandom 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.[9]

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

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

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

[9] 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 will generally add bias terms to our decomposition of β, below. Similarly, we do not consider multicollinearity issues like in Borusyak et al. (2024) by assuming a unique solution to equation (9). For event studies this means we assume some units are never treated, with $A(j) = \infty$.

A. *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 $Y_i = Y_i(D_i) = Y_i(0) + \mathbf{X}_i'\boldsymbol{\tau}_i$ where $\boldsymbol{\tau}_i$ 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 $\boldsymbol{\tau}(\mathbf{W}_i) = E[\boldsymbol{\tau}_i \mid \mathbf{W}_i]$, so that $\tau_k(\mathbf{W}_i)$ is the conditional ATE for the kth treatment. We let $\mathbf{p}(\mathbf{W}_i) = E[\mathbf{X}_i \mid \mathbf{W}_i]$ denote the vector of propensity scores, so that $p_k(\mathbf{W}_i) = \Pr(D_i = k \mid \mathbf{W}_i)$. 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. $p_k(\mathbf{w}) \in \{0, 1\}$ for all $\mathbf{w}$. This is the case in Example 2, since $\mathbf{X}_i$ is a nonrandom function of $\mathbf{W}_i$. We return to practical questions of propensity score support in Section III.

We make two assumptions to interpret β_k in terms of the effects $\boldsymbol{\tau}_i$. First, we assume mean-independence of the potential outcomes and treatment, conditional on the controls:

ASSUMPTION 1: $E[Y_i(k) \mid D_i, \mathbf{W}_i] = E[Y_i(k) \mid \mathbf{W}_i]$ *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:

$$(11) \qquad \big(Y_i(0), \ldots, Y_i(K)\big) \perp D_i \mid \mathbf{W}_i.$$

Such conditional random assignment appears in Example 1. In Example 2, where treatment is a nonrandom function of the unit and time indices in $\mathbf{W}_i$, Assumption 1 holds trivially.

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

ASSUMPTION 2: *Let* $\mu_0(\mathbf{w}) = E[Y_i(0) \mid \mathbf{W}_i = \mathbf{w}]$ *and recall* $p_k(\mathbf{w}) = E[X_{ik} \mid \mathbf{W}_i = \mathbf{w}]$. *Either*

$$(12) \qquad p_k \in \mathcal{G}$$

for all k, *or*

$$(13) \qquad \mu_0 \in \mathcal{G}.$$

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, equation (12) requires the propensity scores to be linear in $\mathbf{W}_i$ (compare with equation (30) in Angrist and Krueger 1999). This condition holds trivially in Example 1, since $\mathbf{W}_i$ is a vector of indicators for groups within which $\mathbf{X}_i$ 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 $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbf{p}(\mathbf{W}_i)$.

In Example 2, with $\mathbf{X}_i$ being a deterministic function of unit and time indices and $g(\mathbf{W}_i)$ including unit and time fixed effects, equation (12) fails because the propensity scores are binary. They cannot be captured by a linear combination of the TWFEs. However, equation (13) is satisfied by a parallel trends assumption: that the average untreated potential outcomes $Y_i(0)$ are linear in the unit and time effects. We elaborate on this setup in online Appendix B.[10]

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 $Y_i = Y_i(0) + \mathbf{X}_i'\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a vector collecting the constant effects. The only source of bias when regressing Y_i on $\mathbf{X}_i$ and controls is then the unobserved variation in the untreated potential outcomes $Y_i(0)$. But it follows from the expression for $\boldsymbol{\beta}$ in equation (10) that there is no such OVB when Assumption 2 holds:

$$\begin{aligned}\boldsymbol{\beta} &= E\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\left(E\left[\tilde{\mathbf{X}}_i Y_i(0)\right] + E\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]\boldsymbol{\tau}\right)\\ &= E\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]^{-1}\underbrace{E\left[\tilde{\mathbf{X}}_i E\left[Y_i(0)\,|\,\mathbf{W}_i\right]\right]}_{=0} + \boldsymbol{\tau} = \boldsymbol{\tau}.\end{aligned}$$

Here the first equality uses the fact that $E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i'\right] = E\left[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i'\right]$ because $\tilde{\mathbf{X}}_i$ is a vector of projection residuals, and the second equality uses the law of iterated expectations and Assumption 1. Under equation (12), $E\left[\tilde{\mathbf{X}}_i\,|\,\mathbf{W}_i\right] = 0$, so that the term in braces is zero by another application of the law of iterated expectations: $E\left[\tilde{\mathbf{X}}_i E\left[Y_i(0)\,|\,\mathbf{W}_i\right]\right] = E\left[E\left[\tilde{\mathbf{X}}_i\,|\,\mathbf{W}_i\right]E\left[Y_i(0)\,|\,\mathbf{W}_i\right]\right] = 0$. It is likewise zero under equation (13) since $\tilde{\mathbf{X}}_i$ is by definition of projection orthogonal to any function in $\mathcal{G}$ such that $E\left[\tilde{\mathbf{X}}_i E\left[Y_i(0)\,|\,\mathbf{W}_i\right]\right] = E\left[\tilde{\mathbf{X}}_i\mu_0(\mathbf{W}_i)\right] = 0$. 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 double robustness property have been previously observed in, for instance, Robins et al. (1992).

When treatment effects are heterogeneous but $\mathbf{X}_i$ contains a *single* treatment indicator, $\boldsymbol{\beta}$ identifies a weighted average of the conditional effects $\boldsymbol{\tau}(\mathbf{W}_i)$. Specifically, since by the previous argument we still have $E\left[\tilde{\mathbf{X}}_i Y_i(0)\right] = 0$, it follows from equation (10) that

$$(14)\quad \boldsymbol{\beta} = \frac{E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\boldsymbol{\tau}_i\right]}{E\left[\tilde{\mathbf{X}}_i^2\right]} = E\left[\lambda_{11}(\mathbf{W}_i)\boldsymbol{\tau}(\mathbf{W}_i)\right],\ \text{with}\ \lambda_{11}(\mathbf{W}_i) = \frac{E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\,|\,\mathbf{W}_i\right]}{E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\right]},$$

where the second equality uses iterated expectations and the identity $E\left[\tilde{\mathbf{X}}_i^2\right] = E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\right]$. Under equation (12), $E\left[\tilde{\mathbf{X}}_i\mathbf{X}_i\,|\,\mathbf{W}_i\right] = E\left[\tilde{\mathbf{X}}_i^2\,|\,\mathbf{W}_i\right] = \text{var}\left[\mathbf{X}_i\,|\,\mathbf{W}_i\right]$, so the weights further simplify to $\lambda_{11}(\mathbf{W}_i) = \frac{\text{var}\left[\mathbf{X}_i\,|\,\mathbf{W}_i\right]}{E\left[\text{var}\left[\mathbf{X}_i\,|\,\mathbf{W}_i\right]\right]} \geq 0$. This extends the

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

Angrist (1998) result to a general control specification; versions of this extension appear in, for instance, Angrist and Krueger (1999); Angrist and Pischke (2009, Chapter 3.3), and Aronow and Samii (2016).

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

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

PROPOSITION 1: *Under Assumptions 1 and 2, the treatment coefficients in* (8) *identify*

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

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

$$\lambda_{kk}(\mathbf{W}_i) = \frac{E\big[\tilde{\tilde{X}}_{ik} X_{ik}\,|\,\mathbf{W}_i\big]}{E\big[\tilde{\tilde{X}}_{ik}^2\big]} = \frac{p_k(\mathbf{W}_i)\big(1 - E^*\big[X_{ik}\,|\,\mathbf{X}_{i,-k} = \mathbf{0}, \mathbf{W}_i\big]\big)}{E\big[\tilde{\tilde{X}}_{ik}^2\big]}, \text{ and}$$

$$\lambda_{k\ell}(\mathbf{W}_i) = \frac{E\big[\tilde{\tilde{X}}_{ik} X_{i\ell}\,|\,\mathbf{W}_i\big]}{E\big[\tilde{\tilde{X}}_{ik}^2\big]} = -\frac{p_\ell(\mathbf{W}_i)\,E^*\big[X_{ik}\,|\,X_{i\ell} = 1, \mathbf{W}_i\big]}{E\big[\tilde{\tilde{X}}_{ik}^2\big]}$$

with $E\big[\lambda_{kk}(\mathbf{W}_i)\big] = 1$ *and* $E\big[\lambda_{k\ell}(\mathbf{W}_i)\big] = 0$. *Furthermore, if equation* (12) *holds,* $\lambda_{kk}(\mathbf{W}_i) \geq 0$.

Proposition 1 shows that the coefficient on X_{ik} in equation (8) is a sum of two terms. The first term is a weighted average of conditional ATEs $\tau_k(\mathbf{W}_i)$, with *own treatment weights* $\lambda_{kk}(\mathbf{W}_i)$ that average to one—generalizing the characterization of the single-treatment case, equation (14). The expression for λ_{kk} implies that these weights are convex if the implicit linear probability model used to compute $\tilde{\tilde{X}}_{ik}$ fits probabilities that lie below one, $E^*\big[X_{ik}\,|\,\mathbf{X}_{i,-k} = \mathbf{0}, \mathbf{W}_i\big] \leq 1$. The second term is a weighted average of treatment effects for *other* treatments $\tau_\ell(\mathbf{W}_i)$, with *contamination weights* $\lambda_{k\ell}(\mathbf{W}_i)$ that average to zero. Because the contamination weights are zero on average, they must be negative for some values of the controls unless they are all identically zero.[11] This is the case when the implicit linear probability model correctly predicts that $X_{ik} = 0$ if $X_{i\ell} = 1$.

[11] 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. In a finite-population setting, Abadie et al. (2020) show that β identifies matrix-weighted averages of individual treatment effect vectors τ_i; however, they do not discuss the interpretation of the estimand.

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

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

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

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

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

Remark 3: The results in Proposition 1 are stated for the case when $\mathbf{X}_i$ are mutually exclusive treatment indicators. In online Appendix A.1 we relax this assumption to allow for combinations of nonmutually exclusive treatments (either discrete or continuous). In this case, the own-treatment weights $\lambda_{kk}(\mathbf{W}_i)$ may be negative even if equation (12) holds.

Remark 4: While we derived Proposition 1 in the context of a causal model, an analogous result follows for descriptive regressions that do not assume

potential outcomes or impose Assumption 1. Consider, specifically, the goal of estimating an average of conditional group contrasts $E[Y_i|D_i = k, \mathbf{W}_i = \mathbf{w}] - E[Y_i|D_i = 0, \mathbf{W}_i = \mathbf{w}]$ with a partially linear model equation (8) and replace condition (13) with an assumption that $E[Y_i|D_i = 0, \mathbf{W}_i = \mathbf{w}] \in \mathcal{G}$. 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 nonconvex 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, $E[Y_i|D_i, \mathbf{W}_i]$.

B. *Implications*

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

Second, in the multiple treatment case, there is a potential for contamination bias from *other* treatment effects, regardless of which condition in Assumption 2 holds. This form of bias is relevant whenever one uses an additive covariate adjustment, no matter how flexibly the covariates are specified. Versions of this problem have been noted in, for example, the Sun and Abraham (2021) analysis of DiD regressions with treatment leads and lags or the Hull (2018b) analysis of mover regressions (see online Appendix B).[12] Proposition 1 shows such contamination bias arises much more broadly, however.

The characterization in Proposition 1 also relates to concerns in interpreting multiple-treatment IV estimates with heterogeneous effects (e.g., Behaghel, Crépon, and Gurgand 2013; Kirkeboen, Leuven, and Mogstad 2016; Kline and Walters 2016; Hull 2018a; Lee and Salanié 2018; Bhuller and Sigstad 2024). This connection comes from viewing equation (8) as the second stage of an model estimated by a

[12] The negative weights issue raised in de Chaisemartin and D'Haultfoeuille (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'Haultfoeuille (2023) (when $K > 1$), arise because the treatment probability is not linear in the unit and time effects. If equation (12) 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, nonnegative.

control function approach; in the linear case, for example, $g(\mathbf{W}_i)$ can be interpreted as giving the residuals from a first-stage regression of $\mathbf{X}_i$ on a vector of valid instruments $\mathbf{Z}_i$. In the single-treatment case, the resulting $\boldsymbol{\beta}$ coefficient has an interpretation of a weighted average of conditional local average treatment effects under the appropriate first-stage monotonicity condition (Imbens and Angrist 1994). But as in Proposition 1 this interpretation fails to generalize when $\mathbf{X}_i$ includes multiple mutually exclusive treatment indicators: each β_k combines the local effects of treatment k with a nonconvex average of the effects of other treatments.

Finally, Proposition 1 has implications for single-treatment estimation with multiple instruments and flexible controls if the first stage has the form of equation (8), where now Y_i is interpreted as the treatment and $\mathbf{X}_i$ 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 nonconvexity, the regression specification may fail to satisfy the effective monotonicity condition even when τ_{ik} is always positive: the cross-instrument contamination of causal effects may cause monotonicity violations, even when specifications with individual instruments do not. This issue is distinct from previous concerns over monotonicity failures in multiple-instrument designs (e.g., Mueller-Smith 2015; Frandsen, Lefgren, and Leslie 2023; Norris 2019; Mogstad, Torgovitsky, and Walters 2021), which are generally also present in such just-identified specifications. It is also distinct from concerns about insufficient flexibility in the control specification when monotonicity holds unconditionally (e.g., Blandhol et al. 2022).

This new monotonicity concern may be especially important in "examiner" 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 estimators can be thought of as implementing versions of a jackknife estimator (Angrist et al. 1999), based on a first stage that uses examiner indicators as instruments, similar to equation (8). Proposition 1 thus raises a new concern with these analyses when controls (such as time fixed effects) are needed to ensure ignorable treatment assignment.

III. Solutions

We now discuss three solutions to the contamination bias problem raised by Proposition 1, each targeting a distinct causal parameter. First, in Section IIIA, we discuss estimation of unweighted ATEs. The other two solutions target weighted averages of individual treatment effects using an easiest-to-estimate weighting (EW) scheme in that the weights minimize the semiparametric efficiency bound for estimating weighted ATEs under homoskedasticity. In the second solution, the weights are allowed to vary across treatments, while in the third, they are constrained to be common across treatments. In Section IIIB we characterize these estimation targets, while in Section IIIC we discuss how to estimate them; we also outline our proposed guidance to researchers in measuring contamination bias.

Implementing the first solution requires strong overlap (i.e. that treatment propensity scores are bounded away from zero and one) while the other two solutions require nonempty overlap, ruling out fully degenerate propensity scores. Solutions

allowing for degenerate propensity scores require either targeting subpopulations of the treated or adding substantive restrictions on conditional means of treated potential outcomes (beyond equation (13), which only restricts untreated potential outcomes). We refer readers to De Chaisemartin and D'Haultfoeuille (2023); Sun and Abraham (2021); Callaway and Sant'Anna (2021); Borusyak, Jaravel, and Spiess (2024); and Wooldridge (2021) for such solutions in the context of DiD regressions.

A. *Estimating Average Treatment Effects*

Many estimators exist for the ATE of binary treatments—see Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018) for reviews. Several of these approaches extend naturally to multiple treatments: including matching on covariates or the propensity score, inverse propensity score weighting, balancing weights, interacted regression, or doubly robust methods (see, among others, Cattaneo 2010; de los Angeles Resa and Zubizarreta 2020; Chernozhukov, Newey, and Singh 2022; and Graham and Campos de Xavier Pinto 2022). Here we summarize the last two approaches.

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

$$Y_i = \mathbf{X}_i'\beta + q_0(\mathbf{W}_i) + \sum_{k=1}^{K} X_{ik}\big(q_k(\mathbf{W}_i) - E\big[q_k(\mathbf{W}_i)\big]\big) + \dot{U}_i, \tag{16}$$

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

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

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

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\tau_k + \mathbf{W}_i'\boldsymbol{\alpha}_{W,0} + \sum_{k=1}^{K} X_{ik}\big(\mathbf{W}_i - \bar{\mathbf{W}}\big)'\boldsymbol{\gamma}_{W,k} + \dot{U}_i. \tag{17}$$

by ordinary least squares (OLS), where $\bar{\mathbf{W}} = \frac{1}{N}\sum_i \mathbf{W}_i$ is the sample average of the covariate vector. More generally, to increase the plausibility of the key assumption that $\mu_k \in \mathcal{G}$, one may constrain $\mathcal{G}$ only by nonparametric smoothness assumptions. Given a sequence of basis functions $\big\{b_j(\mathbf{W}_i)\big\}_{j=1}^{\infty}$, such as polynomials or splines, one

then approximates q_k with a linear combination of the first J terms, with J increasing with the sample size, thus tailoring the model complexity to data availability. Given a choice of J, estimation and inference can proceed as in the parametric case; the only difference is that the baseline covariates $\mathbf{W}_i$ in equation (17) are replaced by the basis vector $\big(b_1(\mathbf{W}_i), \ldots, b_J(\mathbf{W}_i)\big)'$ and $\bar{\mathbf{W}}$ is replaced by the sample average of this expansion. This estimator has been studied in the binary treatment case by Chen, Hong, and Tarozzi (2008) and Imbens, Newey, and Ridder (2007), with the latter providing a detailed analysis of how to choose J and the former showing that this sieve estimator achieves the semiparametric efficiency bound under strong overlap: it is impossible to construct another regular estimator of the ATE with smaller asymptotic variance.

An attractive alternative approach combines the interacted regression with inverse propensity score weighting. Instead of using OLS to estimate equation (16) one uses weighted least squares, weighting observations by the inverse of some estimate $\hat{p}_{D_i}\big(\mathbf{W}_i\big)$ of the propensity score (e.g., Robins, Rotnitzky, and Zhao 1994; Wooldridge 2007; and Słoczyński and Wooldridge 2018). An advantage of this approach is that it is doubly robust: the estimator is consistent so long as either the propensity score estimator is consistent or the outcome model is correct (i.e., $\mu_k \in \mathcal{G}$). A recent literature shows how the double robustness property, when combined with cross-fitting, reduces the sensitivity of the ATE estimate to over fitting or regularization bias in estimating the nuisance functions p_k and μ_k. Cross-fitting also allows for using more flexible methods to approximate p_k and μ_k, including modern machine learning methods (see, e.g. Chernozhukov et al. 2018; Chernozhukov et al. 2022; Chernozhukov, Newey, and Singh 2022).

Either approach should work reliably in conventional stratified RCTs and other settings with strong overlap. But under weak overlap, when propensity scores are not bounded away from zero and one, all of these ATE estimators may be imprecise and have poor finite-sample behavior. This is not a shortcoming of the specific estimator; indeed, Khan and Tamer (2010) show that under weak overlap, $\sqrt{N}$-estimation of the ATE is not possible. Furthermore, if some propensity scores attain values of zero or one, the ATE is not even point-identified. These results formalize the intuition that it is difficult or impossible to estimate the counterfactual outcomes for units with extreme propensity scores.[13] Such extreme propensity scores are common in observational settings. The solutions we discuss next downweight these difficult-to-estimate counterfactuals to address this practical challenge.

B. *Easiest-to-Estimate Averages of Treatment Effects*

Suppose in a sample of observations $i = 1, \ldots, N$ we wish to estimate a weighted average of conditional potential outcome contrasts $\sum_{i=1}^{N}\lambda(\mathbf{W}_i)\sum_{k=0}^{K}c_k\,\mu_k(\mathbf{W}_i)/\sum_{i=1}^{N}\lambda(\mathbf{W}_i)$, where $\mu_k(\mathbf{W}_i) = E\big[Y_i(k)\,|\,\mathbf{W}_i\big]$, $\mathbf{c}$ is a $(K+1)$-dimensional contrast vector with elements c_k, and $\lambda(\mathbf{W}_i)$ is some weighting

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

scheme.[14] We focus on two specifications for the contrast vector, leading to two alternatives the ATE target. First, for separately estimating the effect of each treatment k, we set $c_k = 1$, $c_0 = -1$ and set the remaining entries of $\mathbf{c}$ to 0. The contrast of interest then becomes $\sum_{i=1}^{N} \lambda(\mathbf{W}_i)\, \tau_k(\mathbf{W}_i) / \sum_{i=1}^{N} \lambda(\mathbf{W}_i)$, the weighted ATE of treatment k. Second, we specify $\mathbf{c}$ so as to allow us to simultaneously contrast the effects of all K treatments; we discuss this further below. For each contract vector $\mathbf{c}$, we characterize in this section the easiest-to-estimate weighting (EW) scheme $\lambda(\mathbf{W}_i)$ that leads to the smallest possible standard errors under homoskedasticity. We discuss estimation of the corresponding estimands in Section IIIC.

This optimization problem has four motivations. First, there is a robustness motivation: a researcher would like to estimate a given contrast as precisely as possible, at least under the benchmark of constant treatment effects, while being robust to the possibility that the effects are heterogeneous. While the optimization problem does not impose convexity, it turns out that the EW scheme is convex. Hence, the resulting estimand identifies a convex average of conditional contrasts under heterogeneous treatment effects, and avoids any contamination bias. Such a robustness property presumably underlies the popularity of regression as a tool for estimating the effect of a binary treatment: the regression estimator is efficient under homoskedasticity and constant treatment effects while, by the Angrist (1998) result, retaining a causal interpretation under heterogeneous effects.[15]

Second, the EW scheme gives a bound on the information available in the data: if the scheme yields overly large standard errors, inference on other treatment effects (such as the unweighted ATE) must be at least as uninformative. Computing the EW standard errors thus reveals whether informative conclusions for *any* treatment effect estimand are only possible under additional assumptions or with the aid of additional data. In fact, we show below that in the binary treatment case the EW scheme is exactly the same as that used by regression. Recall that in the binary treatment case, the regression treatment weights are proportional to the conditional variance of treatment, $\text{var}[D_i | \mathbf{W}_i] = p_1(\mathbf{W}_i)(1 - p_1(\mathbf{W}_i))$. Because these weights tend to zero as $p_1(\mathbf{W}_i)$ tends to zero or one, regression 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 for informative inference even when one cannot precisely estimate the unweighted ATE.

Third, the EW scheme can be viewed as offering an intermediate point along a particular robustness-precision "possibility frontier." The ATE estimator based on the interacted specification in equation (16) lies on one end of this frontier, being the most robust to treatment effect heterogeneity (i.e. retaining a clear interpretation regardless of the form of $\boldsymbol{\tau}(w)$ or how it relates to the propensity scores). But this robustness comes at the cost of imprecision and nonstandard inference under weak

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

[15] There are several motivations for the interest in convex weights. First, $\lambda(\mathbf{W}_i) \geq 0$ ensures the estimand captures average effects for *some* well-defined (and characterizable) subpopulation. Second, it prevents what Small et al. (2017) call a sign-reversal: if $\tau_k(\mathbf{w})$ has the same sign for all $\mathbf{w}$ (+,0 or −), then the estimand will also have this sign. Blandhol et al. (2022) call such estimands "weakly causal." Finally, the estimand satisfies a population version of what Robins et al. (2007) call boundedness: the estimand lies in the support of $\tau_k(\mathbf{w})$.

overlap. The regression estimator based on equation (8) lies on the other end of the frontier: it is likely to be precise even when overlap is weak (and is efficient under homoskedasticity if the partly linear model in equation (8) is correct, such that treatment effects are constant). But this precision comes at the cost of contamination bias under heterogeneous treatment effects. The EW scheme lies in between these extremes, purging contamination bias and retaining good performance under weak overlap by giving up explicit control over the treatment effect weighting, letting it be data-determined.[16]

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

We derive the EW scheme in two steps. First, we establish a precision benchmark (a semiparametric efficiency bound) for estimation of a given weighted average of treatment effects under the idealized scenario that the propensity score is known. Second, we determine which weights λ minimize the bound.

The following proposition establishes the first step of our derivation:

PROPOSITION 2: *Suppose equation (11) holds in an i.i.d. sample of size N, with known nondegenerate propensity scores* $p_k(\mathbf{W}_i)$. *Let* $\sigma_k^2(\mathbf{W}_i) = \text{var}\big[Y_i(k) \,|\, \mathbf{W}_i\big]$. *Consider the problem of estimating the weighted average of contrasts*

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

where the weighting function λ *and contrast vector* $\mathbf{c}$ *are both known. Suppose the weighting function satisfies* $E\big[\lambda(\mathbf{W}_i)\big] \neq 0$, *and that the second moments of* $\lambda(\mathbf{W}_i)$ *and* $\mu(\mathbf{W}_i)$ *are bounded. Then, conditional on the controls* $\mathbf{W}_1, \ldots, \mathbf{W}_N$, *the semiparametric efficiency bound is almost-surely given by*

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

As formalized in the online Appendix A.2 proof, $\mathcal{V}_{\lambda,c}$ establishes the lower bound on the asymptotic variance of any regular estimator of $\theta_{\lambda,c}$ under the idealized case of known propensity scores.[17]

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

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

To establish the second step, we minimize equation (18) over λ. Simple algebra shows that the EW scheme is (up to an arbitrary constant) given by

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

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

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

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

$$\lambda^k(\mathbf{W}_i) = \frac{p_0(\mathbf{W}_i)\, p_k(\mathbf{W}_i)}{p_0(\mathbf{W}_i) + p_k(\mathbf{W}_i)}. \tag{20}$$

Per equation (14), the weighting λ^k coincides with the weighting of conditional ATEs from the partially linear model (8) when it is fit only on observations with $D_i \in \{0, k\}$, provided $p_k/(p_k + p_0) \in \mathcal{G}$.[18] Corollary 1 thus gives a precision justification for estimating the effect of any given treatment k by a partially linear regression in the subsample with $D_i \in \{0, k\}$ under a homoskedasticity benchmark, complementing the robustness motivation discussed earlier.[19] To estimate the effects of all treatments one can run K such one-treatment-at-a-time regressions, one for each treatment arm. Plugging equation (20) into equation (18) reveals that the asymptotic variance is bounded so long as the overlap between the covariate distribution in each treatment arm is nonempty: that is, $\Pr\big(\{\mathbf{W}_i : p_k(\mathbf{W}_i) > \varepsilon\} \cap \{\mathbf{W}_i : p_0(\mathbf{W}_i) > \varepsilon\}\big) > \varepsilon$ for some $\varepsilon > 0$.

For binary treatments, Crump et al. (2006, Corollary 5.2) and Li, Morgan, and Zaslavsky (2018, Corollary 1) show that the weighting $p_1(\mathbf{W}_i)\big(1 - p_1(\mathbf{W}_i)\big)$ minimizes the asymptotic variance of a particular class of inverse propensity score weighted estimators. Our Corollary 1 extends the property to all regular estimators, and to multiple treatments.

Remark 5: The one-treatment-at-a-time regression can also be motivated as a direct solution to contamination bias in the partially linear regression in equation (8). In particular, as discussed in Section IIA, contamination bias arises because the implicit linear probability model $E^*[X_{ik} \,|\, \mathbf{X}_{i,-k}, \mathbf{W}_i]$ incorrectly imposes additive

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

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

separability between $\mathbf{X}_{i,-k}$ and $\mathbf{W}_i$. To solve this issue, one can include interactions between the controls and $\mathbf{X}_{i,-k}$. This is similar to the interacted regression in equation (16), except we exclude the interaction $X_{ik}\big(q_k(\mathbf{W}_i) - E\big[q_k(\mathbf{W}_i)\big]\big)$. Simple algebra shows that this regression is equivalent to the one-treatment-at-a-time regression.

Remark 6: The population analog of the estimand implied by the weighting in Corollary *1*, $E\big[\lambda_k(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\big]/E\big[\lambda_k(\mathbf{W}_i)\big]$, also identifies the effect of a particular marginal policy intervention. Consider the effects of a class of policies indexed by a scalar δ that restrict treatments to $\{0,k\}$ by increasing the propensity score of treatment k to $p_k^{\delta}(\mathbf{W}_i)$ and setting $p_0^{\delta}(\mathbf{W}_i) = 1 - p_k^{\delta}(\mathbf{W}_i)$.[20] Then the marginal effect of the increasing the policy intensity δ per unit treated at $\delta = 0$ is given by $E\big[\partial p_k^{\delta}(\mathbf{W}_i)/\partial\delta \cdot \boldsymbol{\tau}(\mathbf{W}_i)\big]/E\big[\partial p_k^{\delta}(\mathbf{W}_i)/\partial\delta\big]$ (see Zhou and Opacic 2022 for derivation and discussion). Thus, the weights $\lambda_k(\mathbf{W}_i) = p_0(\mathbf{W}_i)\,p_k(\mathbf{W}_i)/\big(p_0(\mathbf{W}_i) + p_k(\mathbf{W}_i)\big)$ identify the marginal policy effect when they correspond to the derivative $\partial p_k^{\delta}(\mathbf{W}_i)/\partial\delta$. For example, Zhou and Opacic (2022) show this holds for policies that increase the log odds of a single binary treatment by a constant δ, such as by increasing the intercept in a logit model for treatment.

A shortcoming of the EW scheme in Corollary 1 is that it is treatment-specific, making it difficult to compare the weighted-average effects across treatments.[21] This issue is especially salient when the control group is arbitrarily chosen, such as in teacher VAM regressions which omit an arbitrary teacher from estimation and seek causal comparisons across all teachers.

We thus turn to the question of how Proposition 2 can be used to select a weighting scheme which allows for simultaneous comparisons across all treatment arms. Suppose that the contrast of interest is drawn at random from a given marginal treatment distribution $\Pr(D_i = k) = \pi_k$, so that $c_j = 1$ with probability $\pi_j\big(1 - \pi_j\big)/\big(1 - \sum_{k=0}^{K}\pi_k^2\big)$ and $c_j = -1$ with the same probability.[22] Let F_{π} denote this distribution over the (now random) contrasts. If the researcher wishes to report an accurate contrast estimate but needs to commit to a weighting scheme before knowing the contrast of interest, it is optimal to minimize the expected variance

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

Minimizing this expression over λ is equivalent to minimizing equation (18) with $c_k^2 = 2\pi_k(1-\pi_k)$, which yields equation (19) with this contrast specification as the optimal weighting. Thus, the optimal weights are proportional

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

[21] Formally, for treatments 1 and 2, we estimate the weighted averages $\sum_i\lambda^1(\mathbf{W}_i)\,\tau_1(\mathbf{W}_i)/\sum_i\lambda^1(\mathbf{W}_i)$ and $\sum_i\lambda^2(\mathbf{W}_i)\,\tau_2(\mathbf{W}_i)/\sum_i\lambda^2(\mathbf{W}_i)$. 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 $\tau_1(\mathbf{W}_i) - \tau_2(\mathbf{W}_i)$. This critique also applies to the own-treatment weights in Proposition 1. Thus even without contamination bias one may find the implicit multiple-treatment regression weighting deficient.

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

to $\left[\sum_{k=0}^{K} \frac{\pi_k(1-\pi_k)\sigma_k^2(\mathbf{W}_i)}{p_k(\mathbf{W}_i)}\right]^{-1}$. Specializing to the homoskedastic case leads to the following result.

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

$$\lambda^{CW}(\mathbf{W}_i) = \left[\sum_{k=0}^{K} \frac{\pi_k(1-\pi_k)}{p_k(\mathbf{W}_i)}\right]^{-1}.$$

The easiest-to-estimate common weighting (CW) scheme λ^{CW} generalizes the intuition behind the single binary treatment (Corollary 1), placing lower weight on strata with extreme propensity scores. When the treatment is binary, $K = 1$, the π_k's do not matter and the CW scheme reduces to that in Corollary 1: $\lambda^{CW}(\mathbf{W}_i) = \lambda^1(\mathbf{W}_i) = \lambda^0(\mathbf{W}_i) = p_1(\mathbf{W}_i)p_0(\mathbf{W}_i)$. With multiple treatments, however, the weights λ^{CW} remain the same for every treatment, allowing for simultaneous comparisons across all treatment pairs (k, ℓ).

There are two natural choices for the marginal treatment probabilities π. First, when equally interested in all contrasts, one can set $\pi_k = 1/(K+1)$. This weighting scheme was previously proposed by Li and Li (2019); our characterization of it in terms of optimizing a semiparametric efficiency bound is, to our knowledge, novel. Second, if more common treatments are of greater interest, we may set π_k to the empirical treatment probabilities $N^{-1}\sum_i X_{ik}$. This weighting targets precise estimation of contrasts involving more common treatments at the expense of contrasts involving less common treatments. We use this choice in our empirical applications in Section IV. For either choice of weights, the resulting asymptotic variance in equation (18) remains bounded so long as the overlap between covariate distributions in each treatment arm is not empty: $\Pr\left(\cap_{k=0}^{K}\left\{\mathbf{W}_i : p_k(\mathbf{W}_i) > \varepsilon\right\}\right) > \varepsilon$ for some $\varepsilon > 0$. Nonempty overlap is a substantially weaker assumption than strong overlap, needed for $\sqrt{N}$-estimation of the unweighted ATE, which requires this probability to equal one. For instance, in the nine empirical applications below, nonempty overlap always holds, but strong overlap fails in six applications.

C. *Practical Guidance in Measuring and Avoiding Contamination Bias*

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

For simplicity, we focus on the case where g is linear and equation (8) is estimated by OLS. We suppose 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 $\mathbf{W}_i$. These conditions hold,

for example, when $\mathbf{W}_i$ contains a set of mutually exclusive group indicators. When $\mathcal{G}$ is unrestricted, the recommendations in this section would require nonparametric approximations for g analogous to those discussed in Section IIIA.

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

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

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

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

$$Y_i = \alpha_0 + \sum_{k=1}^{K} X_{ik}\gamma_{0,k} + \mathbf{W}_i'\boldsymbol{\alpha}_{W,0} + \sum_{k=1}^{K} X_{ik}\mathbf{W}_i'\boldsymbol{\gamma}_{W,k} + \dot{U}_i. \tag{22}$$

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

Using the OLS normal equations, we then obtain a sample analog of the population decomposition in Proposition 1:

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{N} \operatorname{diag}\big(\hat{\boldsymbol{\Lambda}}_i\big)\hat{\boldsymbol{\tau}}(\mathbf{W}_i) + \sum_{i=1}^{N}\Big[\hat{\boldsymbol{\Lambda}}_i - \operatorname{diag}\big(\hat{\boldsymbol{\Lambda}}_i\big)\Big]\hat{\boldsymbol{\tau}}(\mathbf{W}_i). \tag{23}$$

The first term estimates the own-treatment effect components, $E\big[\lambda_{kk}(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\big]$, while the second term estimates the contamination bias components, $\sum_{\ell\neq k}E\big[\lambda_{k\ell}(\mathbf{W}_i)\,\tau_\ell(\mathbf{W}_i)\big]$. 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 equation (23) to other weighted averages of own-treatment effects, including the ones discussed next, to gauge the impact of the regression weighting $\operatorname{diag}\big(\hat{\boldsymbol{\Lambda}}_i\big)$.[23]

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

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

effect heterogeneity, small contamination weights, or a low correlation between the two.

Estimation of the unweighted ATE and the EW and CW schemes is also straightforward under the linearity assumptions. First, estimating equation (17) by OLS yields estimates of the unweighted ATEs $\tau_k = E\big[\tau_k(\mathbf{W}_i)\big]$. The estimates are numerically equivalent to $\hat{\tau}_k = \hat{\gamma}_{0,k} + \bar{\mathbf{W}}'\hat{\boldsymbol{\gamma}}_{W,k}$, where $\hat{\gamma}_{0,k}$ and $\hat{\boldsymbol{\gamma}}_{W,k}$ are OLS estimates of equation (22).

Second, the EW scheme from Corollary 1 can be estimated using the uninteracted one-treatment-at-a-time regression

$$Y_i = \ddot{\alpha}_k + X_{ik}\ddot{\beta}_k + \mathbf{W}_i'\ddot{\boldsymbol{\gamma}}_k + \ddot{U}_{ik}, \tag{24}$$

where we only use observations assigned either to treatment k or the control group.

The third solution is to estimate the CW scheme λ^{CW} from Corollary 2. We use inverse propensity score weighting in our applications below: we regress Y_i onto $\mathbf{X}_i$ and a constant, weighting each observation by $\hat{\lambda}^{CW}(\mathbf{W}_i)/\hat{p}_{D_i}(\mathbf{W}_i)$ where $\hat{p}_k(\mathbf{W}_i)$ denotes estimated propensity scores from a multinomial logit model and

$$\hat{\lambda}^{CW}(\mathbf{W}_i) = \left[\sum_{k=0}^{K} \frac{\pi_k(1-\pi_k)}{\hat{p}_k(\mathbf{W}_i)}\right]^{-1} \tag{25}$$

is an estimate of λ^{CW}. When the weights π are uniform, this estimator reduces to the estimator studied in Li and Li (2019). The resulting estimator can be written as

$$\begin{aligned}\hat{\beta}_{\hat{\lambda}^{CW},k} = {} & \frac{1}{\sum_{i=1}^{N} \frac{\hat{\lambda}^{CW}(\mathbf{W}_i)}{\hat{p}_k(\mathbf{W}_i)} X_{ik}} \sum_{i=1}^{N} \frac{\hat{\lambda}^{CW}(\mathbf{W}_i)}{\hat{p}_k(\mathbf{W}_i)} X_{ik} Y_i \\ & - \frac{1}{\sum_{i=1}^{N} \frac{\hat{\lambda}^{CW}(\mathbf{W}_i)}{\hat{p}_0(\mathbf{W}_i)} X_{i0}} \sum_{i=1}^{N} \frac{\hat{\lambda}^{CW}(\mathbf{W}_i)}{\hat{p}_0(\mathbf{W}_i)} X_{i0} Y_i.\end{aligned} \tag{26}$$

When the treatment is binary and $\hat{p}_k$ is obtained via a linear regression, this weighted regression estimator coincides with the usual (unweighted) regression estimator that regresses Y_i onto D_i and $\mathbf{W}_i$.[24] Proposition A.1 in online Appendix A shows that the estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}^{CW}}$ is efficient in the sense that it achieves the semiparametric efficiency bound for estimating $\boldsymbol{\beta}_{\lambda^{CW}} = \sum_i \lambda^{CW}(\mathbf{W}_i)\boldsymbol{\tau}(\mathbf{W}_i)/\sum_i \lambda^{CW}(\mathbf{W}_i)$.

Remark 7: The estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}^{CW}}$ is justified by a parametric model for the propensity score. In order to guard against misspecification of the propensity score, mirroring the discussion in Section IIIA, it may be attractive to instead use a doubly robust version of this estimator that combines propensity score weighting with a regression adjustment using an estimate of μ_k. Another approach is a weighted

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

version of the approach of De los Angeles Resa and Zubizarreta (2020), in which the observations are weighted by $\hat{\lambda}^{CW}$ multiplied by balancing weights (instead of the inverse estimated propensity score).[25] We leave detailed study of these approaches to future research.

Remark 8: Under homoskedasticity, the second and third solutions yield estimates with smaller asymptotic variance than the estimator of the unweighted ATE. These gains in precision are achieved by changing the estimand to a different convex average of conditional treatment effects. In particular, covariate values w where the propensity score $p_k(\mathbf{w})$ is close to zero for some k will be effectively discarded. In practice, explicitly plotting the treatment weights λ^{CW} and λ^{k} may help to identify the types of individuals who are downweighted by these solutions, and to assess the variation in these weights. Plotting them against treatment effect estimates $\hat{\tau}_k$ can help visually assess the extent to which differences in weighting schemes drive differences in between estimates. In particular, the difference between the ATE and any weighted ATE estimand of the effect of treatment k with weights $\lambda(\mathbf{W}_i)$, normalized such that $E\big[\lambda(\mathbf{W}_i)\big] = 1$ is given by $E\big[\lambda(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\big] - E\big[\tau_k(\mathbf{W}_i)\big] = E\big[\lambda(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\big] - E\big[\lambda(\mathbf{W}_i)\big]E\big[\tau_k(\mathbf{W}_i)\big] = \mathrm{cov}\big[\lambda(\mathbf{W}_i), \tau_k(\mathbf{W}_i)\big]$. Thus, if the own treatment weights λ display only a weak covariance with own treatment effect, the weighting will have little effect on the estimand. This is analogous to the observation in Remark 1 that contamination bias reflects the covariance between the contamination weights and treatment effects of the *other* treatments.

IV. Applications

A. *Project STAR Application*

We first illustrate our framework for analyzing and addressing contamination bias with data from Project STAR (Achilles et al. 2008), as studied in Krueger (1999).[26] The Project STAR RCT randomized 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 across 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. We focus on kindergarten effects, where differential attrition and other complications with the experimental analysis are minimal.[27]

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

[25] Under propensity score misspecification, $\hat{\lambda}^{CW}$ would generally converge to a probability limit $\tilde{\lambda}^{CW}$ that may be different from λ^{CW}. Both of these alternative approaches would estimate a weighted average of ATEs weighted by $\tilde{\lambda}^{CW}$ in this case.

[26] Data and code for all empirical results are available at Goldsmith-Pinkham et al. (2024).

[27] 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 probabilities for students entering a participating school in grades 1–3 were different due to uneven availability of slots in small and regular-sized classes Krueger (1999).

TABLE 1—PROJECT STAR CONTAMINATION BIAS AND TREATMENT EFFECT ESTIMATES

	$\hat{\beta}$ (1)	Own (2)	ATE (3)	EW (4)	CW (5)
Panel A. Treatment effect estimates					
Small class size	5.311	5.156	5.515	5.248	5.529
	(0.774)	(0.773)	(0.758)	(0.771)	(0.760)
			[0.740]	[0.739]	[0.738]
Teaching aide	0.205	0.388	0.099	0.292	0.040
	(0.716)	(0.710)	(0.705)	(0.711)	(0.708)
			[0.691]	[0.688]	[0.691]
Number of controls	78				
Sample size	5,902				

		Worst-case bias	
	Bias (1)	Negative (2)	Positive (3)
Panel B. Contamination bias estimates			
Small class size	0.155	−1.643	1.659
	(0.160)	(0.184)	(0.186)
Teaching aide	−0.184	−1.521	1.522
	(0.149)	(0.175)	(0.176)

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

Sources: Achilles et al. (2008); authors' calculations

teaching aide treatments. Specifically, we estimate the partially linear regression (equation (21)) where Y_i is student i's test score achievement at the end of kindergarten, $\mathbf{X}_i = (X_{i1}, X_{i2})$ are indicators for the initial experimental assignment to a small kindergarten class and a regular-sized class with a teaching aide, respectively, and $\mathbf{W}_i$ is a vector of school fixed effects. We follow Krueger (1999) in computing Y_i as the average percentile of student i's math, reading, and word recognition score on the Stanford Achievement Test in the experimental sample. As in the original analysis (Krueger 1999, column 6 of Table V, panel A), we obtain a small class size effect of 5.31 with a heteroskedasticity-robust standard error of 0.77 and a teaching aide effect of 0.21 (standard error: 0.72).[28]

As discussed in Section I, treatment assignment probabilities vary across the schools indicated by the fixed effects in $\mathbf{W}_i$. If treatment effects also vary across schools in a way that covaries with the contamination weights $\lambda_{k\ell}(\mathbf{W}_i)$, we expect

[28] Our estimates are similar to, but not exactly the same as, those in Krueger (1999). The main text reports estimates of an overlap sample that drops one school with no regular classrooms. Full sample estimates, reported in, are nearly identical, but the decomposition in Proposition 1 is not identified in the full sample. We use heteroskedasticity-robust (nonclustered) standard errors throughout this analysis, since the randomization of students to classrooms is at the individual level.

the estimated effect of small class sizes to be partly contaminated by the effect of teaching aides (and vice versa). Panel B reports the contamination bias part of the decomposition in equation (23), which appears minimal for both treatment arms.

It is useful to decompose the contamination bias further into the standard deviation of the school-specific treatment effect $\tau_\ell(\mathbf{W}_i)$, standard deviation of the contamination weights, and their correlation, as discussed in Remark 1. Figure D.2 in online Appendix D does this graphically, plotting estimates of the school-specific treatment effects $\tau_\ell(\mathbf{W}_i)$ against the contamination weights $\lambda_{k\ell}(\mathbf{W}_i)$ for $\ell \neq k$. As can be seen from Figure D.2, the variability of school-specific treatment effects is substantial: Adjusting for estimation error, we estimate the standard deviation of $\tau_k(\mathbf{W}_i)$ to be 10.9 for the small class treatment and of 9.1 for the aide treatment.[29] Both standard deviations are an order of magnitude larger than the standard errors in Table 1. On the other hand, the standard deviations for the contamination weights for the small class and aide treatment are only moderate: 0.14 and 0.11, respectively. Moreover, the correlation between the conditional treatment effects and the contamination weights is weak: 0.10 for the small class effect estimate and -0.13 for the aide effect estimate. The moderate variation in the contamination weights coupled with weak correlation between the weights and the treatment effects explains why the contamination bias is small, even though the treatment effects vary substantially across schools.

Had the experimental design been such that the contamination weights strongly correlate with the treatment effects, sizable contamination bias could have resulted. To illustrate this, we compute worst-case (positive and negative) weighted averages of the estimated $\tau_\ell(\mathbf{W}_i)$ by reordering them across the computed cross-treatment weights $\lambda_{k\ell}(\mathbf{W}_i)$. This exercise highlights potential scenarios in which the randomization strata happened to have been highly correlated with the effect heterogeneity. Columns 2 and 3 in panel B of Table 1 show that both bounds on possible contamination bias are an order of magnitude larger than the actual contamination bias: $[-1.65, 1.67]$ for the small class size treatment and $[-1.53, 1.53]$ for the teaching aide treatment.[30] Overall, for both treatments, the underlying heterogeneity in this setting makes substantial contamination bias possible even though actual contamination bias turns out to be relatively small.

Columns 2–5 of panel A report four treatment effect estimates that are free of contamination bias. Column 2 gives the own-treatment effect component of the decomposition in equation (23), netting out the contamination bias estimate from column 1. This nearly doubles the teaching aide effect estimate, from 0.21 to 0.39, but the estimate remains statistically insignificant; the small classroom estimate moves very little. The remaining columns report the three solutions to contamination bias discussed in Section III. Column 3 estimates the unweighted ATEs of the small class size and teaching aide treatment, by estimating the interacted regression specification in equation (17). Column 4 estimates the one-treatment-at-a-time

[29] 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.

[30] The point estimates and standard errors in columns 4 and 5 in Table 1 do not account for the fact that the reordering is based on estimates of $\tau_k(\mathbf{W}_i)$ rather than the true treatment effects. This biases the reported estimates away from zero, so that they give an upper bound for the worst-case contamination bias.

regressions in equation (24) for $k = 1, 2$. Finally, column 5 runs a weighted regression of Y_i onto $\mathbf{X}_i$ using the CW scheme in equation (25).

There turns out to be little difference between these alternative estimates. The small class size effect varies between 5.2 and 5.5, which is close to the original estimate. The teaching aide effect varies between 0.01 and 0.29. To understand this lack of variation, recall from Remark 8 that the difference between the unweighted ATE and an estimand that uses weights $\lambda(\mathbf{W}_i)$ is given by the covariance between $\lambda(\mathbf{W}_i)$ and the conditional ATEs $\tau_k(\mathbf{W}_i)$. Given the sizable variability in the treatment effect estimates, the covariance will be small only if the correlation between the weights and the treatment effects is small and if the weights display limited variability. This turns out to be the case here, as depicted graphically in Figure D.3 in online Appendix D. The figure shows that the correlations fall below 0.25 in absolute value for all weighting schemes, and that the weights only vary between 0.7 and 1.2.

As a consequence of strong overlap, the standard errors are similar across the columns. Indeed, the efficiency gain of the EW scheme relative to the ATE based on an efficiency bound comparison using equation (18) with $\lambda = \lambda^k$ versus $\lambda = 1$ is less than 1.6 percent for both treatments under homoskedasticity; the gain is even smaller under the CW scheme. The reported standard errors, which allow for heteroskedasticity and don't assume known propensity scores, align with this prediction.[31] As discussed in Remark A.1 in online Appendix A.3, these standard errors are affected by the assumption of known propensity scores, used to derive the weighting schemes underlying the estimates in columns 2 and 3. To gauge the impact of this assumption, we also report a version of the standard errors computed under the assumption that the sample treatment probabilities in each school match the true propensity scores. This changes the standard errors little, showing that there is minimal cost to estimating the weights.

B. *Further Applications*

We next study the broader relevance of contamination bias using data from eight additional studies with multiple-treatment regressions. These studies were identified by a systematic search of papers in the AEA Data and Code Repository from 2013–2022 (see online Appendix C.1 for details). Five studies are experiments like Project STAR; the remaining three use observational regressions to estimate racial disparities across multiple race groups (which we interpret as descriptive, following Remark 4).[32] We replicate a single representative specification for each paper, corresponding to the first relevant regression discussed in the paper's introduction.[33] Table 2 lists the papers and specifications.

[31] The standard errors reported in parentheses in panel B are valid for the population analogs β_k and $\beta_{\lambda^{CW}}$, that is, $E\left[\lambda^k(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\right]/E\left[\lambda^k(\mathbf{W}_i)\right]$ and $E\left[\lambda^{CW}(\mathbf{W}_i)\,\tau_k(\mathbf{W}_i)\right]/E\left[\lambda^{CW}(\mathbf{W}_i)\right]$. Since these standard errors are potentially conservative when viewed as standard errors for β_k and $\beta_{\lambda^{CW}}$, the standard error comparison gives an upper bound on the cost to estimating the weights.

[32] We focus on observational studies of racial disparities as they often include regressions on multiple minority race "treatments," use publicly available data, and are easily identifiable by a keyword search.

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

Table 2—Further Applications

Paper	Journal (1)	Type (2)	Spec. (3)	Sample size: Original (4)	Sample size: Overlap (5)	std($\hat{p}(\mathbf{W})$) (6)
Benhassine et al. (2015)	AEJ:AE	Exp.	5(1)	11,074	6,996	0.14
Cole et al. (2013)	AEJ:AE	Exp.	7(6)	132	73	0.10
de Mel et al. (2013)	AEJ:AE	Exp.	2(2)	520	520	0.02
Drexler et al. (2014)	AEJ:AE	Exp.	2(2)	796	796	0.05
Duflo et al. (2015)	AER	Exp.	2A(1)	9,116	8,664	0.11
Fryer and Levitt (2013)	AER	Obs.	3(4)	8,806	6,623	0.31
Rim et al. (2020a)	AER:P&P	Obs.	2(3)	4,037	620	0.24
Weisburst (2019a)	AER:P&P	Obs.	2A	7,488	7,488	0.31

Notes: This table summarizes the five experimental studies and three observational studies of racial disparities collected from a search of the AEA Data and Code Repository from 2013–2022 (See online Appendix C.1 for details of this search). Column 3 reports the table and panel of the replicated specification with the column or row of the specification in parentheses. Column 6 gives the standard deviation of the estimated propensity score $\hat{p}_k(\mathbf{W}_i)$ for the treatment arm k displaying the greatest propensity score variation; estimates are computed using a multinomial logit model. See online Appendix C.2 for details on the overlap sample and tests for propensity score variation.

Sources: Benhassine et al. (2024); Cole et al. (2019); de Mel, McKenzie, and Woodruff (2019); Drexler, Fischer, and Schoar (2019); Duflo, Dupas, and Kremer (2019); Fryer and Levitt (2019), Rim, Ba, and Rivera (2020b); Weisburst (2019b); authors' calculations

We conduct two preliminary analyses of each study before assessing contamination bias and comparing alternative estimators. First, we ensure that the estimation sample satisfies overlap, since otherwise the decomposition in Proposition 1 is typically not identified. If strong overlap fails, we identify a large subset of each analysis sample where it is satisfied. Columns 4 and 5 of Table 2 list the number of observations in the full and overlap samples (the sample sizes are equal if the original estimation sample satisfies overlap). Second, we check for propensity score variation in each of the studies. In principle, protocol descriptions can reveal whether some regression controls are necessary (and hence generate propensity score variation) or whether the controls are just added to improve precision. In practice, however, this is not always clear from paper descriptions.[34] Column 6 of Table 2 gives a quantitative sense of the variability in the propensity scores by reporting the standard deviation of the estimated propensity score, showing that its variability in the observational studies is substantially higher; the dagger symbol indicates that a hypothesis test for nonzero variation in the population propensity scores was statistically significant. online Appendix C.2 details the overlap sample construction and these tests. We replicate the analyses from Table 1 for each of the eight papers in online Appendix C.3; we summarize the takeaways here.

Figure 1 summarizes the statistical and practical significance of contamination bias in the estimated effect of each treatment for each specification (as estimated in the overlap sample). Column A shows the absolute value of the contamination bias t -statistics for each regression coefficient, obtained from the decomposition in equation (23). In both columns, we sort treatments within papers by this absolute t-statistic

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

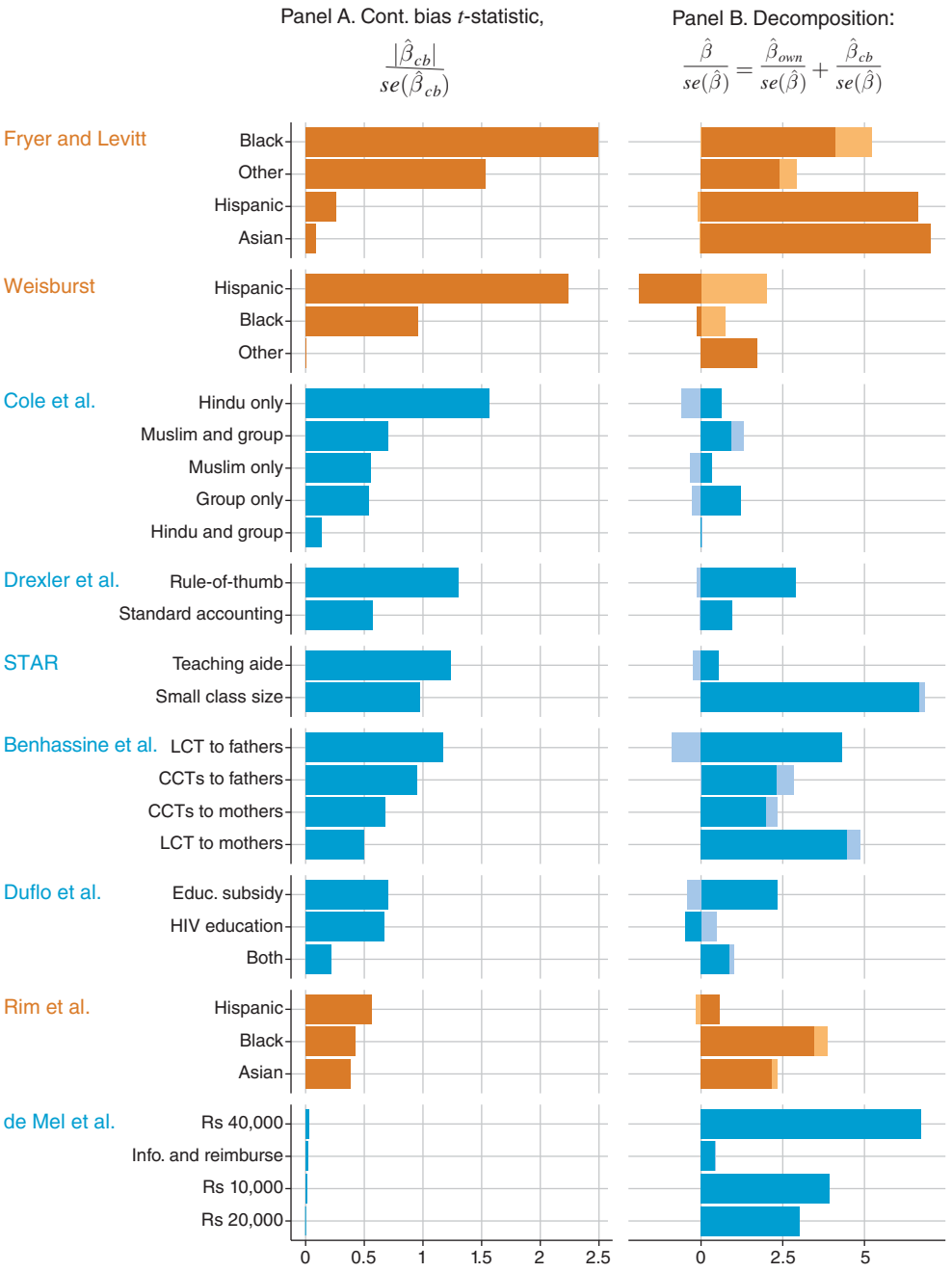

FIGURE 1. CONTAMINATION BIAS ACROSS ALL APPLICATIONS

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

Sources: Achilles et al. (2008); Benhassine et al. (2024); Cole et al. (2019); de Mel, McKenzie, and Woodruff (2019); Drexler, Fischer, and Schoar (2019); Duflo, Dupas, and Kremer (2019); Fryer and Levitt (2019); Rim, Ba, and Rivera (2020b); Weisburst (2019b); authors' calculations

and sort papers by the maximum absolute *t*-statistic across treatments. Column B shows a normalized version of the decomposition that divides each term by the standard error of the regression coefficient. The darker bar shows the own-treatment effect component of the decomposition, while the lighter bar denotes the contamination bias component (which can be of the same or opposite sign).

The figure shows economically and statistically meaningful contamination bias in two of the three observational studies while showing no evidence for bias in any of the experimental studies. This aligns with the intuition that the large propensity score variability in observational studies generates much larger variability in the contamination weights. Specifications from both the de Mel, McKenzie, and Woodruff (2013) and Drexler, Fischer, and Schoar (2014) experiments have some of the smallest contamination bias and also smallest propensity score variation, consistent with the theoretical results that contamination bias requires variation in the contamination weights which in turn requires variation in the propensity scores. On the other hand, the two studies with statistically significant contamination bias (Fryer and Levitt 2013 and Weisburst 2019a) also display the greatest variation in propensity scores. These results highlight the importance of testing for contamination bias, especially in observational settings where the included covariates are likely to drive sizable variation in propensity scores and hence contamination weights.

Figure 2 plots estimates of the treatment effects for each estimator from Table 1, again normalizing by the standard error of the regression coefficient. We include a line between the estimates from OLS regression and from the CW estimator we propose. Among observational studies, we see substantial variation across the different estimates and a much larger difference between the OLS estimator and the CW estimator. In the experimental papers, the difference is much smaller.[35] This is consistent with the larger propensity score variability in observational studies magnifying the impact of the choice of weighting scheme.

V. Conclusion

Regressions with multiple treatments and flexible controls are common across a wide range of empirical settings in economics. We show that such regressions generally fail to estimate a convex weighted average of treatment effects: coefficients on each treatment are 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. We then discuss three alternative estimators that are free of this bias.

Our analysis of nine empirical applications finds economically and statistically meaningful contamination bias in observational studies. Contamination bias in experimental studies is more limited, even in papers that display statistically significant variation in the propensity scores. We also find that the choice among alternative estimators that are free of contamination bias matters more in the

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

FIGURE 2. TREATMENT EFFECT ESTIMATES WITH USING DIFFERENT ESTIMATORS

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

Sources: Achilles et al. (2008); Benhassine et al. (2024); Cole et al. (2019); de Mel, McKenzie, and Woodruff (2019); Drexler et al. (2019), Duflo, Dupas, and Kremer (2019); Fryer and Levitt (2019); Rim, Ba, and Rivera. (2020b); Weisburst (2019b); authors' calculations

observational studies. Overall, our analysis highlights the importance of testing the empirical relevance of theoretical concerns with how regression combines heterogeneous effects, particularly in observational studies.

REFERENCES

Abadie, Alberto, and Matias D. Cattaneo. 2018. "Econometric Methods for Program Evaluation." *Annual Review of Economics* 10 (1): 465–503.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–96.

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc. 2021. "Mortality Effects and Choice Across Private Health Insurance Plans." *Quarterly Journal of Economics* 136 (3): 1557–1610.

Achilles, C. M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. *Tennessee's Student Teacher Achievement Ratio* (*STAR*) *project* (*Version V1*). Harvard Dataverse. https://doi.org/10.7910/DVN/SIWH9F.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871–919.

Angrist, Joshua, Peter Hull, Parag A. Pathak, and Christopher Walters. 2024. "Credible School Value-Added with Undersubscribed School Lotteries." *Review of Economics and Statistics* 106 (1): 1–19.

Angrist, Joshua D., Guido W. Imbens, and Alan Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14 (1): 57–67.

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

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

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60 (1): 250–67.

Athey, Susan, and Guido W. Imbens. 2022. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics* 226 (1): 62–79.

Behaghel, Luc, Bruno Crépon, and Marc Gurgand. 2013. "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial." Unpublished.

Benhassine, Najy, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. 2015. "Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education." *American Economic Journal: Economic Policy* 7 (3): 86–125.

Benhassine, Najy, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. 2024. *Replication data for: "Turning a Shove Into a Nudge? A "Labeled Cash Transfer" for Education* (Version V2)." Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E114579V2co.

Bhuller, Manudeep, and Henrik Sigstad. 2024. "2SLS with Multiple Treatments." *Journal of Econometrics* 242 (1): 105785.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When is TSLS actually LATE?" Unpublished.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *Review of Economic Studies*. https://doi.org/10.1093/restud/rdae007.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–30.

Cattaneo, Matias D. 2010. "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability." *Journal of Econometrics* 155 (2): 138–54.

Chattopadhyay, Ambarish, and Jose R. Zubizarreta. 2021. "On the Implied Weights of Linear Regression for Causal Inference." Unpublished.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2008. "Semiparametric Efficiency in GMM Models with Auxiliary Data." *Annals of Statistics* 36 (2): 808–43.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–C68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4): 1501–35.

Chernozhukov, Victor, Whitney K. Newey, and Rahul Singh. 2022. "Automatic Debiased Machine Learning of Causal and Structural Effects." *Econometrica* 90 (3): 967–1027.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–2632.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickery. 2013. "Barriers to Household Risk Management: Evidence from India." *American Economic Journal: Applied Economics* 5 (1): 104–35.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickery. 2019. *Replication data for: "Barriers to Household Risk Management: Evidence from India.* (Version V1)."Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E116379V1.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand." NBER Working Paper 0330.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2023. "Two-Way Fixed Effects and Differences- in-Differences Estimators with Several Treatments." *Journal of Econometrics* 236 (2): 105480.

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

de los Angeles Resa, María, and José R. Zubizarreta. 2020. "Direct and Stable Weight Adjustment in Non-Experimental Studies with Multivalued Treatments: Analysis of the Effect of an Earthquake on Post-Traumatic Stress." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183 (4): 1387–1410.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2013. "The Demand For, and Consequences of, Formalization Among Informal Firms in Sri Lanka." *American Economic Journal: Applied Economics* 5 (2): 122–50.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2019. *Replication data for: "The Demand For, and Consequences of, Formalization Among Informal Firms in Sri Lanka* (Version V1)." Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E113847V1.

Dobbie, Will, and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection." *American Economic Review* 105 (3): 1272–1311.

Drexler, Alejandro, Greg Fischer, and Antoinette Schoar. 2014. "Keeping it Simple: Financial Literacy and Rules of Thumb." *American Economic Journal: Applied Economics* 6 (2): 1–31.

Drexler, Alejandro, Greg Fischer, and Antoinette Schoar. 2019. *Replication data for: "Keeping it simple: Financial Literacy and Rules of Thumb* (Version V1)." Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E113888V1.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "Education, HIV, and early Fertility: Experimental Evidence from Kenya." *American Economic Review* 105 (9): 2757–97.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2019. Replication data for: "*Education, HIV, and Early Fertility: Experimental Evidence from Kenya* (Version V1)."Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E112899V1.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113 (1): 253–77.

Fryer, Roland G., and Steven D. Levitt. 2013. "Testing for Racial Differences in the Mental Ability of Young Children." *American Economic Review* 103 (2): 981–1005.

Fryer, Roland G., and Steven D. Levitt. 2019. *Replication data for: "Testing for racial differences in the mental ability of young children* (*Version V1*)."Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E112609V1.

Geruso, Michael, Timothy J. Layton, and Jacob Wallace. 2020. "What Difference Does a Health Plan Make? Evidence from Random Plan Assignment in Medicaid." NBER Working Paper 27762.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. *Data and Code for: "Contamination Bias in Linear Regressions."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E207983V1.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Graham, Bryan S., and Cristine Campos de Xavier Pinto. 2022. "Semiparametrically Efficient Estimation of the Average Linear Regression Function." *Journal of Econometrics* 226 (1): 115–38.

Hull, Peter D. 2018a. "Isolateing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons." https://dx.doi.org/10.2139/ssrn.2705108.

Hull, Peter D. 2018b. "Estimating Treatment Effects in Mover Designs." https://doi.org/10.48550/arXiv.1804.06721.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Imbens, Guido W., Whitney Newey, and Geert Ridder. 2007. "Mean-Squared-Error Calculations for Average Treatment Effects." Unpublished.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

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

Khan, Shakeeb, and Elie Tamer. 2010. "Irregular Identification, Support Conditions, and Inverse Weight Estimation." *Econometrica* 78 (6): 2021–42.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." *Quarterly Journal of Economics* 131 (3): 1057–1111.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Quarterly Journal of Economics* 131 (4): 1795–1848.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Lee, Sokbae, and Bernard Salanié. 2018. "Identifying Effects of Multivalued Treatments." *Econometrica* 86 (6): 1939–63.

Li, Fan, and Fan Li. 2019. "Propensity Score Weighting for Causal Inference with Multiple Treatments." *Annals of Applied Statistics* 13 (4): 2389–2415.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Mestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review* 103 (5): 1797–1829.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, and M. Susan Marquis. 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review* 77 (3): 251–77.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." *Econometrica* 86 (5): 1589–1619.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2021. "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables." *American Economic Review* 111 (11): 3663–98.

Mountjoy, Jack, and Brent R. Hickman. 2021. "The Returns to College(S): Estimating Value-Added and Match Effects in Higher Education." NBER Working Paper 29276.

Mueller-Smith, Michael. 2015. "The Criminal and Labor Market Impacts of Incarceration." Unpublished.

Norris, Samuel. 2019. "Examiner Inconsistency: Evidence from Refugee Appeals." Unpublished.

Rim, Nayoung, Bocar Ba, and Roman Rivera. 2020a. "Disparities in Police Award Nominations: Evidence from Chicago." *AEA Papers and Proceedings* 110: 447–51.

Rim, Nayoung, Bocar Ba, and Roman Rivera. 2020b. *Replication data for: "Disparities In Police Award Nominations: Evidence from Chicago* (*Version V1*)."Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E120749V1.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights are Highly Variable." *Statistical Science* 22 (4): 544–59.

Robins, James M., Steven D. Mark, and Whitney K. Newey. 1992. "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders." *Biometrics* 48 (2): 479–95.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors are not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–54.

Słoczyński, Tymon, and Jeffrey M. Wooldridge. 2018. "A General Double Robustness Result for Estimating Average Treatment Effects." *Econometric Theory* 34 (1): 112–33.

Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science* 32 (4): 561–79.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.

van der Vaart, A. W. 1998. Asymptotic Statistics. Cambridge: Cambridge University Press.

Weisburst, Emily K. 2019a. "Police Use of Force as an Extension of Arrests: Examining Disparities across Civilian and Officer Race." *AEA Papers and Proceedings* 109: 152–56.

Weisburst, Emily K. 2019b. *Replication data for: "Police Use of Force as an Extension of Arrests: Examining Disparities across Civilian and Officer Race* (Version V1)."Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E114511V1.

Wooldridge, Jeffrey M. 2007. "Inverse Probability Weighted Estimation for General Missing Data Problems." *Journal of Econometrics* 141 (2): 1281–1301.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, The Two-Way Mundlak Regression, and Difference-in-Differences Estimators." Unpublished.

Yang, Shu, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. 2016. "Propensity Score Matching and Subclassification in Observational Studies with Multi-Level Treatments." *Biometrics* 72 (4): 1055–65.

Zhou, Xiang, and Aleksei Opacic. 2022. "Marginal Interventional Effects." Unpublished.