Supplemental Appendix for "Contamination Bias in Linear Regressions"

Paul Goldsmith-Pinkham
Yale University

Peter Hull
Brown University

Michal Kolesár
Princeton University

July 16, 2024

Appendix A collects all proofs and extensions. Appendix B discusses the connection between our contamination bias characterization and that in the difference-in-differences (DiD) literature. Details on the applications and additional exhibits are given in Appendices C and D.

Appendix A Proofs and Additional Results

A.1 Proof of Proposition 1

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

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

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

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

First note that by iterated expectations and conditional mean-independence, $E[\tilde{\tilde{X}}_{ik} Y_i] =$

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

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

This proves eq. (15).

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

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

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

Finally, if X_i are mutually exclusive treatment indicators, write $E^*[X_{ik} \mid X_{i,-k}, W_i] = X_{i,-k}'\tilde{\delta}_k + \tilde{g}_k(W_i)$. Since $X_{ik}X_{i,-k} = 0$, we may write

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

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

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

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

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

Note $p_0(W_i) = 1 - \sum_{k=1}^K p_k(W_i)$ and $p_k(W_i)p_{-k}(W_i) = v_{-k}(W_i)\iota - p_0(W_i)p_{-k}(W_i)$, where ι

denotes a $(K-1)$-vector of ones. Thus, the numerator can be written as

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

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

A.2 Proof of Proposition 2

The parameter of interest $\theta_{\lambda,c}$ depends on the realizations of the controls. We therefore derive the semiparametric efficiency bound conditional on the controls; i.e. we show that eq. (18) is almost-surely the variance bound for estimators that are regular conditional on the controls. Relative to the earlier results in Hahn (1998) and Hirano et al. (2003), we need to account for the fact that the data are no longer i.i.d. once we condition on the controls.

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

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

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

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

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

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

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

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

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

Since the scores $\Delta_{N,s} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} s(Y_i, D_i \mid w_i)$ are exactly linear in s, condition (3.12) also holds. It follows that the model is LAN on $\mathcal{E}$.

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

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

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

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

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

A.3 Efficiency of the CW estimator

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

Proposition A.1. *Suppose eq.* (11) *holds in an i.i.d. sample of size N, with known non-degenerate propensity scores $p_k(W_i)$. Let $\beta^*_{\lambda^{\mathrm{CW}},k} = E[\lambda^{\mathrm{CW}}(W_i)\tau_k(W_i)]/E[\lambda^{\mathrm{CW}}(W_i)]$, and $\alpha^*_k = \beta^*_{\lambda^{\mathrm{CW}},k} + E[\lambda^{\mathrm{CW}}(W_i)\mu_0(W_i)]/E[\lambda^{\mathrm{CW}}(W_i)]$. Suppose that the fourth moments of $\lambda^{\mathrm{CW}}(W_i)$ and $\mu(W_i)$ are bounded, and that $p_k \in \mathcal{G}$, $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^{\mathrm{CW}}(W_i)^2}{p_{k'}(W_i)^2} \in \mathcal{G}$, and $(\mu_k(W_i) - \alpha^*_k)\frac{\lambda^{\mathrm{CW}}(W_i)}{p_k(W_i)} \in \mathcal{G}$ for all k, k'. Then, provided it is asymptotically linear and regular, $\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}}$ achieves the semiparametric efficiency bound for estimating $\beta_{\lambda^{\mathrm{CW}}}$, with diagonal elements of its asymptotic variance of:*

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

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

Remark A.1. The asymptotic variance of the estimator $\hat{\beta}_{\lambda^{\mathrm{CW}}}$ is larger than the asymptotic variance of the infeasible estimator that replaces the estimated weights $\hat{\lambda}^{\mathrm{CW}}(W_i)/\hat{p}_{D_i}(W_i)$ in eq. (26) with the infeasible weights $\lambda^{\mathrm{CW}}(W_i)/p_{D_i}(W_i)$. The latter achieves the asymptotic variance implied by Corollary 2,

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

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

A.4 Proof of Proposition A.1

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

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

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

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

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

In line with this result, we define the tangent space to consist of all functions $s(y, d \mid w) = \sum_{k=0}^{K} \mathbb{1}\{d = k\}(s_{y,k}(y \mid w) + s_{p,k}(w))$ such that $s_{y,k}$ and $s_{p,k}$ satisfy the above restrictions. Define the inner product on this space by the marginal expectation $\langle s_1, s_2 \rangle = E[s_1(Y_i, D_i \mid W_i)s_2(Y_i, D_i \mid W_i)]$. We will prove the efficiency bound on the event $\mathcal{E}$ that (i) $\frac{1}{N}\sum_{i=1}^{N} E[s(Y_i, D_i \mid W_i)^2 \mid W_i] \to E[s(Y, D_i \mid W_i)^2]$; (ii) $N^{-1}\sum_i \lambda^{\mathrm{CW}}(W_i) \to E[\lambda^{\mathrm{CW}}(W_i)]$; (iii) $N^{-1}\sum_i \lambda^{\mathrm{CW}}(W_i)\sum_{k=0}^{K} c_k E[Y_i(k) \cdot s_{y,k}(Y_i \mid W_i) \mid W_i] \to \sum_{k=0}^{K} c_k E[\lambda^{\mathrm{CW}}(W_i)Y_i(k) \cdot s_{y,k}(Y_i(k) \mid W_i)]$; (iv) $N^{-1}\sum_{i=1}^{N} \lambda^{\mathrm{CW}}(W_i)^2 \sum_{k,k'} c_{k'}\mu_{k'}(W_i)\frac{s_{p,k}(W_i)}{p_k(W_i)} \to E[\lambda^{\mathrm{CW}}(W_i)^2 \cdot \sum_{k,k'} c_{k'} \cdot \mu_{k'}(W_i)\frac{s_{p,k}(W_i)}{p_k(W_i)}]$; (v) $N^{-1}\sum_{i=1}^{N} \lambda^{\mathrm{CW}}(W_i)^2 \sum_{k=0}^{K} \frac{s_{p,k}(W_i)}{p_k(W_i)} \to E[\lambda^{\mathrm{CW}}(W_i)^2 \sum_{k=0}^{K} \frac{s_{p,k}(W_i)}{p_k(W_i)}]$; and (vi) $\beta_{\lambda^{\mathrm{CW}}} \to \beta^*_{\lambda^{\mathrm{CW}}}$. Under the proposition assumptions and the law of large numbers, $\mathcal{E}$ is a probability-one set.

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

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

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

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

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

where $\lambda_i^{\mathrm{CW}} = \lambda^{\mathrm{CW}}(W_i)$, the limit on the second line holds on the event $\mathcal{E}$, and the third line uses $E[X_{ki}(Y_i - \mu_k(W_i)) s(Y_i, D_i \mid W_i) \mid W_i] = p_k(W_i) E[Y_i(k) s_{y,k}(Y_i(k) \mid W_i) \mid W_i]$ and $E[X_{ki} s(Y_i, D_i \mid W_i) \mid W_i] = p_k(W_i) s_{p,k}(W_i)$. Since for any function $a(W_i)$, $E[a(W_i) s(Y_i, D_i \mid W_i)] = 0$, subtracting $\frac{1}{E[\lambda_i^{\mathrm{CW}}]} \sum_{k=0}^{K} E[(\lambda_i^{\mathrm{CW}})^2 \frac{\tilde{h}(W_i)}{p_k(W_i)} s(Y_i, D_i \mid W_i)] = 0$ from the preceding display implies $\sqrt{N}(\beta_N(\theta(\cdot \mid \cdot; 1/\sqrt{N})) - \beta_N(\theta)) = E[\psi(Y_i, D_i, W_i) s(Y_i, D_i \mid W_i)] + o(1)$, where

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

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

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

$$\sqrt{N}(\beta_{\lambda^{\mathrm{CW}}} - \beta^*_{\lambda^{\mathrm{CW}}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi^*(W_i) + o_p(1) \text{ where } \psi^*(W_i) = \frac{\lambda_i^{\mathrm{CW}}}{E[\lambda_i^{\mathrm{CW}}]} (\tau(W_i) - \beta^*_{\lambda^{\mathrm{CW}}}). \quad \text{(A.2)}$$

Second, we show that

$$\sqrt{N}(\hat{\beta}_{\hat{\lambda}^{\mathrm{CW}}} - \beta^*_{\lambda^{\mathrm{CW}}}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Y_i, D_i, W_i) + o_p(1), \quad \text{(A.3)}$$

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

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

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

Equation (A.2) follows directly under the assumptions of the proposition by the law of large numbers and the fact that the variance of $\lambda_i^{\text{CW}}(\tau(W_i) - \beta^*_{\lambda^{\text{CW}}})$ is bounded. To show eq. (A.3), write $\hat{\beta}_{\hat{\lambda}^{\text{CW}},k} = \hat{\alpha}_k - \hat{\alpha}_0$, where $\hat{\alpha}$ is a two-step method of moments estimator based on the $(K+1)$ dimensional moment condition $E[m(Y_i, D_i, W_i, \alpha^*, p)] = 0$ with elements $m_k(Y_i, D_i, W_i, \alpha^*, p) = \lambda_i^{\text{CW}} \frac{X_{ki}}{p_k(W_i)}(Y_i - \alpha_k^*)$, and α^* is a $(K+1)$ dimensional vector with elements $\alpha_k^* = E[\lambda_i^{\text{CW}} \mu_k(W_i)]/E[\lambda_i^{\text{CW}}]$.

Consider a one-dimensional path F_t such that the distribution of the data is given by F_0. Let $p_{k,t}(W_i) = E_{F_t}[X_{ki} \mid W_i]$ denote the propensity score along this path. The derivative of $E[m_k(Y_i, D_i, W_i, \alpha^*, p_t)]$ with respect to t evaluated at $t = 0$ is

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

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

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

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

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

which yields eq. (A.3).

Appendix B Connections to the DiD Literature

In this appendix we elaborate on the connections between Proposition 1 and the recent literature studying potential biases from heterogeneous treatment effects in DiD regressions and related specifications (e.g. Goodman-Bacon, 2021; Sun & Abraham, 2021; Hull, 2018; de Chaisemartin & D'Haultfœuille, 2020; De Chaisemartin & D'Haultfœuille, 2023; Callaway & Sant'Anna, 2021; Borusyak et al., 2024; Wooldridge, 2021). We first show how our framework fits a two-way fixed effects (TWFE) regression with a general treatment specification. We then show how Proposition 1 applies to three particular specifications: a static binary treatment, a

dynamic "event study" treatment, and a static multivalued treatment (or "movers regression"). In each case we discuss whether there is a potential for bias—either contamination bias or own-treatment negative weighting—and give a numerical illustration.

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

To study these specifications, we follow de Chaisemartin and D'Haultfœuille (2020) and Borusyak et al. (2024) in considering the n observed units as fixed, and we condition on their treatment status (results when the units are sampled from a large population are analogous). For each unit j, we observe a random T-vector of outcomes $Y_j = (Y_{j1}, \ldots, Y_{jT})$ and a fixed T-vector of $(\mathcal{K}+1)$-valued treatments $\mathcal{D}_j = (\mathcal{D}_{j1}, \ldots, \mathcal{D}_{jT})$. These treatments are used to construct a vector of $(K+1)$-valued "treatments states" $D_j = (D_{j1}, \ldots, D_{jT})$, with $D_{jt} \in \{0, \ldots, K\}$. Setting $D_j = \mathcal{D}_j$ covers scenarios with static treatments; as we show below, other choices of D_j allows us to cover scenarios with dynamic treatment effects. As in the main text, X_{jt} denotes a K-vector of treatment status indicators derived from D_{jt}.

We make two assumptions. First, we assume that potential outcomes $Y_{jt}(d_t)$ depend on the T-vector of treatments only through the current value d_t of the treatment state, such that $Y_{jt} = Y_{jt}(D_{jt})$.[B.1] Second, we make a parallel trends assumption by writing the untreated potential outcomes as

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

for fixed α_j and λ_t, and assuming

$$E[\eta_{jt}] = 0. \tag{B.1}$$

Together these expressions imply $E[Y_{jt}(0)] = \alpha_j + \lambda_t$, which is how parallel trends is sometimes formalized (c.f. Assumption 1 in Borusyak et al. (2024); weaker versions of the parallel trends assumption yield analogous results). We do not restrict the dependence of η_{jt} across units or time, nor do we make restrictions on the potentially random treatment effects $\tau_{jt,k} = Y_{jt}(k) - Y_{jt}(0)$. Collecting these effects in a vector τ_{jt}, we have

$$Y_{jt} = X'_{jt}\tau_{jt} + \alpha_j + \lambda_t + \eta_{jt}. \tag{B.2}$$

This outcome model reduces to a conventional TWFE model under the assumption of constant

[B.1] This assumption rules out misspecification of the treatment states, such as when there are dynamic effects but $D_{jt} = \mathcal{D}_{jt}$ only indexes contemporaneous treatment status, as noted in footnote 9.

treatment effects: $\tau_{jt} = \beta$ for all (j,t).

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

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

$$\begin{aligned} \beta &= \left(\sum_{j=1}^{n}\sum_{t=1}^{T} E_{\tau,\eta}[\tilde{X}_{jt}\tilde{X}_{jt}']\right)^{-1} \sum_{j=1}^{n}\sum_{t=1}^{T} E_{\tau,\eta}[\tilde{X}_{jt}Y_{jt}] \\ &= \left(\sum_{j=1}^{n}\sum_{t=1}^{T} \tilde{X}_{jt}\tilde{X}_{jt}'\right)^{-1} \sum_{j=1}^{n}\sum_{t=1}^{T} \tilde{X}_{jt}X_{jt}'E[\tau_{jt}], \end{aligned} \tag{B.3}$$

where the second equality uses eqs. (B.1) and (B.2), and the fact that only η_{jt} and τ_{jt} are stochastic. Proposition 1 implies that the coefficient on the kth element on X_{jt} is given by

$$\beta_k = \sum_{j,t} \lambda_{kk}(j,t)E[\tau_{jt,k}] + \sum_{\ell \neq k}\sum_{j,t} \lambda_{k\ell}(j,t)E[\tau_{jt,\ell}] \tag{B.4}$$

where

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

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

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

Static binary treatment Consider a DiD setting where units adopt (and potentially drop) a binary treatment at different time periods—as studied by de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021). For example, different states j may choose to roll out a policy in different years and a researcher wishes to estimate the average effect of this policy using this staggered adoption. We assume that the treatment is static,

setting $D_{jt} = \mathcal{D}_{jt}$, with $K = \mathcal{K} = 1$. Since the treatment is binary, $X_{jt} = D_{jt}$ is a scalar with $\tilde{\tilde{X}}_{jt,1} = \tilde{X}_{jt}$, and the second term in eq. (B.4) drops; the weights on the first term simplify to

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

which coincides with the expression in Theorem 1 of de Chaisemartin and D'Haultfœuille (2020). These treatment weights are not guaranteed to be convex since eq. (12) does not hold.[B.2] In contrast, Athey and Imbens (2022) consider staggered DiD regressions where eq. (12) holds because intervention timing is assumed to be random (in place of the parallel trends assumption). Under this design-based assumption, Proposition 1 shows the treatment weights (corresponding to those in Theorem 1(iv) of Athey and Imbens (2022)) are convex.

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

$$1 - \max_j \overline{X}_j - \overline{X}_T + \overline{X} \geq 0. \tag{B.5}$$

Condition (B.5) holds in the canonical DiD case with a single intervention date, where the first $n_1 < n$ units treated in the last $T_1 < T$ periods and untreated in the earlier periods $1,\ldots,T-T_1$. The remaining units are never treated, so that $D_{jt} = \mathcal{D}_{jt} = \mathbb{1}\{j \leq n_1, t \geq T - T_1\}$. This nests the simplest DiD specification where $T = 2$ and $T_1 = 1$. In this case, when units in the treatment group are treated, $1-\overline{X}_j-\overline{X}_t+\overline{X} = (1-n_1/n)(1-T_1/T)$ so that the weights $\lambda_{11}(j,t)$ are non-negative, and eq. (B.4) simplifies to:

$$\beta_1 = \sum_{j,t}\lambda_{11}(j,t)E[\tau_{jt,1}], \qquad \lambda_{11}(j,t) = \frac{(1-\frac{n_1}{n})(1-\frac{T_1}{T})X_{jt}}{(1-\frac{n_1}{n})(1-\frac{T_1}{T})\frac{n_1T_1}{nT}} = \frac{X_{jt}}{n_1T_1/N},$$

which is simply the average treatment effect for the n_1T_1 treated observations.

However, in presence of multiple treatment adoption dates, eq. (B.5) may fail. To illustrate, consider a case with three time periods ($T = 3$) and three groups of units: $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{N}$, with respective sizes n_E, n_L, and n_N. Units $j \in \mathcal{E}$ are "early adopters", and are treated beginning in period 2. Units $j \in \mathcal{L}$ are "late adopters", and are treated only in period 3. Units in the last group are never treated.[B.3] In this case, eq. (B.5) simplifies to $1-2/3-(n_E+n_L)/n+(2/3n_E+1/3n_L)/n = (n_N-n_L)/3n$, which is negative if there are more late adopters than

[B.2] Since $E[X_{jt} \mid W_{jt}] = X_{jt} \in \{0,1\}$, if eq. (12) held, then the residual $\tilde{X}_{jt}$ must be zero (this is true if, e.g., all units have the same treatment adoption date). But that would generate a multicollinearity issue, precluding the researcher from including unit and time effects in the regression.

[B.3] This example is a special case of the example discussed in Figure 2 of Goodman-Bacon (2021).

never adopters; otherwise, if $n_L < n_N$, all weights are positive. Indeed, some algebra shows

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

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

Condition (B.5) is generally quite restrictive. Consider, for instance, a setting in which no units are treated in the first period and a fraction $1/T$ of observations adopts the treatment in period $t = 2, \ldots, T$. Then for the group adopting treatment in period 2, eq. (B.5) becomes $(3-T)/2T$, which is negative if $T \geq 4$. Similarly, condition (B.5) fails if there exists an always-treated group, or if everyone is treated in the last period.

Dynamic binary treatment with staggered adoption Next, consider an "event study" setting in which each unit j starts being treated in period $A(j) \in \{1, 2, \ldots, T\} \cup \infty$ and remains treated thereafter, with $A(j) = \infty$ denoting a unit that is never treated. Thus, $\mathcal{D}_{jt} = \mathbb{1}\{t > A(j)\}$, with $\mathcal{K} = 1$. Unlike in previous cases, we allow for dynamic effects by letting $D_{jt} = t - A(j)$ index the number of periods since the treatment adoption date (breaking with our usual indexing convention of $D_{jt} \geq 0$), assuming no anticipation effect one period before adoption, and correspondingly normalizing $D_{jt} = -1$ for the never-treated group. X_{jt} then consists of indicators for all leads and lags relative to the adoption date: $X_{jt} = (\mathbb{1}\{D_{jt} = -(T-1)\}, \ldots, \mathbb{1}\{D_{jt} = -2\}, \mathbb{1}\{D_{jt} = 0\}, \ldots, \mathbb{1}\{D_{jt} = T-1\})'$, with the indicator for the period just prior to adoption ($D_{jt} = -1$) excluded. This specification avoids perfect collinearity when all treatment adoption dates are represented in the data (including the never-treated group). Sun and Abraham (2021) and Borusyak et al. (2024) study such "fully-dynamic" event study specifications.

Since X_{jt} is now a vector with $K = 2(T-1)$, the second contamination bias term in eq. (B.4) will generally be present. As such, Sun and Abraham (2021) and Borusyak et al. (2024) study the potential for contamination across estimates of post- and pre-treatment effects (with the latter used in conventional pre-trend specification tests). Furthermore, like in the previous case with static treatment, the own-treatment weights in the first term are potentially negative. While random treatment timing assumptions may solve the issue of negative own treatment weights, contamination bias remains a concern even under such assumptions.

To illustrate the potential for contamination bias, consider again the example with early, late, and never adopters and $T = 3$, except we now allow the treatment effect to be dynamic. Let $\tau_{jts} = Y_{jt}(s) - Y_{jt}(-\infty)$, $s \in \{-2, 1, 0, 1\}$ denote the effect on unit j in time period t of adopting the treatment s periods ago. If s is negative, we interpret this as the anticipation

effect of adopting the treatment $-s$ periods from now. Under our assumptions $\tau_{jt,-1} = 0$, such that there is no anticipation effect immediately before treatment adoption. To test whether the two-period-ahead anticipation effect is zero, and whether the effect of the treatment fades out over time, we let $X_{jt} = (\mathbb{1}\{D_{jt} = -2\}, \mathbb{1}\{D_{jt} = 0\}, \mathbb{1}\{D_{jt} = T-1\})'$. Thus, for instance, $X_{j1} = (1,0,0)'$ for late adopters while $X_{j2} = (0,1,0)'$ for early adopters. Let $\tau_{E,ts} = n_E^{-1} \sum_{j \in \mathcal{E}} E[\tau_{jts}]$ denote the average effect among early adopters, and define $\tau_{L,ts}$ similarly. Then some rather tedious algebra shows that

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

where

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

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

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

Partialling out the unit and time effects therefore yields

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

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

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

Hence, by eq. (B.3), TWFE regression identifies the average treatment for the treated, $\beta = \frac{1}{n_1}\sum_{j=1}^{n_1}(\tau_{j1,-(T-1)}, \ldots, \tau_{j,T_1-2,-2}, \tau_{jT_1,1}, \ldots, \tau_{jT,T-T_1})$. Intuitively, since the contamination weights sum to zero and there is only one group of adopters, the contamination weights must be identically zero.

Mover regressions: multiple treatments with multiple transitions. Finally, consider a "mover regression" in a setting with a static multivalued treatment $\mathcal{D}_{jt} \in \{0, \ldots, K\}$ with multiple transitions of units between treatment states, leading to multiple treatment paths. We focus on the static treatment case, setting $D_{jt} = \mathcal{D}_{jt}$. This setting has been studied by Hull (2018) and De Chaisemartin and D'Haultfœuille (2023). Our Proposition 1 shows that such specifications can suffer from two distinct sources of bias: own-treatment negative weighting from multiple transitions and contamination bias from the multiple treatments. As before the former bias disappears under random treatment timing (as in Athey and Imbens (2022)), or other assumptions which make eq. (12) hold.

To illustrate this case, consider a setting with $T = 3$ periods, $K = 3$ treatments, and three groups of units, $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{N}$. Units in the first group start out untreated, move to treatment 2 in period 1, and move to treatment 3 in period 3. Units in the second group start in treatment 1, move to being untreated in period 2, and move to treatment 2 in period 3. Units in group $\mathcal{N}$ are never treated. This example is isomorphic to the previous event study example, in that it leads to the same regression specification and the same eq. (B.4) characterization of regression coefficients. Thus, there are no negative own-treatment weights in this example, but there are potentially large contamination weights depending on the relative group sizes.

Appendix C Details on the Further Applications

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

C.1 Article Search Protocol

We scraped the American Economic Association (AEA) website for a list of all published articles across all AEA journals over 2013–2022. This search included all articles from the following journals: *American Economic Review*, *American Economic Review: Insights*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Microeconomics*, *Journal of Economic Literature* (excluding articles with "review" in the title and articles labeled as Front Matter, Doctoral Dissertations, and Annotated Listings), *Journal of Economic Perspectives*, and *AER/AEA Papers and Proceedings* (excluding articles with "report" or "minutes" in the title). We limited this search to articles with online replication packages which include at least one data file.[C.1]

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

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

We focused on racial disparities as a set of possible examples because these papers typically have three or more categories, and they were easily identifiable based on keywords, giving us a systematic way to identify them. These searches yielded a total of 1,848 experiments and 67 observational studies on race. To further narrow down experiments, we restricted attention to papers where one of the keywords appears in the paper's title, abstract, or associated tweet.

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

C.2 Overlap Sample and Propensity Score Variation

For each main specification, we identify a subset of the analysis sample with full treatment overlap using the following procedure. First, we define a primary strata variable (when not otherwise obvious from the paper) as the discrete variable with the greatest number of unique levels. In the experimental applications this is always the randomization strata; in the observational applications this is the "finest" fixed effect. We then drop observations for the levels

[C.1] Here "data files" refers to those with any of the following extensions: Stata (`'dta'`), Excel (`'xls'` or `'xlsx'`), Matlab (`'mat'`), R (`'rdata'`, `'rda'`, `'rds'`), HDFS (`'h5'`, `'hdf5'`), Apache (`'parquet'`, `'arrow'`), SAS (`'sas7bdat'`), and delimited files (`'csv'`, `'tsv'`).

	Wald			LM		
	Statistic	(d.f.)	p-value	Statistic	(d.f.)	p-value
Project STAR	309.8	(155)	0.000	336.6	(156)	0.000
Benhassine et al.	207.2	(159)	0.006	217.2	(194)	0.121
Cole et al.	22.7	(39)	0.983	70.3	(54)	0.067
de Mel et al.	0.9	(392)	1.000	1.1	(392)	1.000
Drexler et al.	12.4	(14)	0.574	12.6	(14)	0.555
Duflo et al.	109.6	(254)	1.000	94.5	(258)	1.000
Fryer and Levitt	3947.6	(630)	0.000	4164.0	(681)	0.000
Rim et al.	1403.5	(88)	0.000	233.0	(234)	0.506
Weisburst	2350.0	(69)	0.000	223.2	(48)	0.000

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

Table C.1: Tests of Propensity Score Variation

of this variable which do not exhibit all levels of the treatment. Finally, in the remaining sample, we drop any additional controls which have no within-treatment variation.

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

C.3 Full Results

In Tables C.2–C.10, we report the estimated effects for each application. Panel A of each table first reports the $\hat{\beta}$ estimates from the multiple-treatment regression as reported in the original paper and corresponding standard errors. We also report the own-treatment effect component from the decomposition in eq. (23) along with three alternative estimators: the average treatment effect (ATE) estimator, easiest-to-estimate weighting (EW) and the easiest-to-estimate common weighting (CW) estimator. Panel B reports the difference between $\hat{\beta}$ and these 4 alternative estimators. The $\hat{\beta}$, EW and CW estimators are consistent even without overlap. However, if full overlap fails in the full sample, the own-treatment effect component from the decomposition in eq. (23) may not be identified for all treatments, and the ATE is not identified. If identification of the decomposition fails for the full treatment vector, we

subset to the overlap sample, as described in Appendix C.2 above, and report the full set of estimates from the different estimators.

	Full sample					Overlap				
A. Estimates	$\hat{\beta}$	Own	ATE	EW	CW	$\hat{\beta}$	Own	ATE	EW	CW
Small class size	5.267			5.248	5.530	5.311	5.156	5.515	5.248	5.529
	(0.773)			(0.771)	(0.760)	(0.774)	(0.773)	(0.758)	(0.771)	(0.760)
				[0.739]	[0.738]			[0.740]	[0.739]	[0.738]
Teaching aide	0.242			0.292	0.040	0.205	0.388	0.099	0.292	0.040
	(0.716)			(0.711)	(0.708)	(0.716)	(0.710)	(0.705)	(0.711)	(0.708)
				[0.688]	[0.691]			[0.691]	[0.688]	[0.691]
Number of controls	78					77				
Sample size	5,902					5,868				
B. Bias										
Small class size				0.020	−0.262		0.155	−0.204	0.063	−0.219
				(0.139)	(0.196)		(0.160)	(0.219)	(0.134)	(0.192)
Teaching aide				−0.050	0.202		−0.184	0.106	−0.087	0.165
				(0.128)	(0.195)		(0.149)	(0.187)	(0.124)	(0.192)

Notes: This table reports estimates from the STAR application, as described in Appendix C.3. Robust standard errors are reported in parentheses. Standard errors assuming known propensity scores are reported in square brackets.

Table C.2: Full results: STAR

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

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

Table C.3: Full results: Benhassine et al. (2015)

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

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

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

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

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

Table C.5: Full results: de Mel et al. (2013)

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

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

Table C.6: Full results: Drexler et al. (2014)

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

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

Table C.7: Full results: Duflo et al. (2015)

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

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

Table C.8: Full results: Fryer and Levitt (2013)

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

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

Table C.9: Full results: Rim et al. (2020)

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

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

Table C.10: Full results: Weisburst (2019)

Appendix D Additional Figures

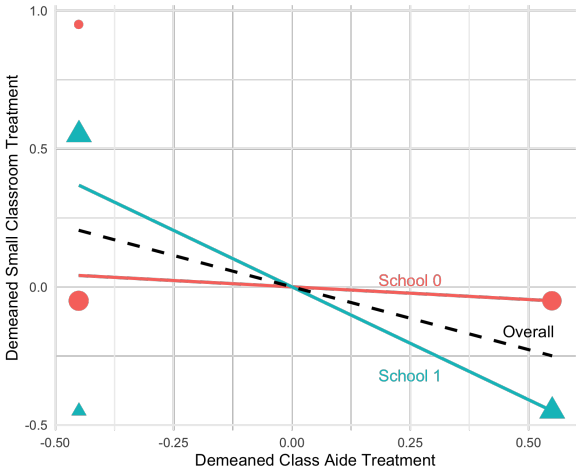

Figure D.1: Regression of Small Classroom Treatment on Class Aide Treatment

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

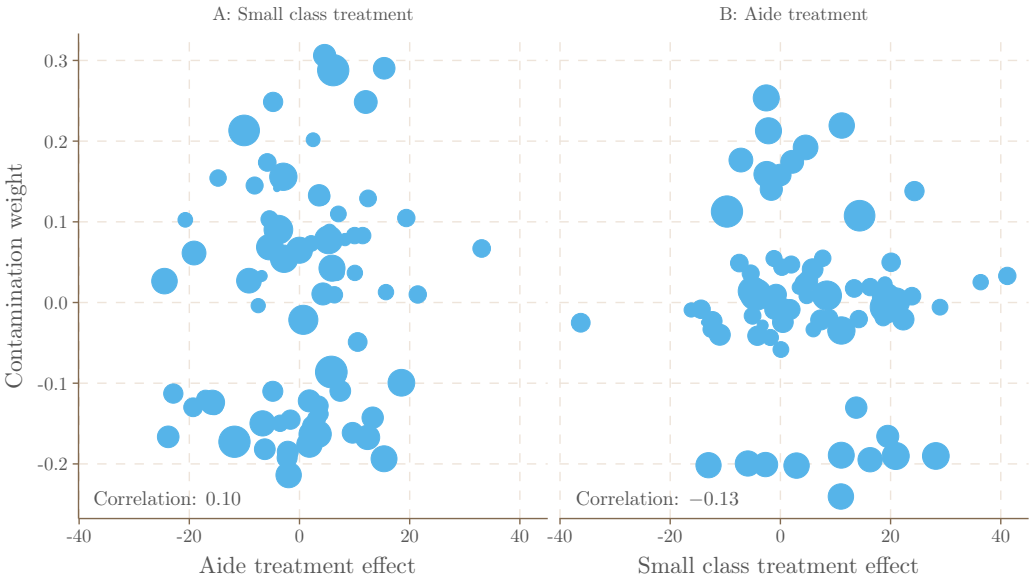

Figure D.2: Project STAR contamination weights.

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

Figure D.3: Project STAR treatment weights

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

References

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

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2015). Turning a shove into a nudge? a "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, *7*(3), 86–125. https://doi.org/10.1257/pol.20130225

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

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*. https://doi.org/10.1093/restud/rdae007

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cole, S., Giné, X., Tobacman, J., Topalova, P., Townsend, R., & Vickery, J. (2013). Barriers to household risk management: Evidence from India. *American Economic Journal: Applied Economics*, *5*(1), 104–135. https://doi.org/10.1257/app.5.1.104

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2006). *Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand* (Working Paper No. 330). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/t0330

De Chaisemartin, C., & D'Haultfœuille, X. (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. *Journal of Econometrics*, *236*(2), 105480. https://doi.org/10.1016/j.jeconom.2023.105480

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

de Mel, S., McKenzie, D., & Woodruff, C. (2013). The demand for, and consequences of, formalization among informal firms in Sri Lanka. *American Economic Journal: Applied Economics*, *5*(2), 122–150. https://doi.org/10.1257/app.5.2.122

Drexler, A., Fischer, G., & Schoar, A. (2014). Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics*, *6*(2), 1–31. https://doi.org/10.1257/app.6.2.1

Duflo, E., Dupas, P., & Kremer, M. (2015). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review*, *105*(9), 2757–2797. https://doi.org/10.1257/aer.20121607

Fryer, R. G., & Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, *103*(2), 981–1005. https://doi.org/10.1257/aer.103.2.981

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

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

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

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

McNeney, B., & Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *Journal of Statistical Planning and Inference*, *91*(2), 441–480. https://doi.org/10.1016/S0378-3758(00)00193-2

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*(6), 1349–1382. https://doi.org/10.2307/2951752

Rim, N., Ba, B., & Rivera, R. (2020). Disparities in police award nominations: Evidence from Chicago. *AEA Papers and Proceedings*, *110*, 447–451. https://doi.org/10.1257/pandp.20201118

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

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

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

van der Vaart, A., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer. https://doi.org/10.1007/978-1-4757-2545-2

Weisburst, E. K. (2019). Police use of force as an extension of arrests: Examining disparities across civilian and officer race. *AEA Papers and Proceedings*, *109*, 152–156. https://doi.org/10.1257/pandp.20191028

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