Doubly Robust Estimators with Weak Overlap

Yukun Ma* Pedro H. C. Sant'Anna† Yuya Sasaki‡ Takuya Ura§

April 8, 2026

Abstract

Doubly robust (DR) estimators guard against model misspecification but remain sensitive to weak covariate overlap. We show that trimming extreme propensity scores reduces variance but eliminates double robustness. We introduce DR estimators that retain double robustness after trimming through bias correction, preserving the original causal targets across unconfoundedness, instrumental variables, and difference-in-differences designs. In four applications, the proposed estimator yields substantially more precise estimates: ruling out large mortality effects of Medicaid expansion, detecting workforce growth from mental health reform, recovering the Black–White test score gap without strong functional form restrictions, and recovering a positive 401(k) savings effect consistent with the prior literature.

JEL Codes: C10, C14, C21, C23, C26.

Keywords: Difference-in-Differences; LATE; Unconfoundedness; Instrumental Variables; Propensity Score; Trimming; Bias Correction.

*University of Rochester. Email: yma69@ur.rochester.edu

†Emory University. Email: pedro.santanna@emory.edu

‡Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu

§University of California, Davis. Email: takura@ucdavis.edu

1 Introduction

Accounting for confounders is essential for reliable observational causal inference, but doing so poses practical challenges due to model misspecification and weak covariate overlap. Doubly robust (DR) estimators address the first concern: they remain consistent for a prespecified causal estimand even when one component of the working model is incorrect (e.g., Robins, Rotnitzky and Zhao, 1994).[1] Yet DR estimators remain vulnerable to the second problem: when propensity scores are close to zero or one, they can become unreliable. A handful of observations with extreme inverse probability weights can dominate the entire estimate, inflating standard errors by factors of two to six, or worse, flipping the sign of point estimates. Weak overlap does not merely reduce precision; it can render observational studies uninformative or misleading.

As we highlight in different applications, this concern is indeed warranted. For instance, we find that the standard DR 95% simultaneous confidence bands in a staggered difference-in-differences (DiD) analysis of Medicaid expansion on mortality span on average 53 deaths per 100,000 people over the post-treatment horizons—too wide to inform policy. In a 401(k) savings application, the standard DR local average treatment effect (LATE) estimate implies that participation *reduces* net financial assets by $13,042—a conclusion at odds with the prior literature—because two ineligible observations with extreme instrument propensity scores carry effective weights of roughly 31 and 98 each, dominating the remaining observations. Removing the direct influence of these observations recovers a sensible point estimate; the bias correction we develop below allows this trimmed estimator to target the *original* LATE with valid inference, rather than a redefined trimmed estimand. A natural first response is to trim observations with extreme propensity scores. Although intuitive, we highlight that trimming a DR estimator eliminates its double robustness, the very property that justified using DR methods in the first place.

We demonstrate this via simulations: when the outcome regression is misspecified but the propensity score is correct, the standard trimmed DR estimator should remain consistent, yet its 95% coverage collapses to 34% at a trimming threshold of $h = 0.10$ (Section 2.2). The double robustness guarantee is gone. The alternatives to trimming also come with important caveats. Changing the target parameter of interest (e.g., Crump, Hotz, Imbens and Mitnik, 2006; Li, Morgan and Zaslavsky, 2018; Słoczyński, 2022, 2026; Blandhol, Bonney, Mogstad and Torgovitsky, 2026) addresses the variance problem by redefining the causal question being answered—paraphrasing the song *When I'm Not Near the Girl I Love*, "When I'm not near the parameter I love, I love the parameter I'm near."[2] Imposing a linear specification and interpreting the regression treatment coefficient as a weighted average of heterogeneous treatment effects is another common response, but the resulting estimand can have unattractive properties such as placing negative weight on

1 See Słoczyński and Wooldridge (2018) for general identification results with doubly robustness in the causal inference.

2 Our understanding is that Art Goldberger is the source of this paraphrase. We thank Matt Masten for bringing this to our attention.

some subpopulations and being hard to motivate via policy justification, and must be derived on a case-by-case (and design-by-design) basis (Słoczyński, 2022, 2026; Mogstad and Torgovitsky, 2024; Caetano and Callaway, 2024).

This paper shows that robustness to weak covariate overlap need not come at the cost of re-defining the causal estimand. We make three contributions. First, we establish a negative result: trimming without bias correction eliminates the double robustness property of DR estimators. The trimmed estimator is consistent only if the outcome regression is correctly specified, losing robustness to the very misspecification it was designed to guard against (Section 2.2). Second, we propose a class of bias-corrected DR estimators that is also robust against weak overlap. We show that a key condition for bias correction—that the conditional mean of the quantity being averaged vanishes as the propensity score approaches its boundary—holds under *either* route to double robustness. When the propensity score is correctly specified, no observations of the relevant type exist at the boundary; when the outcome regression is correctly specified, the conditional expectation of the DR residual is zero everywhere. Either way, the conditional mean of the ratio has a well-defined limit at the boundary, and a low-dimensional polynomial can reconstruct the contribution of the trimmed observations, so the bias correction preserves double robustness. Third, we highlight that our unified framework covering the ATE, LATE, and staggered DiD designs, and derive a closed-form influence function that enables standard inference. Implementation adds a single polynomial regression step to any existing DR procedure.

Rich conditioning sets are often needed for identification—interviewer fixed effects for the test score gap, county characteristics for parallel trends, income controls for instrument validity—but the same covariates that make identification credible can make estimation infeasible through weak overlap (D'Amour, Ding, Feller, Lei and Sekhon, 2021). Our estimator addresses this tension: researchers can condition on whatever covariates identification requires while substantially reducing the variance penalty. Four empirical applications—two DiD, one under unconfoundedness, and one using instrumental variables—demonstrate that the proposed estimator produces tighter confidence intervals and recovers informative estimates: in a DiD analysis of Medicaid expansion, the 95% simultaneous confidence bands narrow from roughly 53 to 17 deaths per 100,000 on average over the post-treatment horizons; in an IV analysis of 401(k) participation, the estimator recovers a positive and significant savings effect of $8,864, consistent with prior findings; and under unconfoundedness, it detects the Black–White test score gap at age two more precisely than linear regressions while directly targeting the ATE without imposing linearity. Across the four applications, standard errors from existing DR methods are 1.8 to 6.6 times larger than those from our proposed doubly robust bias-corrected (DR-BC) method. Since reducing the standard error by a factor of c translates into a c^2-fold reduction in the sample size needed to detect a given effect at conventional power, these gains expand what applied researchers can learn from existing datasets.

Our approach builds on and extends several strands of work on causal inference under weak

overlap (e.g., Crump, Hotz, Imbens and Mitnik, 2009; Khan and Tamer, 2010).[3] We are mostly related to Yang and Ding (2018) and Heiler and Kazak (2021), who study DR-type estimators in unconfoundedness settings; however, Yang and Ding (2018) relies on trimming strategies that redefine the estimand, while Heiler and Kazak (2021) does not consider trimming at all. None of these papers considers more general research designs as we do. Sasaki and Ura (2022) propose bias-corrected inverse probability weighting estimators for a single ratio moment under weak overlap, relying on the boundary condition when the propensity score is correctly specified. Our framework differs in that it accommodates doubly robust estimands, establishes boundary conditions under both DR routes, and handles the nonlinear multi-moment structure required for the ATE, LATE, and DiD designs. In particular, the Hájek-type normalization that is standard in causal inference creates a denominator dependence across moments—requiring us to jointly bias-correct the normalization moments—and the residual-versus-normalization moment asymmetry we exploit under the outcome-correct route has no analog in the unnormalized, single-moment setting of Sasaki and Ura (2022). See, e.g., Słoczyński, Uysal and Wooldridge (2025) for a recent discussion of why Hájek-type estimators are predominant in causal inference problems.

Notation. We write $\mathbb{E}[\cdot]$ for the expectation operator and $\mathbb{E}_n[\cdot] = n^{-1}\sum_{i=1}^{n}(\cdot)_i$ for the sample average. The indicator function is $\mathbb{1}\{\cdot\}$. For a nuisance parameter γ, we let γ_0 denote the true value, $\hat{\gamma}$ an estimator, and γ^* the probability limit of $\hat{\gamma}$.

2 Overview and practical relevance

This section introduces the setup, demonstrates that trimming compromises double robustness, presents four empirical applications, and describes the proposed estimator.

2.1 Setup and DR estimands

Our framework covers different research designs commonly used in empirical work, including unconfoundedness (selection on observables), LATE (instrumental variables), and DiD. Although the notation and identification assumptions differ across designs, they share a common structure: each involves a propensity score that can take values near its boundary, leading to weak overlap, and each admits a DR estimand that combines an outcome regression with inverse probability weighting. We present the DR estimands for these three designs using normalized weights that sum to one within each treatment group.

3 See Sasaki and Ura (2022) for more discussions on this literature.

2.1.1 Unconfoundedness

A common goal in observational studies is to estimate the average effect of a binary treatment on an outcome, while adjusting for pre-treatment covariates to make the identification assumptions more plausible. Let $Y(1)$ and $Y(0)$ respectively denote the potential outcomes with and without treatment, $D \in \{0,1\}$ the treatment indicator, and X a vector of pre-treatment confounding variables. The propensity score is $p_D(X) = \mathbb{E}[D \mid X]$, and the outcome regression is $m_Y(d,X) = \mathbb{E}[Y \mid D = d, X]$. Under unconfoundedness ($\{Y(0), Y(1)\} \perp\!\!\!\perp D \mid X$) and overlap ($0 < p_D(X) < 1$), the ATE $= \mathbb{E}[Y(1) - Y(0)]$ is identified and admits the following DR representation (see, e.g., Słoczyński and Wooldridge, 2018, and references therein)

$$\text{ATE}_{\text{DR}} = \mathbb{E}\left[m_Y(1,X) - m_Y(0,X) + w^{\text{ATE}}_{D=1}(D,X)\,(Y - m_Y(1,X)) - w^{\text{ATE}}_{D=0}(D,X)\,(Y - m_Y(0,X))\right], \quad (2.1)$$

where

$$w^{\text{ATE}}_{D=1}(D,X) = \frac{D/p_D(X)}{\mathbb{E}\left[D/p_D(X)\right]}, \quad w^{\text{ATE}}_{D=0}(D,X) = \frac{(1-D)/(1-p_D(X))}{\mathbb{E}\left[(1-D)/(1-p_D(X))\right]}. \quad (2.2)$$

When $p_D(X)$ is close to zero or one for some values of X, the weights in (2.2) can become large. This is the weak overlap problem. Standard two-step estimators for (2.1) does not achieve the $\sqrt{n}$-consistency in such weak-overlap cases (Khan and Tamer, 2010).

2.1.2 Instrumental variables and LATE

When treatment is endogenous, and unconfoundedness is not empirically plausible, researchers often use an instrument to identify the average treatment effect among units whose treatment status is shifted by the instrument, i.e., the compliers (Imbens and Angrist, 1994). Consider a setting with a binary instrument Z, binary treatment D, and observed covariates X, with potential treatments $D(1)$ and $D(0)$ and potential outcomes $Y(1), Y(0)$. Define the instrument propensity score as $p_Z(X) = \mathbb{E}[Z \mid X]$, the reduced-form and first-stage regressions as $m^{\text{LATE}}_Y(z,X) = \mathbb{E}[Y \mid Z = z, X]$ and $m^{\text{LATE}}_D(z,X) = \mathbb{E}[D \mid Z = z, X]$, respectively. Under the standard IV-LATE assumptions, the local average treatment effect LATE $= \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0)]$ is identified and the DR estimand for the LATE (Belloni, Chernozhukov, Fernández-Val and Hansen, 2017; Słoczyński, Uysal and Wooldridge, 2022; Słoczyński et al., 2025) can be written as

$$\text{LATE}_{\text{DR}} = \frac{\mathbb{E}\left[m^{\text{LATE}}_Y(1,X) - m^{\text{LATE}}_Y(0,X) + w^{\text{LATE}}_{Z=1}(Z,X)\left(Y - m^{\text{LATE}}_Y(1,X)\right) - w^{\text{LATE}}_{Z=0}(Z,X)\left(Y - m^{\text{LATE}}_Y(0,X)\right)\right]}{\mathbb{E}\left[m^{\text{LATE}}_D(1,X) - m^{\text{LATE}}_D(0,X) + w^{\text{LATE}}_{Z=1}(Z,X)\left(D - m^{\text{LATE}}_D(1,X)\right) - w^{\text{LATE}}_{Z=0}(Z,X)\left(D - m^{\text{LATE}}_D(0,X)\right)\right]} \quad (2.3)$$

where

$$w^{\text{LATE}}_{Z=1}(Z,X) = \frac{Z/p_Z(X)}{\mathbb{E}\left[Z/p_Z(X)\right]}, \quad w^{\text{LATE}}_{Z=0}(Z,X) = \frac{(1-Z)/(1-p_Z(X))}{\mathbb{E}\left[(1-Z)/(1-p_Z(X))\right]}.$$

The numerator identifies the reduced-form effect and the denominator the compliance share, so the ratio recovers the LATE. Weak overlap arises when $p_Z(X)$ is close to zero or one, potentially implying a high variance of two-step estimators based on (2.3).

2.1.3 Difference-in-Differences

Another widely used research design that accommodates selection on unobservables is the DiD design. Consider a staggered DiD panel data setup with T time periods (Callaway and Sant'Anna, 2021). Let G denote the period of first treatment ($G = \infty$ if never treated), X be a vector of pre-treatment covariates, and let $Y_t(g)$ be the potential outcome in period t if a unit was first treated in period g. Let $\delta \geqslant 0$ denote the number of anticipation periods. Under limited anticipation and a conditional parallel trends assumption (Callaway and Sant'Anna, 2021), the group-time average treatment effect $\mathrm{ATT}(g,t) = \mathbb{E}[Y_t(g) - Y_t(\infty) \mid G = g]$ is identified for $t \geqslant g - \delta$. The more aggregated event-study estimand

$$\mathrm{ES}(e) = \sum_g w^{\mathrm{es}}_{g,e} \cdot \mathrm{ATT}(g, g+e), \tag{2.4}$$

where $w^{\mathrm{es}}_{g,e} = \mathbb{P}(G = g \mid G + e \leqslant T, G \leqslant T)$ is also identified. Estimation of each $\mathrm{ATT}(g,t)$ relies on a comparison group $\mathcal{C}_{g,t}$ (either never-treated or not-yet-treated units) with indicator $C_{g,t}$, generalized propensity score $p_{g,t}(X) = \mathbb{P}(G = g \mid X, \mathbb{1}\{G = g\} + C_{g,t} = 1)$, and outcome regression $m_{g,t}(X) = \mathbb{E}[Y_t - Y_{g-\delta-1} \mid X, C_{g,t} = 1]$. The DR estimand for the $\mathrm{ATT}(g,t)$ parameters is given by

$$\mathrm{ATT}_{\mathrm{DR}}(g,t) = \mathbb{E}\left[\left(w^{\mathrm{DiD}}_{G=g}(G) - w^{\mathrm{DiD}}_{g,t}(G,X)\right)\left(Y_t - Y_{g-\delta-1} - m_{g,t}(X)\right)\right], \tag{2.5}$$

with weights given by

$$w^{\mathrm{DiD}}_{G=g}(G) = \frac{\mathbb{1}\{G=g\}}{\mathbb{E}\left[\mathbb{1}\{G=g\}\right]}, \quad w^{\mathrm{DiD}}_{g,t}(G,X) = \frac{C_{g,t}\, p_{g,t}(X)/(1-p_{g,t}(X))}{\mathbb{E}\left[C_{g,t}\, p_{g,t}(X)/(1-p_{g,t}(X))\right]}.$$

DR estimands for $\mathrm{ES}(e)$ are defined by replacing $\mathrm{ATT}(g, g+e)$ in (2.4) with $\mathrm{ATT}_{\mathrm{DR}}(g, g+e)$ as defined in (2.5). Weak overlap arises when $p_{g,t}(X)$ is close to one, potentially implying a high variance of DiD and ES estimators based on (2.5).

2.1.4 The common challenge

Across all three designs, the DR estimand involves weights that divide by either the propensity score or one minus the propensity score (for the ATE, both treated and control weights can be extreme; for the LATE, both instrument groups; for DiD, only the comparison-group weights). When any such denominator is near zero, the resulting weights become very large. To illustrate, consider a DiD setup: if a comparison unit has an estimated propensity score of $\widehat{p}_{g,t}(X) = 0.99$, its weight is proportional to $0.99/(1-0.99) = 99$, as influential as 99 observations with $\widehat{p}_{g,t}(X) = 0.50$. A handful of such observations can dominate the entire estimate, leading to a high variance of the resulting estimator. Trimming these observations reduces variance but introduces bias for the original estimand. As we show next, trimming a standard DR estimator can *eliminate its double robustness property.*

2.2 Trimming compromises double robustness

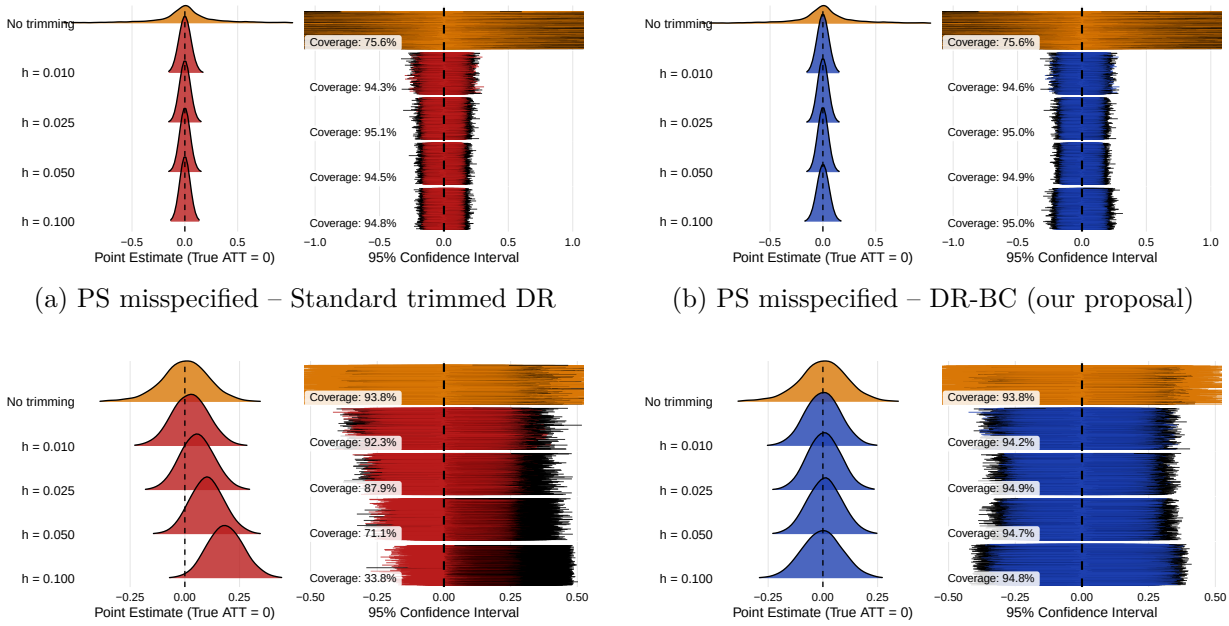
A key appeal of the DR estimands presented above is their robustness to model misspecification: estimators based on them remain consistent for the causal parameter of interest even if one of the two working models (the propensity score or the outcome regression) is misspecified. A natural question is: Does this property survive when one trims observations with extreme propensity scores? The answer is no, as we illustrate via a simulation exercise.

Consider a DiD setup with two periods and two groups, with $n = 10{,}000$ units and 4 covariates drawn from independent Student-t distributions with 10 degrees of freedom. Treatment group membership is determined by a logistic model, and potential outcomes are linear in transformations of the covariates, along with covariate-specific time trends and unit-level unobserved heterogeneity. The true ATT in our setup is zero. Our data-generating process is similar to Kang and Schafer (2007) and Sant'Anna and Zhao (2020), and we create weak overlap by design so that a fraction of comparison-group units have a propensity score close to one. To assess double robustness, we consider two complementary configurations: in the first, the propensity score model is misspecified while the outcome regression is correctly specified; in the second, the propensity score is correctly specified while the outcome regression is misspecified. Double robustness should protect us in both cases: the correctly specified component should ensure consistent estimation despite misspecification of the other. We repeat each exercise 10,000 times; full details of the data-generating process are in Appendix A. Results for additional configurations, including cases where both models are correctly specified, are qualitatively similar and reported in Appendix A. We present the DiD design here; analogous simulations for the ATE and LATE designs yield the same conclusions and are omitted for brevity.

We use the standard DR DiD estimator based on (2.5), estimate the propensity score via logit regression, and estimate the outcome regression via ordinary least squares (OLS). To assess the impact of trimming comparison-group observations with propensity scores above $1 - h$, we vary the trimming threshold over $h \in \{0, 0.01, 0.025, 0.05, 0.10\}$, where $h = 0$ corresponds to the no-trimming case. Figure 1 summarizes the results. Each panel shows ridge density plots alongside zipper confidence interval plots, where black segments indicate non-coverage. The ideal scenario would be density plots centered at zero with moderate spread, and tight confidence intervals with correct coverage.

The left column of Figure 1 presents the results for the standard trimmed DR DiD estimator. When the propensity score is misspecified but the outcome regression is correct (Panel (a)), the trimmed estimator performs well: bias is negligible and coverage remains near 95% across all positive thresholds, as one would expect from a doubly robust estimator. However, the reverse case (Panel (c)) reveals the problem: when the outcome regression is misspecified and the propensity score is correct, trimming introduces growing bias. At $h = 0.05$, the 95% empirical coverage drops to 71%, and at $h = 0.10$, it collapses to 34%. The double robustness property is lost, as the

Figure 1: Effect of trimming on DR DiD estimators under model misspecification



(a) PS misspecified – Standard trimmed DR

(b) PS misspecified – DR-BC (our proposal)

(c) Outcome misspecified – Standard trimmed DR

(d) Outcome misspecified – DR-BC (our proposal)

Notes. Simulation design as discussed in Appendix A, with $n = 10{,}000$ and 10,000 Monte Carlo repetitions. The true ATT is zero. Top row (Panels (a)–(b)): propensity score misspecified, outcome regression correct. Bottom row (Panels (c)–(d)): propensity score correct, outcome regression misspecified. In both cases, double robustness should ensure consistent estimation. Left column (Panels (a), (c)): standard trimmed DR estimator. Right column (Panels (b), (d)): our bias-corrected estimator (DR-BC). Ridge plots show the density of point estimates; zipper plots show 95% confidence intervals across 10,000 Monte Carlo draws, with black segments indicating non-coverage.

reliability of this procedure depends on the outcome model being correctly specified. We expand on this rationale in Section 2.4.

The right column of Figure 1 (Panels (b) and (d)) shows that our proposed weak-overlap-robust DR estimator resolves this trade-off. It trims comparison-group observations with extreme propensity scores, just as the standard trimmed estimator does, but adds a bias-correction term that compensates for the trimmed observations. In both configurations, whether the propensity score or the outcome regression is the correctly specified component, the estimator remains well-centered at the truth, and coverage stays near the nominal 95% level (94–95%). Confidence intervals at $h = 0.05$ are narrower than at $h = 0$, so the bias-corrected estimator retains double robustness while achieving the variance reduction that trimming provides. Thus, although the standard trimming procedure compromises the DR property, our proposed bias-correction procedure restores it and yields estimators with attractive statistical guarantees.

2.3 Empirical applications

Our simulation results clearly illustrated that trimming DR estimators can lead to unreliable inferences in the presence of model misspecifications. Thus, in practice, one may argue that the two natural candidate estimators to estimate a particular treatment effect parameter are the

standard (untrimmed) DR estimator and our proposed DR bias-corrected (DR-BC) estimator. One may wonder if our DR-BC estimator indeed improves upon the standard DR estimators in empirical applications. In this section, we address this question. We compare DR-BC ($h = 0.05$) with the standard untrimmed DR estimator ($h = 0$) in four applications spanning all three research designs: unconfoundedness, IV and DiD. Because both estimators target the same causal parameter, differences in standard errors reflect genuine precision gains, not changes in what is being estimated or the causal question being addressed.[4] Because we focus on applications where weak overlap is a concern, the precision gains below are large; in settings with adequate overlap, the gains would be more modest.

2.3.1 Difference-in-Differences

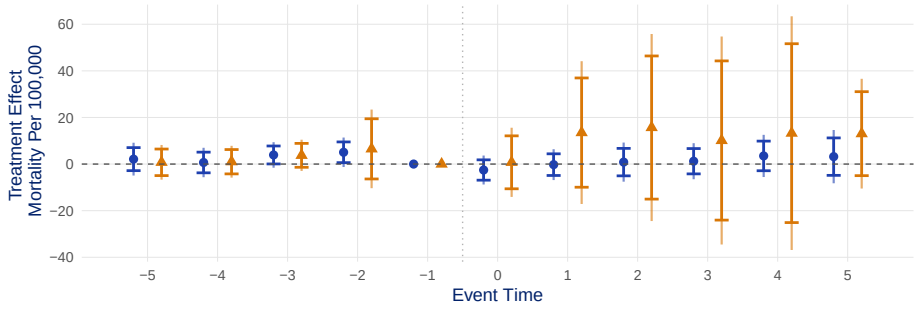
We apply our method to two staggered DiD settings using the DR framework of Callaway and Sant'Anna (2021). In both cases, a rich set of covariates is used to make conditional parallel trends more plausible, but conditioning on them leads to weak overlap.

Medicaid expansion and mortality. Does Medicaid expansion save lives? Baker, Callaway, Cunningham, Goodman-Bacon and Sant'Anna (2025) provides a DiD practical guide and uses this question as their running example, using county-level all-cause mortality rates for adults ages 20–64 and the staggered adoption of Medicaid expansion across states from 2009 to 2019. We re-analyze their data using the DR DiD framework of Callaway and Sant'Anna (2021) with never-treated counties as the comparison group, conditioning on six county-level covariates (poverty rate, unemployment rate, median household income, percent female, percent White, and percent Hispanic) and using population weights. Because these covariates differ between expanding and non-expanding states, the generalized propensity score is close to 1 for some comparison counties, resulting in weak overlap; at $h = 0.05$, up to 3.3% of comparison-group observations are trimmed in the most affected group-time cell (and 0.8% across post-treatment cells, or 177 counties in total). Panel (a) of Figure 2 shows the event-study estimates. Post-treatment standard errors are on average 3.5 times larger for the standard DR estimator, and as large as 4.6 times larger for the 2014 expansion cohort at the four-year horizon. While neither estimator rejects a zero effect, the standard DR 95% simultaneous confidence bands span on average 53 deaths per 100,000 (and up to 74 at longer horizons), which is too wide to be policy relevant. DR-BC narrows these bands to roughly 17 deaths per 100,000 on average, ruling out large effects in either direction and providing a much more informative bound on the plausible magnitude of Medicaid expansion's mortality impact. Overall, this application demonstrates that our proposed DR procedure can

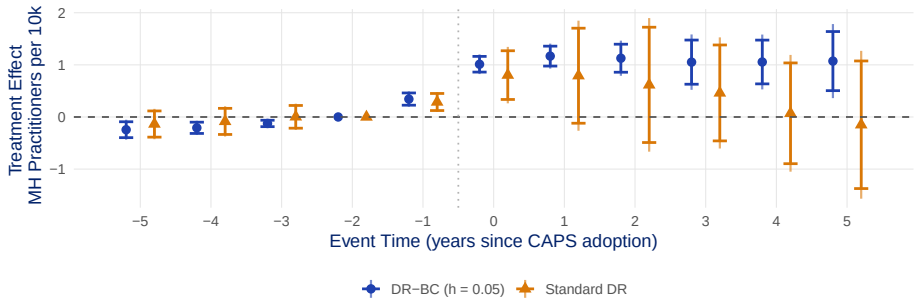
4 In all applications, propensity scores are estimated via logit regression and outcome regressions via least squares. Standard errors are computed analytically from the influence function. For the DiD event studies, standard errors are clustered at the county level (Medicaid) and municipality level (Dias–Fontes), and simultaneous confidence bands are based on multiplier bootstrap.

yield substantial gains in precision in realistic scenarios.

Figure 2: DR DiD event-study estimators across applications



(a) Medicaid expansion on mortality



(b) CAPS adoption on mental health practitioners

Notes. Panel (a) shows population-weighted event-study estimates with staggered treatment timing using the doubly robust estimation method of Callaway and Sant'Anna (2021). The comparison group consists of counties that remained untreated by 2019. The outcome is the crude mortality rate for adults ages 20–64. Data are from Baker et al. (2025). Panel (b) shows event-study estimates with not-yet-treated municipalities as the comparison group. The outcome is mental health practitioners per 10,000 population. Covariates include state fixed effects and 29 baseline characteristics. Three small northern states with fewer than 10 treated municipalities are excluded. We allow for a one-year anticipation. Data are from Dias and Fontes (2024). In both panels, DR-BC is our proposed estimator with trimming threshold $h = 0.05$; Standard DR is the untrimmed estimator of Callaway and Sant'Anna (2021). Points denote point estimates, vertical bars denote 95% simultaneous confidence bands, and error bars with caps denote 95% pointwise confidence intervals.

Mental health reform in Brazil. Did Brazil's 2002 psychiatric reform expand community mental health capacity? Dias and Fontes (2024) study this question using the staggered rollout of community-based Psychosocial Care Centers (CAPS) across municipalities, which replaced centralized hospital-based care. The conditional parallel trends argument in this setting naturally motivates an estimator that accommodates covariate-specific trends, since the timing of CAPS adoption was correlated with municipal characteristics: larger and wealthier municipalities tended to adopt earlier. We re-analyze their data using the DR DiD framework of Callaway and Sant'Anna (2021), focusing on mental health practitioners per 10,000 population; results for additional outcomes (service utilization, hospital spending, violence) are in Appendix B. We restrict the sample to states with at least 10 treated municipalities and adjust for state fixed effects and 29 baseline covariates that are motivated by the original study. Because this rich covariate set leaves limited overlap between treated and not-yet-treated municipalities, a small number of comparison observations receive estimated propensity scores near one and therefore extremely large weights;

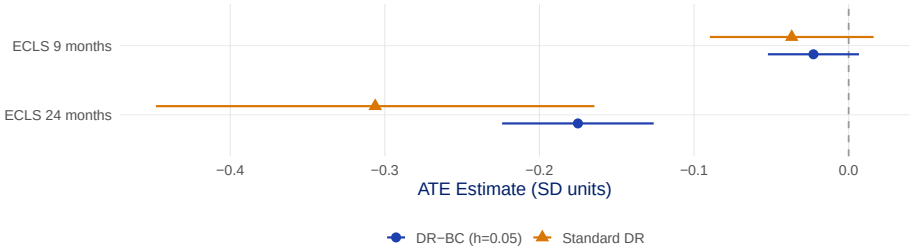
at $h = 0.05$, up to 0.2% of comparison-group observations are trimmed in the most affected group-time cell (158 municipality-year observations in total, or 0.03% across post-treatment cells). Panel (b) of Figure 2 presents the event-study estimates. Post-treatment standard errors are on average 2.8 times larger for the standard DR estimator, and at longer horizons the standard DR confidence intervals become too wide to be informative. DR-BC yields estimates precise enough to trace the workforce expansion over time, with a sustained increase of roughly 1 mental health practitioner per 10,000 population following CAPS adoption, consistent with the findings of Dias and Fontes (2024). As Dias and Fontes (2024) analyzes the impact of CAPS on multiple outcomes, Appendix B compares DR and DR-BC event-study estimates for these outcomes as well. See also Appendix D for a sensitivity analysis with respect to the choice of h.

2.3.2 Unconfoundedness

How large is the Black–White test score gap at age two, and can we estimate it without imposing functional form restrictions? Using data from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Fryer and Levitt (2013) document that the gap is negligible at 9 months but grows to roughly 0.2–0.4 standard deviations by age two, conditioning on a rich set of family background variables via OLS. Because their OLS specification imposes linearity and does not interact race dummies with covariates, it may not recover the ATE under treatment effect heterogeneity (Słoczyński, 2022). We re-analyze their data using DR estimation, treating race as the "treatment" in an operational sense: following Holland (1986), race is not a cause in the counterfactual sense, and the resulting "ATE" is a nonlinear covariate-adjusted descriptive contrast between Black and White children, not a causal effect of race. Their conditioning set includes interviewer fixed effects, fifth-degree polynomials in parental age, and numerous interactions. The interviewer fixed effects create many small cells — interviewers who observed children of mostly one race generate near-zero or near-one propensity scores — pushing the propensity score to its boundaries for a non-negligible share of the sample. At $h = 0.05$, a small number of observations with extreme arm-specific weights drive the instability in the untrimmed DR estimator (see Appendix C for the exact decomposition). Because trimming is defined at the covariate-cell level, roughly 38% and 35% of observations at 9 and 24 months fall within a trimmed covariate cell, but the vast majority of these carry benign effective weights close to one and enter the standard DR estimator without instability (see Appendix D).

Figure 3 presents the results. At 24 months, the standard error drops from 0.072 to 0.025 (a 2.9-fold reduction), while at 9 months the reduction is 1.8-fold (from 0.027 to 0.015). At 24 months, both estimators detect a significant gap, but the standard DR confidence interval spans 0.28 standard deviations compared to 0.10 for DR-BC. For comparison, the original OLS estimate of Fryer and Levitt (2013) at 24 months is -0.213 (SE = 0.032). DR-BC yields -0.175 (SE = 0.025) under a much more flexible specification that does not impose linearity, while the

Figure 3: DR estimates of the Black–White test score gap



Notes. Forest plot of ATE estimates for the Black–White test score gap at ages 9 months and 24 months from the ECLS-B. DR-BC is our proposed estimator with $h = 0.05$; Standard DR is the untrimmed estimator. Covariates and sample restrictions follow Fryer and Levitt (2013). Error bars denote 95% confidence intervals based on analytical standard errors.

standard DR estimate of -0.306 (SE $= 0.072$) illustrates the precision cost of flexibility without overlap correction. DR-BC thus delivers OLS-level precision while directly targeting the ATE, without strong functional-form restrictions.

The large share of trimmed covariate cells in this application is driven by interviewers who observe almost only one racial group, mechanically pushing the propensity score to its boundary. These interviewers are scattered across the covariate distribution, so the trimmed cells do not represent a distinct demographic subpopulation. This explains why the polynomial reconstruction we use for bias correction is adequate despite the large share of the sample lying in trimmed cells, and why the h-sensitivity analysis in Appendix D (Figure OA-6) confirms that point estimates remain stable across h. Appendix C also reports the estimated propensity score distribution (Figure OA-5).

2.3.3 Instrumental variables, LATE, and ITT

Does 401(k) participation increase household savings? A central challenge is that participation is endogenous. A widely used identification strategy exploits employer-determined *eligibility* as an instrument for participation (Abadie, 2003; Chernozhukov and Hansen, 2004; Benjamin, 2003). Using the 1991 Survey of Income and Program Participation (SIPP) data and the same covariates as Abadie (2003), we estimate both the ITT effect of eligibility and the LATE of participation for two outcomes: net financial assets and total wealth. We consider two samples: a full sample requiring only positive income (Chernozhukov and Hansen, 2004; Benjamin, 2003), and a restricted sample with income between \$10,000 and \$200,000 (Abadie, 2003). The income restriction was motivated precisely by weak overlap: as Abadie (2003, p. 249) notes, "outside this interval, 401(k) eligibility in the sample is rare." Indeed, there are a few variance-inflating observations in this dataset: at $h = 0.05$, only 7 of 9,910 observations (0.07%) fall in a trimmed cell in the full sample and 4 of 9,275 (0.04%) in the restricted sample—all with $\hat{p}_Z(X) > 0.95$. Within each trimmed cell, only the observations on the active (ineligible) arm carry large weights: in the full sample, the 2 ineligible observations in the $\hat{p}_Z(X) > 0.95$ tail have effective weights $1/(1 - \hat{p}_Z(X))$ of roughly 31 and 98,

while the 5 eligible observations in the same tail have effective weights $1/\hat{p}_Z(X) \approx 1.03$ and do not destabilize the standard DR estimator on their own. It is this small set of high-weight ineligible observations that drives the variance inflation, yet removing their direct contribution via trimming is enough to reverse the sign of the untrimmed estimate. Moreover, 4 of the 7 extreme upper-tail observations survived Abadie's income restriction, illustrating that ad hoc sample restrictions may fail to address instability.

Figure 4 presents the results. The full-sample LATE for net financial assets illustrates how weak overlap can render standard DR estimates unreliable: the standard DR point estimate is −$13,042 (SE = $16,223), suggesting that 401(k) participation *reduces* savings—a conclusion that is at odds with standard economic theory and the existing literature. The untrimmed DR estimator is uninformative in this sample, with a confidence interval spanning from −$45,000 to +$19,000.

DR-BC yields $8,864 (SE = $2,471), recovering a positive, statistically significant effect consistent with prior findings (Abadie, 2003). The sign reversal is not a change in estimand—both estimators target the same LATE—but reflects removing the outsized influence of the two ineligible observations whose effective weights dominate the untrimmed estimator. Decomposing the Wald ratio reveals that the instability is entirely in the reduced form: the first-stage estimate is stable at 0.68 regardless of trimming, but the intention-to-treat (ITT) swings from −$8,879 (SE = $11,044) to $6,035 (SE = $1,686).

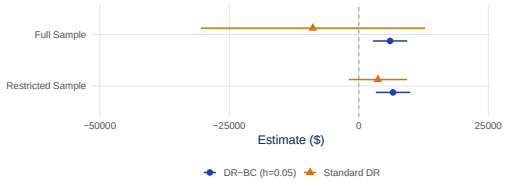
The primary mechanism in this application is variance reduction through trimming, not extrapolation through bias correction. The bias correction is nonetheless essential for theoretical integrity: it ensures the estimator targets the original LATE rather than a trimmed estimand, and it provides the correct influence function for inference. The h-sensitivity analysis in the Supplemental Appendix confirms that DR-BC point estimates are stable across $h \in [0.03, 0.10]$. For total wealth, the standard error reduction is 3.2-fold. In the restricted sample, where overlap is better, gains are more modest (1.35- to 1.70-fold across outcomes and estimands) but DR-BC point estimates are stable across the two samples, suggesting results are not driven by the income restriction.

2.4 Trimming with bias correction to retain the DR property

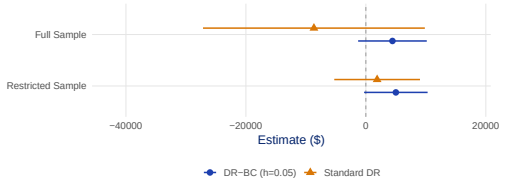
The previous sections show that our DR-BC estimators can yield significant improvements in concrete applications, making inference more precise. But, so far, we have not yet described *how* our DR-BC estimator operates; just that it involves trimming and a bias-correction term that guarantees the DR property for the original causal estimand. In this section, we provide a more informal but intuitive explanation of how our proposed DR-BC estimator works and why it has appealing statistical guarantees.

Each of the DR estimands in Section 2.1, whether for the ATE, LATE, or DiD ATT, can be written as a known function of one or more ratio moments $\alpha_\ell \equiv \mathbb{E}[B_\ell/A_\ell]$, where $\ell \in \{1, \cdots, L\}$

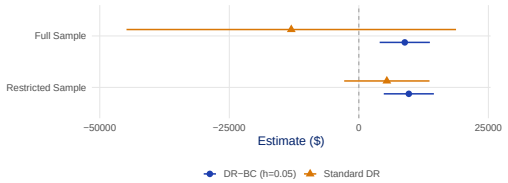
Figure 4: Effect of 401(k) retirement plans on asset accumulation



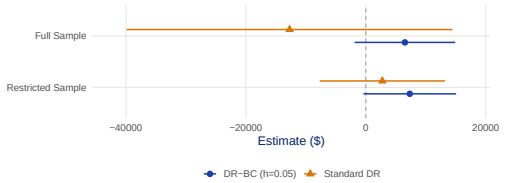
(a) DR ITT estimates of having access to 401(k) on net financial assets



(b) DR ITT estimates of having access to 401(k) on total wealth



(c) DR LATE estimates of 401(k) participation on net financial assets



(d) DR LATE estimates of 401(k) participation on total wealth

Notes. Full sample restricts to households with positive income (Benjamin, 2003; Chernozhukov and Hansen, 2004). Restricted sample further restricts to income between \$10,000 and \$200,000 (Abadie, 2003). DR-BC trims observations with estimated instrument propensity scores $\hat{p}_Z(X)$ outside $[0.05, 0.95]$; Standard DR is the untrimmed DR estimator based on (2.3). Covariates: income, age, age squared, marital status, family size. Points denote estimates; horizontal lines denote 95% confidence intervals.

indexes the ratio moments entering the estimand. Here, A_ℓ involves the propensity score and can be close to zero, and B_ℓ contains a group indicator—multiplied by an outcome residual for the effect-estimating moments, or standing alone for the companion normalization moments. The specific (A_ℓ, B_ℓ) decomposition depends on the design, but the weak overlap problem is always the same: when A_ℓ is near zero, B_ℓ/A_ℓ is unstable with high variance. Standard trimming discards observations with $|A_\ell| < h$ for some threshold $h > 0$. This stabilizes the estimator but changes what is being estimated:

$$\mathbb{E}\left[\frac{B_\ell}{A_\ell} \cdot \mathbb{1}\{|A_\ell| \geqslant h\}\right] \neq \mathbb{E}\left[\frac{B_\ell}{A_\ell}\right].$$

The trimming bias equals $\mathbb{E}\left[(B_\ell/A_\ell) \cdot \mathbb{1}\{|A_\ell| < h\}\right]$, which does not vanish for fixed h. Our bias-correction term uses the behavior of the conditional expectation $\xi_\ell(a) \equiv \mathbb{E}\left[B_\ell \mid A_\ell = a\right]$ near the propensity score boundary to compensate for the discarded observations.

The key observation is that, under either condition for double robustness, the relevant conditional mean at the propensity score boundary equals zero. There are two routes. *First*, when the propensity score is correctly specified, $A_\ell = 0$ places the propensity score at its boundary, so one group has probability zero at that covariate value. The group indicator in B_ℓ then forces $B_\ell = 0$ almost surely conditional on $A_\ell = 0$, giving $\xi_\ell(0) = 0$. *Second*, when the outcome regression is correctly specified, the residual-containing moments have a conditional mean of zero given X. Since A_ℓ is a function of X, the law of iterated expectations gives $\xi_\ell(a) = 0$ for those moments, and in particular $\xi_\ell(0) = 0$. The companion normalization moments need not satisfy this stronger statement, but they are irrelevant along this route because the corresponding residual moments

are already zero. One can verify this for each estimand in Section 2.1; the formal general statement is Assumption 2 in Section 3.

Note the asymmetry: when the outcome regression is correct, $\xi_\ell(a) = 0$ for *all* a for the residual-containing moments, so trimming does not change those estimands and no bias correction is needed. This explains why there is no bias in Panel (a) of Figure 1. The bias correction is operative only when the outcome model is incorrect and the propensity score is correct, which is precisely the case where trimming introduces bias—which, again, explains the bias in Panel (c) of Figure 1. Weak overlap inflates variance under either specification route, but when the outcome regression is correct, trimming removes no information about the estimand, so no bias correction is needed. In either case, $\xi_\ell(0) = 0$ implies that $\xi_\ell(a)/a$ has a removable singularity at zero and can be approximated by a polynomial in a.

Our proposed estimator replaces each problematic ratio moment $\mathbb{E}[B_\ell/A_\ell]$ with its bias-corrected counterpart:

$$\widehat{\alpha}_\ell(h) = \underbrace{\mathbb{E}_n\left[\frac{B_\ell}{A_\ell}\cdot \mathbb{1}\{|A_\ell| \geqslant h\}\right]}_{\text{trimmed mean}} + \underbrace{\sum_{\kappa=1}^{k} \frac{\mathbb{E}_n\left[A_\ell^{\kappa-1}\cdot \mathbb{1}\{|A_\ell| < h\}\right]}{\kappa!}\cdot \widehat{\xi}_\ell^{(\kappa)}(0)}_{\text{bias correction}},$$

where $\widehat{\xi}_\ell^{(\kappa)}(0)$ is the κ-th derivative of a sieve estimator for $\xi_\ell(\cdot)$, evaluated at zero, using a shifted Legendre polynomial basis of maximum degree K, and k is the order of the bias-correction terms. The first part is the standard trimmed estimator. The second part uses the observations excluded by trimming ($|A_\ell| < h$), together with the estimated derivatives, to reconstruct their contribution. The final estimator applies a known function Λ to the corrected moments $(\widehat{\alpha}_1(h), \ldots, \widehat{\alpha}_L(h))$; with the normalized weights in Section 2.1, Λ involves ratios for all three designs.

This structure preserves double robustness. Under either DR condition, $\xi_\ell(0) = 0$ holds for the residual-containing moments under correct outcome regression, and all moments under correct propensity score, so the polynomial approximation of $\xi_\ell(a)/a$ is valid for those moments. As discussed above, normalization moments for which this condition may fail are irrelevant along the outcome-regression route because their companion residual moments are identically zero. Under standard regularity conditions on the propensity score density and the smoothness of ξ_ℓ, the resulting estimator is consistent and asymptotically normal for the original estimand, with a closed-form influence function that enables standard inference; the standard errors account for all sources of estimation uncertainty, including the sieve-based bias correction (Section 3).

In practice, we recommend $h = 0.05$, $k = 1$ (first-order bias correction), and $K = 3$. Figure 1 provides evidence: at $h = 0.05$, coverage remains near the nominal 95% level under both misspecification configurations, while confidence intervals are narrower than at $h = 0$. We find that $k = 1$ with $K = 3$ delivers stable coverage across a wide range of sample sizes and data-generating processes; higher-order bias corrections ($k \geqslant 2$) degrade finite-sample performance because the required sieve derivatives are imprecisely estimated with a fixed low-degree basis. Sensitivity of the

empirical results to the choice of h is examined in Appendix D. Section 3 presents the formal regularity conditions, rate requirements on h, and establishes consistency and asymptotic normality under weak overlap.

2.5 Practical diagnostics and when to use DR-BC

DR-BC and the standard (untrimmed) DR estimator target the same causal parameter, and DR-BC reduces to standard DR when no observations are trimmed ($h = 0$ or no propensity scores near the boundary). The cost of using DR-BC when overlap is adequate is therefore minimal: it adds a polynomial regression step that has a negligible impact on the estimates. The benefit when overlap is weak can be substantial, as the applications above demonstrate. In general, when using our DR-BC estimator, we recommend the following diagnostic checks:

1. *Propensity score distribution.* Plot the estimated propensity score separately for treated and comparison groups. If no observations have $\hat{p}(X)$ near 0 or 1, weak overlap is not a concern and DR-BC will coincide with the standard DR estimator.

2. *Trimming fraction.* Report the share of observations trimmed at the chosen h. When the trimming fraction is small, the bias correction has little to reconstruct and the precision gains come primarily from variance reduction.

3. *Sensitivity to h.* Plot point estimates and confidence intervals as a function of h (as in Figures OA-6–OA-9 of Appendix D). When weak overlap is present, the DR-BC estimate will typically differ from the standard DR estimate at $h = 0$, sometimes substantially—this is expected, since the $h = 0$ estimate is the one inflated by extreme weights. The relevant check is whether DR-BC estimates are stable across moderate values of h (e.g., $h \in [0.03, 0.10]$). Stability across this range is reassuring; instability may indicate model misspecification or that the conditions underlying Assumption 2 are not met, and researchers should exercise caution.

3 Theory

This section formalizes the general estimator introduced in Section 2.4, states the conditions for asymptotic normality, and verifies them for each research design.

3.1 General framework

The DR estimands in Section 2.1 share a common structure. Each can be written as a known function of L ratio moments:

$$\theta_0 = \Lambda\left(\mathbb{E}\left[\frac{B_1(\gamma_0)}{A_1(\gamma_0)}\right], \ldots, \mathbb{E}\left[\frac{B_L(\gamma_0)}{A_L(\gamma_0)}\right]\right),$$

where $(A_\ell(\gamma), B_\ell(\gamma))_{\ell=1}^L$ are known functions of observed data, $\gamma_0 = (m, p)$ collects the true outcome regression m and propensity score p, and Λ is a known function.[5] Weak overlap means some $A_\ell(\gamma_0)$ can be close to zero, making B_ℓ/A_ℓ potentially unstable with high variance.

Given i.i.d. observations $W_1, \ldots, W_n$, a preliminary estimator $\hat{\gamma} = (\hat{m}, \hat{p})$ with probability limit $\gamma^* = (m^*, p^*)$ (so $m^* = m$ when the outcome model is correctly specified, or $p^* = p$ when the propensity score model is correctly specified), a trimming threshold $h > 0$, and positive integers $k \leqslant K$, our estimator is $\hat{\theta} = \Lambda(\hat{\alpha}_1(h, \hat{\gamma}), \ldots, \hat{\alpha}_L(h, \hat{\gamma}))$, where

$$\hat{\alpha}_\ell(h, \gamma) = \mathbb{E}_n\left[\frac{B_\ell(\gamma)}{A_\ell(\gamma)} \mathbb{1}\{|A_\ell(\gamma)| \geqslant h\}\right] + \sum_{\kappa=1}^{k} \frac{\mathbb{E}_n\left[A_\ell(\gamma)^{\kappa-1} \mathbb{1}\{|A_\ell(\gamma)| < h\}\right]}{\kappa!} \cdot \hat{\xi}_\ell^{(\kappa)}(0; \gamma). \tag{3.1}$$

Here $\hat{\xi}_\ell^{(\kappa)}(0; \gamma)$ is the κ-th derivative at zero of a sieve estimator for $\xi_\ell(a; \gamma) \equiv \mathbb{E}\left[B_\ell(\gamma) \mid A_\ell(\gamma) = a\right]$, the conditional mean of B_ℓ given A_ℓ. The dependence on γ was suppressed in Section 2.4. The sieve uses a shifted Legendre polynomial basis $q_K(\cdot)$ of degree K, natural since propensity scores lie in $[0, 1]$; with K fixed, the sieve basis is well-conditioned. The asymptotic theory allows K to grow with n, and we recommend $K = 3$ as a practical default for typical sample sizes in applications. Details are in Appendix E.

3.2 Asymptotic theory

We decompose $\hat{\theta} - \theta_0 = (\hat{\theta} - \theta_h) + (\theta_h - \theta_0)$, where $\theta_h \equiv \theta_h(\gamma_0)$ with $\theta_h(\gamma) = \Lambda(\alpha_1(h, \gamma), \ldots, \alpha_L(h, \gamma))$ is the population analog of $\hat{\theta}$. The first term is the estimation error; the second is the trimming bias. We require the following assumptions.

Assumption 1 (Approximate double robustness)**.** *If either $m^* = m$ or $p^* = p$ (i.e., the outcome regression or propensity score is correctly specified), then $\theta_h(\gamma^*) = \theta_0 + o(h^k)$.*

Assumption 2 (Conditional mean at zero)**.** *For each $\ell = 1, \ldots, L$ with $0 \in \text{support}(A_\ell(\gamma^*))$:* (i) *$\xi_\ell(0; \gamma^*) = 0$ if $p^* = p$, and $\xi_\ell(a; \gamma^*) = 0$ for all a for the residual moments if $m^* = m$;* (ii) *$\xi_\ell(\cdot; \gamma^*)$ is $(k+1)$-times continuously differentiable near 0.*

Assumption 3 (Smoothness of Λ)**.** *$\Lambda(\cdot)$ is twice continuously differentiable near $(\alpha_1(0, \gamma^*), \ldots, \alpha_L(0, \gamma^*))$.*

Assumptions 1–3 are the key substantive conditions. Assumption 1 extends double robustness from θ_0 to its trimmed counterpart θ_h; we verify it for each design in Section 3.3. Assumption 2(i) is the key condition from Section 2.4. When the propensity score is specified correctly, this gives $\xi_\ell(0; \gamma^*) = 0$. When the outcome regression model is specified correctly, the residual-containing moments satisfy $\xi_\ell(a; \gamma^*) = 0$ for all a, so in particular $\xi_\ell(0; \gamma^*) = 0$; the companion normalization moments need not satisfy this statement. Part (ii) controls the polynomial approximation error.

5 Λ may also depend on γ through regression adjustments $\mathbb{E}\left[m(d, X)\right]$ that are smooth, root-n estimable, and not subject to weak overlap; see Section 3.3.

Assumption 3 holds whenever the population moments $\alpha_\ell(0,\gamma^*)$ that appear as denominators in Λ are bounded away from zero; this is a condition on the target estimand (e.g., compliance share bounded away from zero for the LATE, group share bounded away from zero for DiD), not on individual observations, and therefore remains compatible with the weak-overlap regime.

Assumption 4 (Regularity conditions)**.** *The following conditions hold for each $\ell = 1,\dots,L$:*

(a) *First-stage influence function.* $\alpha_\ell(h,\hat\gamma) - \alpha_\ell(h,\gamma^*) = (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\phi_\ell] + o_p(n^{-1/2})$*, where ϕ_ℓ is the influence function of $\alpha_\ell(h,\gamma)$ at γ^*.*

(b) *Sieve estimation. For each $\kappa = 1,\dots,k$,*

$$\hat\xi_\ell^{(\kappa)}(0;\gamma^*) - \xi_\ell^{(\kappa)}(0;\gamma^*) - (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\psi_{\ell,\kappa}(\gamma^*)] = o_p(n^{-1/2}h^{1-\kappa}),$$

where $\psi_{\ell,\kappa}(\gamma) = q_K^{(\kappa)}(0)'\,\mathbb{E}\left[q_K(A_\ell(\gamma))q_K(A_\ell(\gamma))'\right]^{-1} q_K(A_\ell(\gamma))(B_\ell(\gamma) - \xi_\ell(A_\ell(\gamma);\gamma))$.

(c) *Moment bound.* $\mathbb{E}[\omega_\ell(h,\gamma^*)^2] = o(n^{1/2})$*, where ω_ℓ is defined in (3.2) below.*

(d) *Stochastic equicontinuity.* $\hat\alpha_\ell(h,\hat\gamma) - \alpha_\ell(h,\hat\gamma) - \hat\alpha_\ell(h,\gamma^*) + \alpha_\ell(h,\gamma^*) = o_p(n^{-1/2})$.

(e) *Rate condition.* $nh^{2k} = O(1)$ *as* $n \to \infty$.

Parts (a)–(d) require the first-stage estimator, sieve regression, influence function moments, and the sample criterion to behave well; lower-level sufficient conditions (e.g., for parametric first stages) are in Appendix E. Part (e) imposes an upper bound on h that controls the trimming bias; a complementary lower bound of the form $nh^4 \to \infty$, needed to control the variance of the sieve bias-correction term, is stated in Appendix F.

The influence function ω_ℓ of each moment $\hat\alpha_\ell$ has four components:

$$\omega_\ell(h,\gamma) = \underbrace{\frac{B_\ell(\gamma)}{A_\ell(\gamma)}\mathbb{1}\{|A_\ell(\gamma)| \geqslant h\}}_{\text{trimmed ratio}} + \underbrace{\sum_{\kappa=1}^{k} \frac{A_\ell(\gamma)^{\kappa-1}\mathbb{1}\{|A_\ell(\gamma)| < h\}}{\kappa!} \cdot \xi_\ell^{(\kappa)}(0;\gamma)}_{\text{bias correction}}$$
$$+ \underbrace{\sum_{\kappa=1}^{k} \frac{\mathbb{E}\left[A_\ell(\gamma)^{\kappa-1}\mathbb{1}\{|A_\ell(\gamma)| < h\}\right]}{\kappa!} \cdot \psi_{\ell,\kappa}(\gamma)}_{\text{sieve estimation}} + \underbrace{\phi_\ell}_{\text{first stage}}. \tag{3.2}$$

Let Λ_ℓ denote the partial derivative of Λ with respect to its ℓ-th argument, and define the overall influence function

$$\varphi = \sum_{\ell=1}^{L} \Lambda_\ell(\alpha_1(h,\gamma^*),\dots,\alpha_L(h,\gamma^*))\,\omega_\ell(h,\gamma^*). \tag{3.3}$$

Note that Λ_ℓ is evaluated at the population moments $\alpha_\ell(h,\gamma^*)$—the probability limits of the bias-corrected moment estimators—rather than at $\alpha_\ell(0,\gamma^*)$; these are the correct expansion points for the delta method because, under the outcome-correct route, $\alpha_\ell(h,\gamma^*)$ need not equal $\alpha_\ell(0,\gamma^*)$ for

the normalization moments. Under Assumption 1, when h goes to zero at the rate in Assumption 4(e), $\alpha_\ell(h, \gamma^*)$ converges to $\alpha_\ell(0, \gamma^*)$ and the influence function targets θ_0.

Theorem 3.1. *Under Assumptions 1–4:*

(i) $\hat{\theta} - \theta_0 = (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\varphi] + o_p(n^{-1/2})$.

(ii) *If in addition* $\mathbb{E}[\varphi^2]$ *is bounded away from zero and* $\mathbb{E}[|\varphi - \mathbb{E}[\varphi]|^{2+\eta}]/(n^{\eta/2}\,\mathbb{E}[(\varphi - \mathbb{E}[\varphi])^2]^{(2+\eta)/2}) = o(1)$ *for some* $\eta > 0$, *then* $(\hat{\theta} - \theta_0) \Big/ \sqrt{\mathbb{E}[(\varphi - \mathbb{E}[\varphi])^2]/n} \xrightarrow{d} \mathcal{N}(0, 1)$.

A proof is in Appendix G. A key difficulty relative to Sasaki and Ura (2022) is that θ_0 is a nonlinear function of multiple ratio moments with heterogeneous convergence rates, so the standard delta method does not directly apply. The variance is estimated by $\hat{\sigma}^2 = \mathbb{E}_n[(\hat{\varphi} - \mathbb{E}_n[\hat{\varphi}])^2]$, where $\hat{\varphi}$ plugs sample analogs into (3.2)–(3.3).

Remark 3.1 (Diverging variance)**.** Assumption 4(c) allows $\mathbb{E}[\omega_\ell^2]$ to diverge, accommodating heavy tails from B_ℓ/A_ℓ when h is small. The Lyapunov condition ensures the CLT holds despite this.

3.3 Design-specific verification

We now verify Assumptions 1–3 for each design from Section 2.1. Table 1 summarizes the (A_ℓ, B_ℓ) decompositions; the resulting estimators specialize (3.1) to each case.

Table 1: (A_ℓ, B_ℓ) Decompositions by Research Design

ℓ	$B_\ell(\gamma^*)$	$A_\ell(\gamma^*)$
	Panel A. Average Treatment Effect (ATE)	
1	$D\big(Y - m_Y^*(1, X)\big)$	$p_D^*(X)$
2	D	$p_D^*(X)$
3	$(1 - D)\big(Y - m_Y^*(0, X)\big)$	$1 - p_D^*(X)$
4	$1 - D$	$1 - p_D^*(X)$
	Panel B. Local Average Treatment Effect (LATE)	
1	$Z\big(Y - m_Y^{*\mathrm{LATE}}(1, X)\big)$	$p_Z^*(X)$
2	Z	$p_Z^*(X)$
3	$(1 - Z)\big(Y - m_Y^{*\mathrm{LATE}}(0, X)\big)$	$1 - p_Z^*(X)$
4	$1 - Z$	$1 - p_Z^*(X)$
5	$Z\big(D - m_D^{*\mathrm{LATE}}(1, X)\big)$	$p_Z^*(X)$
6	Z	$p_Z^*(X)$
7	$(1 - Z)\big(D - m_D^{*\mathrm{LATE}}(0, X)\big)$	$1 - p_Z^*(X)$
8	$1 - Z$	$1 - p_Z^*(X)$
	Panel C. Difference-in-Differences (DiD)	
1	$\mathbb{1}\{G = g\}\big(Y_t - Y_{g-\delta-1} - m_{g,t}^*(X)\big)$	1
2	$p_{g,t}^*(X)\, C_{g,t}\big(Y_t - Y_{g-\delta-1} - m_{g,t}^*(X)\big)$	$1 - p_{g,t}^*(X)$
3	$\mathbb{1}\{G = g\}$	1
4	$p_{g,t}^*(X)\, C_{g,t}$	$1 - p_{g,t}^*(X)$

Example 1 (ATE under unconfoundedness)**.** The ATE decomposition uses $L = 4$; see Table 1. Weak overlap arises when $p_D(X)$ is near zero or one.

Proposition 3.1. *Suppose:* (i) *for each* $d \in \{0, 1\}$,

$$\mathbb{E}\left[p_D(X)^d \left(1 - p_D(X)\right)^{1-d} \big(m_Y(d, X) - m_Y^*(d, X)\big) \,\middle|\, p_D^*(X) = 1 - d\right] = 0;$$

and (ii) *the functions*

$$s \mapsto \mathbb{E}\left[p_D(X)\big(m_Y(1, X) - m_Y^*(1, X)\big) \mid p_D^*(X) = s\right],$$
$$s \mapsto \mathbb{E}\left[\left(1 - p_D(X)\right)\big(m_Y(0, X) - m_Y^*(0, X)\big) \mid 1 - p_D^*(X) = s\right]$$

are $(k + 1)$*-times continuously differentiable near* 0*. Then Assumptions 1–3 hold for the ATE.*

The bias-corrected ATE estimator is $\hat{\theta} = \mathbb{E}_n\left[\hat{m}_Y(1, X) - \hat{m}_Y(0, X)\right] + \hat{\alpha}_1/\hat{\alpha}_2 - \hat{\alpha}_3/\hat{\alpha}_4$, where each $\hat{\alpha}_\ell = \hat{\alpha}_\ell(h, \hat{\gamma})$ is as in (3.1). Trimming discards observations with $\hat{p}_D(X) < h$ or $1 - \hat{p}_D(X) < h$.

Example 2 (LATE under instrumental variables)**.** The LATE decomposition uses $L = 8$, with moments $\ell = 1, \ldots, 4$ for the reduced form and $\ell = 5, \ldots, 8$ for the first stage; see Table 1. The function Λ is the Wald ratio.

Proposition 3.2. *Suppose:* (i) *for each* $z \in \{0, 1\}$ *and each* $R \in \{Y, D\}$,

$$\mathbb{E}\left[p_Z(X)^z \left(1 - p_Z(X)\right)^{1-z} \big(m_R^{LATE}(z, X) - m_R^{*LATE}(z, X)\big) \,\middle|\, p_Z^*(X) = 1 - z\right] = 0;$$

(ii) *the functions*

$$s \mapsto \mathbb{E}\left[p_Z(X)\big(m_R^{LATE}(1, X) - m_R^{*LATE}(1, X)\big) \mid p_Z^*(X) = s\right],$$
$$s \mapsto \mathbb{E}\left[\left(1 - p_Z(X)\right)\big(m_R^{LATE}(0, X) - m_R^{*LATE}(0, X)\big) \mid 1 - p_Z^*(X) = s\right]$$

are $(k + 1)$*-times continuously differentiable near* 0 *for each* $R \in \{Y, D\}$*; and* (iii) *the compliance share* $\mathbb{E}\left[D(1) - D(0)\right]$ *is bounded away from zero. Then Assumptions 1–3 hold for the LATE.*

The structure parallels the ATE case, with the instrument Z replacing D and the instrument propensity score $p_Z(X)$ replacing $p_D(X)$. Condition (i) holds under either DR condition: when $p_Z^* = p_Z$, the boundary $p_Z(X) = 0$ or $p_Z(X) = 1$ forces Z or $1 - Z$ to zero a.s., so $B_\ell = 0$; when $m_R^{*\mathrm{LATE}} = m_R^{\mathrm{LATE}}$, the outcome residual in B_ℓ has conditional mean zero given X, so $\xi_\ell(a; \gamma^*) = 0$ for all a. Condition (iii) ensures the Wald ratio denominator is well-defined.

The bias-corrected LATE estimator is the Wald ratio of two bias-corrected DR expressions:

$$\hat{\theta} = \frac{\mathbb{E}_n\left[\hat{m}_Y^{\mathrm{LATE}}(1, X) - \hat{m}_Y^{\mathrm{LATE}}(0, X)\right] + \hat{\alpha}_1/\hat{\alpha}_2 - \hat{\alpha}_3/\hat{\alpha}_4}{\mathbb{E}_n\left[\hat{m}_D^{\mathrm{LATE}}(1, X) - \hat{m}_D^{\mathrm{LATE}}(0, X)\right] + \hat{\alpha}_5/\hat{\alpha}_6 - \hat{\alpha}_7/\hat{\alpha}_8}.$$

The ITT effect is the numerator alone.

Example 3 (Staggered DiD)**.** The DiD ATT(g, t) decomposition uses $L = 4$, with $A_1 = A_3 = 1$ (trivial denominators); only α_2 and α_4 involve weak overlap through $A_\ell = 1 - p_{g,t}(X)$. See

Table 1. Staggered DiD identification relies on irreversible treatment, δ-limited anticipation, and conditional parallel trends as discussed in Callaway and Sant'Anna (2021). Let $\delta \geqslant 0$ denote the number of anticipation periods. The key assumptions are: (i) irreversibility ($D_{t-1} = 1 \Rightarrow D_t = 1$ a.s.); (ii) limited anticipation (for $t < g - \delta$, $Y_t(g) = Y_t(\infty)$ a.s.); and (iii) conditional parallel trends (for the relevant comparison group $\mathcal{C}_{g,t}$, $\mathbb{E}\left[Y_t(\infty) - Y_{g-\delta-1}(\infty) \mid X, G = g\right] = \mathbb{E}\left[Y_t(\infty) - Y_{g-\delta-1}(\infty) \mid X, C_{g,t} = 1\right]$ a.s.). One can rely on never-treated or not-yet-treated comparison groups, and on weighted versions of these assumptions as discussed in Baker et al. (2025); we omit the details to avoid repetition.

Proposition 3.3. *Suppose that* $0 < \mathbb{E}\left[\mathbb{1}\{G = g\}\right] < 1$ *and:* (i)

$$\mathbb{E}\left[C_{g,t}\big(m_{g,t}(X) - m^*_{g,t}(X)\big) \mid p^*_{g,t}(X) = 1\right] = 0;$$

and (ii) *the relevant conditional expectations are* $(k+1)$*-times continuously differentiable near* 0. *Then Assumptions 1–3 hold for* $ATT(g,t)$ *and for* $ES(e) = \sum_g w^{es}_{g,e} \cdot ATT(g, g+e)$.

When $p^*_{g,t} = p_{g,t}$, $p_{g,t}(X) = 1$ implies $C_{g,t} = 0$ a.s., so $B_2 = B_4 = 0$. When $m^*_{g,t} = m_{g,t}$, the residual moment $\ell = 2$ has conditional mean zero among comparison units, so $\xi_2(0; \gamma^*) = 0$. The companion normalization moment $\ell = 4$ need not satisfy this stronger statement.

The bias-corrected estimator for each group-time ATT is

$$\hat{\theta}_{g,t} = \frac{\mathbb{E}_n\left[\mathbb{1}\{G = g\}(Y_t - Y_{g-\delta-1} - \hat{m}_{g,t}(X))\right]}{\mathbb{E}_n\left[\mathbb{1}\{G = g\}\right]} - \frac{\hat{\alpha}_2(h, \hat{\gamma}_{g,t})}{\hat{\alpha}_4(h, \hat{\gamma}_{g,t})},$$

where $\hat{\alpha}_2$ and $\hat{\alpha}_4$ trim on $1 - \hat{p}_{g,t}(X) < h$. Event-study parameters aggregate across groups: $\hat{\theta}(e) = \sum_g \hat{w}_{g,e}\hat{\theta}_{g,g+e}$, with standard errors from $\hat{\varphi}(e) = \sum_g \hat{w}_{g,e}\hat{\varphi}_{g,g+e}$, where inference is conducted conditional on the observed cohort shares (i.e., treating $\hat{w}_{g,e}$ as fixed). Simultaneous confidence bands are constructed via multiplier bootstrap following (Callaway and Sant'Anna, 2021).

4 Conclusion

This paper proposes a class of doubly robust estimators that use trimming for variance reduction while preserving the original causal estimand through sieve-based bias correction. The key insight is that under either double robustness condition, the conditional expectation of the moment numerator given the denominator equals zero at the trimming boundary, enabling polynomial reconstruction of the trimmed observations' contribution. The resulting estimators apply across unconfoundedness, instrumental variables, and DiD designs without changing the causal question.

In our applications, standard errors from the untrimmed DR estimator are 1.8 to 6.6 times larger than those from DR-BC, corresponding to variance reductions of up to 43-fold. These gains translate directly into tighter confidence intervals and higher power, and are especially valuable in settings where weak overlap would otherwise render DR estimates uninformative.

We recommend $h = 0.05$, $k = 1$, and $K = 3$ as practical defaults. Practitioners should verify that the residual bias $O(h^2)$ is small relative to the standard error and examine sensitivity to h when the trimming fraction is large. For DiD, we integrate DR-BC into the `did` R package (Callaway and Sant'Anna, 2021). We also plan to release easy-to-use software for LATE and unconfoundedness-based DR estimators.

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 2003, *113* (2), 231–263.

Baker, Andrew, Brantly Callaway, Scott Cunningham, Andrew Goodman-Bacon, and Pedro HC Sant'Anna, "Difference-in-Differences Designs: A Practitioner's Guide," *Journal of Economic Literature*, 2025. Forthcoming.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

Benjamin, Daniel J., "Does 401(k) eligibility increase saving? Evidence from propensity score subclassification," *Journal of Public Economics*, 2003, *87* (5-6), 1259–1290.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, "When is TSLS Actually LATE?," *Review of Economic Studies*, 2026. Forthcoming.

Caetano, Carolina and Brantly Callaway, "Difference-in-Differences when Parallel Trends Holds Conditional on Covariates," *arXiv:2406.15288*, 2024.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Chernozhukov, Victor and Christian Hansen, "The impact of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis," *The Review of Economics and Statistics*, 2004, *86* (3), 735–751.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A Mitnik, "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand," Technical Working Paper 330, National Bureau of Economic Research, Cambridge, MA September 2006.

_ , _ , _ , and _ , "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 2009, *96* (1), 187–199.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon, "Overlap in observational studies with high-dimensional covariates," *Journal of Econometrics*, 2021, *221* (1), 644–654.

Dias, Mateus and Luiz Felipe Fontes, "The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil," *American Economic Journal: Economic Policy*, 2024, *16* (3), 257–289.

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

Heiler, Phillip and Ekaterina Kazak, "Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores," *Journal of Econometrics*, 2021, *222* (2), 1083–1108.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, 1986, *81* (396), 945–960.

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

Kang, Joseph D. Y. and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.," *Statistical Science*, 2007, *22* (4), 569–573.

Khan, Shakeeb and Elie Tamer, "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

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

Mogstad, Magne and Alexander Torgovitsky, "Chapter 1 - Instrumental variables with unobserved heterogeneity in treatment effects," in Christian Dustmann and Thomas Lemieux, eds., *Handbook of Labor Economics*, Vol. 5, Elsevier, 2024, pp. 1–114.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 1994, *89* (427), 846–866.

Sant'Anna, Pedro HC and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

Sasaki, Yuya and Takuya Ura, "Estimation and inference for moments of ratios with robustness against large trimming bias," *Econometric Theory*, 2022, *38* (1), 66–112.

Słoczyński, Tymon, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economics and Statistics*, 2022, *104* (3), 501–509.

_ , "When Should We (Not) Interpret Linear IV Estimands as LATE?," *Review of Economic Studies*, 2026. Accepted.

_ **and Jeffrey M. Wooldridge**, "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 2018, *34* (1), 112–133.

_ **, S. Derya Uysal, and Jeffrey M. Wooldridge**, "Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment," 2022. Working paper, arXiv:2208.01300.

_ **,** _ **, and** _ , "Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 2025, *43* (1), 164–177.

Yang, S and P Ding, "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores," *Biometrika*, 2018, *105* (2), 487–493.

Doubly Robust Estimators with Weak Overlap: Supplemental Appendix

Yukun Ma Pedro H.C. Sant'Anna Yuya Sasaki Takuya Ura

April 8, 2026

This Supplemental Appendix provides additional results for "Doubly Robust Estimators with Weak Overlap." Section A presents the data-generating process for the simulation study in Section 2.2 of the main text. Section B reports event-study estimates for additional outcomes in the Dias–Fontes DiD application discussed in Section 2.3.1. Section C plots the propensity score distribution in Fryer-Levitt application. Section D presents sensitivity analyses of the empirical results to the choice of trimming threshold h. Sections E–H contain the technical material: sieve estimation details and lower-level sufficient conditions for Assumption 4, the proof of Theorem 3.1, and proofs of Propositions 3.1–3.3.

A Simulation Design

This section provides the details of the data-generating process (DGP) used in Figure 1. Our design builds on Kang and Schafer (2007) and Sant'Anna and Zhao (2020), adapted to a panel DiD setting with two periods and two groups.

A.1 Data-generating process

We generate $n = 10{,}000$ independent units. For each unit i, draw $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ with each component independently Student-t with 10 degrees of freedom. Define the nonlinear transformations

$$Z_i = \big(Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}\big)' = \left(\frac{X_{i1}}{\sigma_1},\ \frac{X_{i1}^2 - X_{i2}^2}{\sigma_2},\ \frac{X_{i3}^3}{\sigma_3},\ \frac{X_{i4}^3}{\sigma_4}\right)',$$

where σ_j is the standard deviation of the j-th component (computed from the Student-t moments), so that each Z_{ij} has unit variance. Define two index functions:

$$f_{\text{reg}}(W) = 1 + W_1 + W_2 + W_3 + W_4,$$
$$f_{\text{ps}}(W) = 1.5 + W_1 + W_2 + W_3 + W_4,$$

for a generic argument $W = (W_1, W_2, W_3, W_4)'$.

Treatment group membership is determined by a logistic model: $D_i = \mathbb{1}\{p(W_i^{\text{ps}}) \geqslant U_i\}$, where $p(w) = \exp(f_{\text{ps}}(w))/(1+\exp(f_{\text{ps}}(w)))$ and $U_i \sim \text{Uniform}(0,1)$. The potential outcomes

follow a panel structure:

$$Y_{i,0} = f_{\text{reg}}(W_i^{\text{reg}}) + \upsilon_i + \varepsilon_{i,0},$$
$$Y_{i,1}(d) = 2\, f_{\text{reg}}(W_i^{\text{reg}}) + \upsilon_i + \varepsilon_{i,1}(d), \quad d \in \{0,1\},$$

where $\varepsilon_{i,0}$, $\varepsilon_{i,1}(0)$, $\varepsilon_{i,1}(1) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and $\upsilon_i \sim \mathcal{N}(D_i \cdot f_{\text{reg}}(W_i^{\text{reg}}), 1)$ captures unit-level unobserved heterogeneity that depends on the treatment status and the regression index. The observed outcomes are $Y_{i,0}$ (pre-treatment) and $Y_{i,1} = D_i\, Y_{i,1}(1) + (1 - D_i)\, Y_{i,1}(0)$ (post-treatment). By construction, $\mathbb{E}[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1] = 0$, so the true ATT is zero.

A.2 Misspecification configurations

The key feature of this design is that X and Z are nonlinearly related: a researcher who uses X when the true model depends on Z (or vice versa) faces misspecification. We consider four configurations:

Configuration	W^{ps}	W^{reg}	Interpretation
DGP 1	Z	Z	Both models correctly specified
DGP 2	X	Z	PS misspecified, outcome correct
DGP 3	Z	X	PS correct, outcome misspecified
DGP 4	X	X	Both models misspecified

In each configuration, the researcher observes (X, D, Y_0, Y_1) and estimates the propensity score and outcome regression using Z (not X).

A.3 Weak overlap

The logistic index f_{ps} with an intercept of 1.5, combined with the heavy-tailed Student-t covariates, produces a non-negligible fraction of comparison-group units with propensity scores close to one. With $df = 10$ degrees of freedom, approximately 5–10% of comparison units have $\hat{p}(X) > 0.95$, creating the weak overlap that motivates trimming.

A.4 Estimation details

For each Monte Carlo draw, we estimate the ATT using the DR DiD estimator of Callaway and Sant'Anna (2021):

1. Estimate the propensity score via logistic regression of D on Z.

2. Estimate the outcome regression via OLS of $Y_1 - Y_0$ on Z among controls ($D = 0$).

3. Compute the DR DiD ATT using the sample analogs of (2.5), with comparison-group weights determined by the estimated propensity score.

We compare three estimators across a grid of trimming thresholds

$$h \in \{0,\ 0.01,\ 0.025,\ 0.05,\ 0.10\}:$$

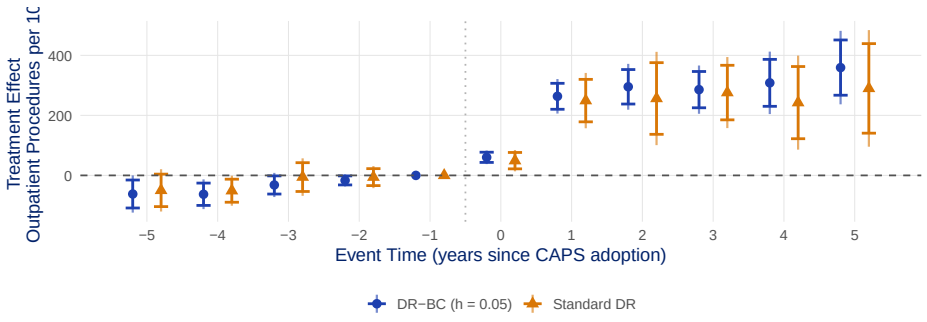
(1) the standard (untrimmed) DR estimator ($h = 0$); (2) the trimmed DR estimator without bias correction (DR-Trim); and (3) our proposed DR-BC estimator with $k = 1$ and $K = 3$.

Figure 1 in the main text presents results for DGPs 2 and 3 (the two single-misspecification cases that test double robustness). Results for DGPs 1 and 4 are qualitatively similar: under DGP 1 (both models correct), all estimators perform well, and DR-BC improves precision; under DGP 4 (both models misspecified), all estimators are inconsistent, as expected. We run 10,000 Monte Carlo repetitions.

B Additional DiD Outcomes: Dias–Fontes Application

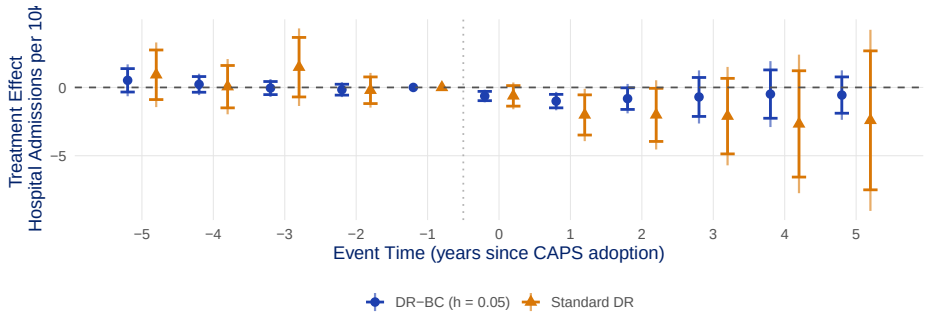
The main paper presents event-study estimates for mental health practitioners (Figure 2, Panel (b)), where we set $\delta = 1$ because the original analysis of Dias and Fontes (2024) finds evidence of anticipation for that outcome. For the remaining four outcomes, figures OA-1–OA-4, no such evidence is present, so we set $\delta = 0$. In all specifications, the comparison group consists of not-yet-treated municipalities, and the covariate set includes state fixed effects and 29 baseline characteristics.

Figure OA-1: Effect of CAPS adoption on outpatient mental health procedures per 10,000 population



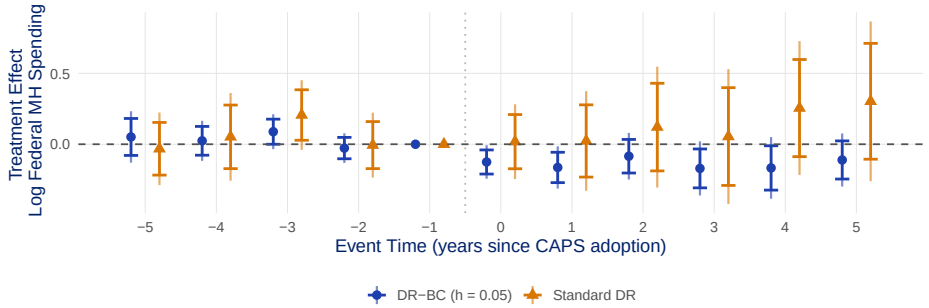
Notes. Event-study estimates using the DR DiD framework of Callaway and Sant'Anna (2021) with not-yet-treated municipalities as the comparison group and no anticipation ($\delta = 0$). DR-BC is our proposed estimator with $h = 0.05$; Standard DR is the untrimmed estimator. Points denote point estimates, vertical bars denote 95% simultaneous confidence bands, and error bars with caps denote 95% pointwise confidence intervals. Data from Dias and Fontes (2024). Sample starts in 2008.

Figure OA-2: Effect of CAPS adoption on mental health hospital admissions per 10,000 population



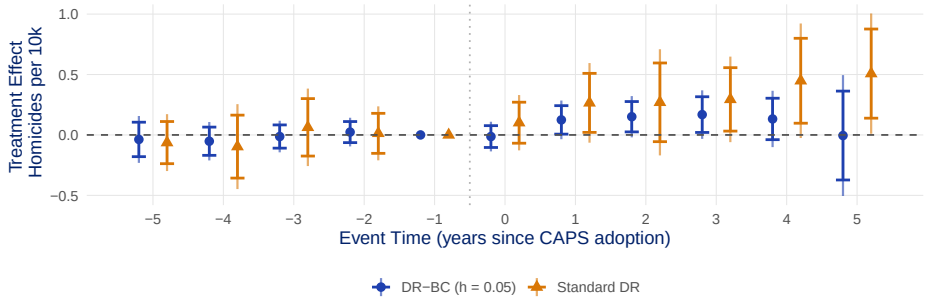
Notes. Event-study estimates using the DR DiD framework of Callaway and Sant'Anna (2021) with not-yet-treated municipalities as the comparison group and no anticipation ($\delta = 0$). DR-BC is our proposed estimator with $h = 0.05$; Standard DR is the untrimmed estimator. Points denote point estimates, vertical bars denote 95% simultaneous confidence bands, and error bars with caps denote 95% pointwise confidence intervals. Data from Dias and Fontes (2024). Sample starts in 2002.

Figure OA-3: Effect of CAPS adoption on log federal mental health hospital spending



Notes. Event-study estimates using the DR DiD framework of Callaway and Sant'Anna (2021) with not-yet-treated municipalities as the comparison group and no anticipation ($\delta = 0$). DR-BC is our proposed estimator with $h = 0.05$; Standard DR is the untrimmed estimator. Points denote point estimates, vertical bars denote 95% simultaneous confidence bands, and error bars with caps denote 95% pointwise confidence intervals. Data from Dias and Fontes (2024). Sample starts in 2002.

Figure OA-4: Effect of CAPS adoption on homicides per 10,000 population

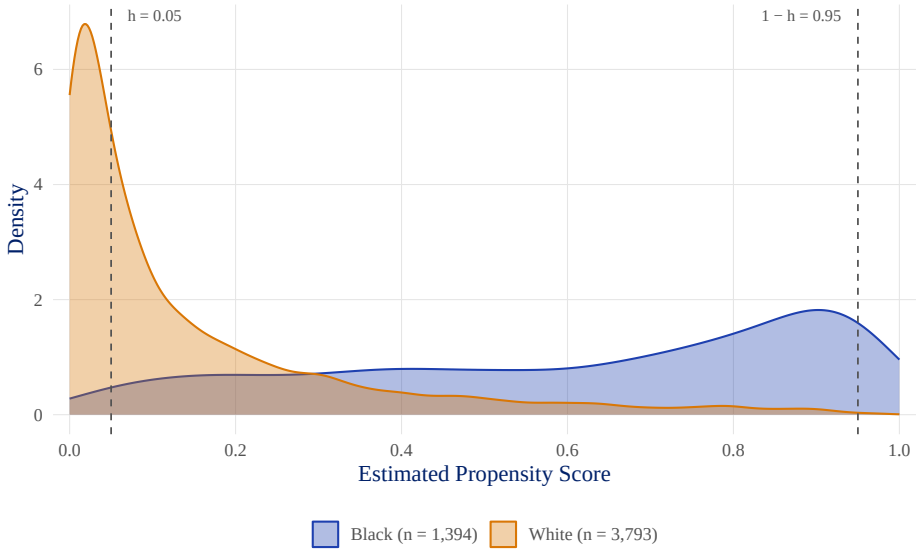


Notes. Event-study estimates using the DR DiD framework of Callaway and Sant'Anna (2021) with not-yet-treated municipalities as the comparison group and no anticipation ($\delta = 0$). DR-BC is our proposed estimator with $h = 0.05$; Standard DR is the untrimmed estimator. Points denote point estimates, vertical bars denote 95% simultaneous confidence bands, and error bars with caps denote 95% pointwise confidence intervals. Data from Dias and Fontes (2024). Sample starts in 2002.

C Propensity Score Distribution: Fryer–Levitt Application

Figure OA-5 shows the estimated propensity score distribution for the Fryer–Levitt application. The interviewer fixed effects create many small cells in which only one race is observed, pushing the propensity score toward its boundaries for a non-negligible share of observations. The trimming thresholds $h = 0.05$ and $1 - h = 0.95$ are indicated by dashed vertical lines.

Figure OA-5: Propensity score distribution: Black–White test score gap (Fryer–Levitt)



Notes. Density of estimated propensity scores from a logit model of race (Black vs. White) on the full set of covariates used by Fryer and Levitt (2013), including interviewer fixed effects. Dashed vertical lines mark the trimming thresholds $h = 0.05$ and $1 - h = 0.95$. Observations with estimated propensity scores below h or above $1 - h$ fall in a covariate cell where trimming operates; at $h = 0.05$, roughly 38% and 35% of observations at 9 and 24 months fall within such cells, driven by interviewer fixed effects that make race nearly deterministic within interviewer cells. However, only a small fraction carry unstable inverse-probability weights: 36 observations (0.69%) at 9 months and 38 observations (0.73%) at 24 months lie on the "active" arm of the trimmed moment (i.e., treated observations with $\hat{p}(X) < 0.05$ or control observations with $\hat{p}(X) > 0.95$). The remaining observations in the trimmed region share a covariate cell with these few high-weight units but carry benign weights close to one.

D Sensitivity to h

Figures OA-6–OA-9 display estimates and 95% confidence intervals as functions of $h \in [0, 0.10]$ for each application.

A common pattern across applications is a substantial change in point estimates between $h = 0$ (the standard untrimmed DR estimator) and small positive h (e.g., $h = 0.02$ or $h = 0.03$), followed by approximate stability for $h \geqslant 0.03$. This pattern is expected and informative: the jump reflects the removal of a small number of observations with extreme

weights that dominate the untrimmed estimator, while the subsequent stability indicates that the bias correction is adequately reconstructing the trimmed observations' contribution across a range of thresholds. It is important to distinguish two notions of "trimming." The share of observations whose arm-specific IPW weight is unstable—treated units with $\hat{p}(X) < h$ or control units with $\hat{p}(X) > 1 - h$—is what actually drives the variance of the standard DR estimator and the sharp initial jump. This share is small across applications: 36–38 observations (0.69–0.73%) in Fryer–Levitt, 2 ineligible observations with effective weights above 30 in the full-sample 401(k), and 177 comparison observations (0.8% on average, up to 3.3% in the most affected group-time cell) in Medicaid. The share of observations whose covariate cell places them in the trimmed region is typically larger—as in Fryer–Levitt, where 35–38% of the sample shares a covariate cell with a high-weight unit—because trimming operates at the covariate-cell level. These additional observations carry near-unit effective weights and would not destabilize the standard DR estimator on their own; they are reconstructed by the bias correction to preserve the target estimand. The relevant diagnostic is not whether the DR-BC estimate at $h = 0.05$ matches the standard DR estimate at $h = 0$—it generally will not, since the standard DR estimate is inflated by extreme weights—but whether the DR-BC estimate is stable across h for moderate values of h. Stability across $h \in [0.03, 0.10]$ suggests the polynomial approximation is adequate and the results are not driven by the specific choice of threshold. Instability across this range would suggest model misspecification or that the conditions underlying Assumption 2 are not met, and researchers should exercise caution in such cases.

E Sieve Estimation Details

The sieve estimator for $\xi_\ell(\cdot;\gamma)$ uses the shifted orthonormal Legendre polynomial basis of degree K:

$$q_K(a) = \big(1,\ \sqrt{3}(2a-1),\ \sqrt{5}(6a^2-6a+1),\ \sqrt{7}(20a^3-30a^2+12a-1),\ \ldots\big)'.$$

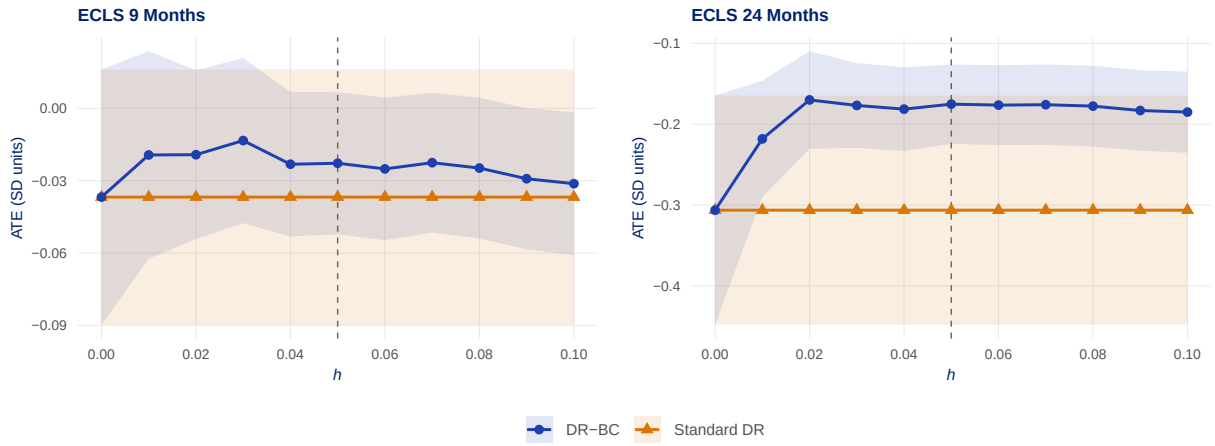
These polynomials are orthonormal on $[0,1]$, making q_K natural for propensity scores. The sieve estimator for the κ-th derivative at zero is

$$\hat{\xi}_\ell^{(\kappa)}(0;\gamma) = q_K^{(\kappa)}(0)'\,\mathbb{E}_n\left[q_K(A_\ell(\gamma))q_K(A_\ell(\gamma))'\right]^{-1}\mathbb{E}_n\left[q_K(A_\ell(\gamma))B_\ell(\gamma)\right].$$

The influence function of this sieve estimator, centered at its population analog $\xi_\ell^{(\kappa)}(0;\gamma^*)$, is

$$\psi_{\ell,\kappa}(\gamma^*) = q_K^{(\kappa)}(0)'\,\mathbb{E}\left[q_K(A_\ell(\gamma^*))q_K(A_\ell(\gamma^*))'\right]^{-1} q_K(A_\ell(\gamma^*))\big(B_\ell(\gamma^*) - \xi_\ell(A_\ell(\gamma^*);\gamma^*)\big),$$

Figure OA-6: Sensitivity to h: Black–White Test Score Gap (Fryer–Levitt)



Notes. DR-BC (blue) and Standard DR (gold) ATE estimates with 95% confidence intervals (shaded bands) as a function of $h \in [0, 0.10]$. The dashed vertical line marks $h = 0.05$, the recommended default. DR-BC estimates are approximately flat for $h \geqslant 0.03$, while the Standard DR estimate (at $h = 0$) has a much wider confidence interval. At 24 months, the SE drops from 0.072 to 0.025 as h increases from 0 to 0.05.

which is the $\psi_{\ell,\kappa}(\gamma^*)$ appearing in Assumption 4(b) and the influence function (3.2).

The following lemma, based on Belloni, Chernozhukov, Chetverikov and Kato (2015), gives lower-level sufficient conditions for Assumption 4(b). See also Chen and Christensen (2015).

Lemma E.1. *For each $\ell = 1, \ldots, L$ and $\kappa = 1, \ldots, k$, suppose:*

(i) *The eigenvalues of $\mathbb{E}\left[q_K(A_\ell(\gamma^*))q_K(A_\ell(\gamma^*))'\right]$ are bounded above and away from zero.*

(ii) $\sqrt{\log K}\,(K + K^{5/2-s})\,\|q_K^{(\kappa)}(0)\| = o(h^{1-\kappa}n^{1/2})$, *where s is the smoothness order of $\xi_\ell(\cdot;\gamma^*)$.*

(iii) $K^{1-s}\|q_K^{(\kappa)}(0)\| = o(h^{1-\kappa})$.

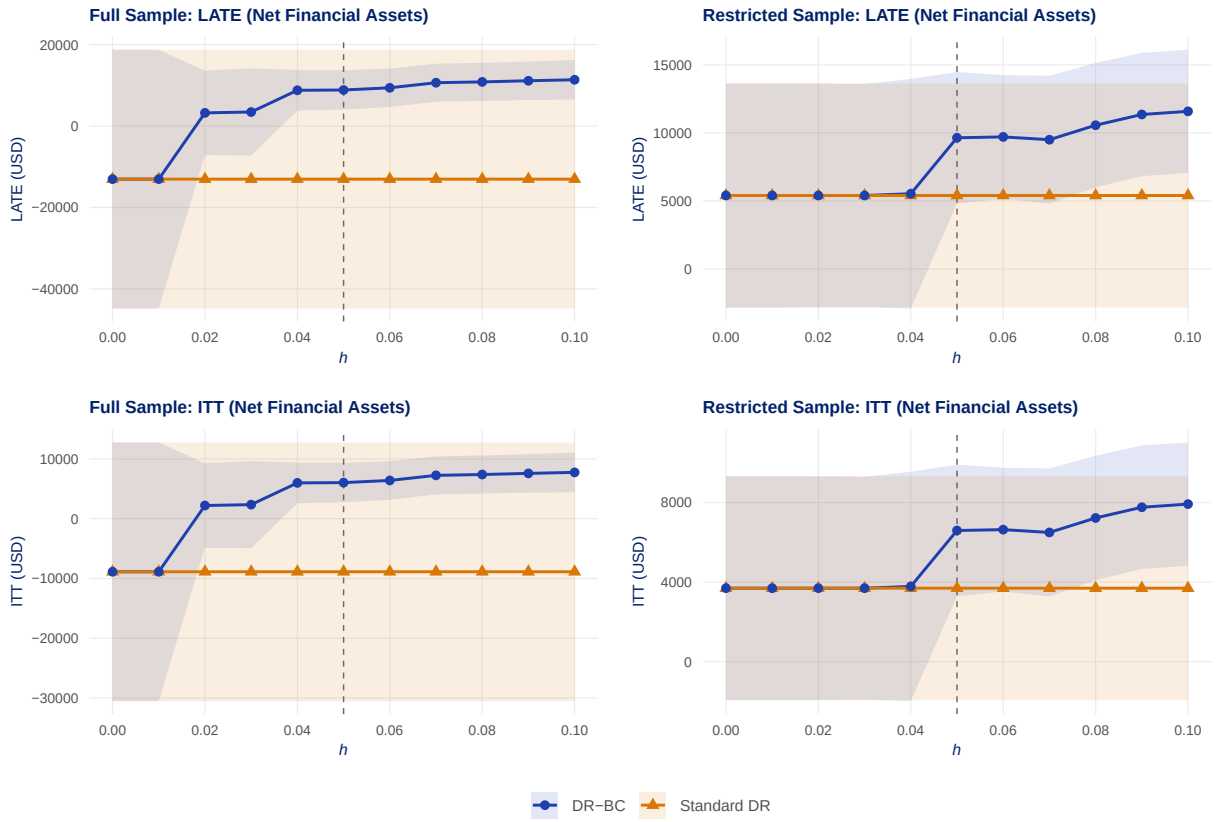
(iv) *The sieve approximation error $r_{K,\ell}^{(\kappa)}(0) = o(h^{1-\kappa}n^{-1/2})$, where*

$$r_{K,\ell}^{(\kappa)}(0) = \xi_\ell^{(\kappa)}(0;\gamma^*) - q_K^{(\kappa)}(0)'\,\mathbb{E}\left[q_K(A_\ell(\gamma^*))q_K(A_\ell(\gamma^*))'\right]^{-1}\mathbb{E}\left[q_K(A_\ell(\gamma^*))\xi_\ell(A_\ell(\gamma^*);\gamma^*)\right].$$

Then Assumption 4(b) holds for the index (ℓ, κ).

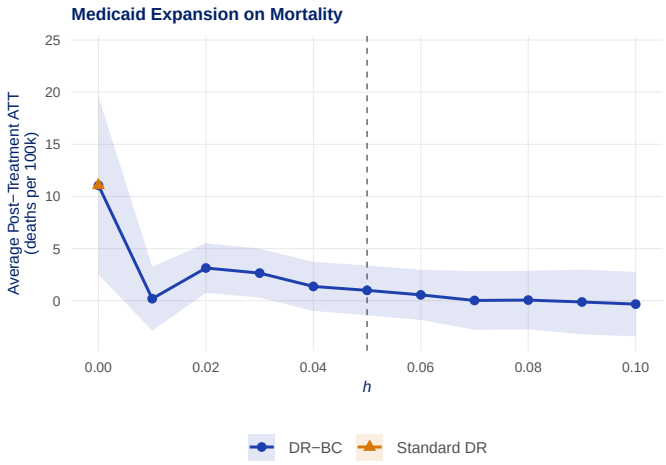
For parametric first-stage estimators $\hat{\gamma}$ (e.g., logistic propensity score, linear outcome regression), Assumption 4(a) and (d) hold under standard regularity conditions (see, e.g., Newey, 1997). Specifically, if $\hat{\gamma}$ is root-n consistent with influence function ϕ, and if $\alpha_\ell(h, \gamma)$ is differentiable in γ with bounded derivative, then (a) holds with $\phi_\ell = (\partial/\partial\gamma')\alpha_\ell(h, \gamma^*) \cdot \phi$, and (d) follows from a uniform law of large numbers.

Figure OA-7: Sensitivity to h: 401(k) Eligibility on Net Financial Assets



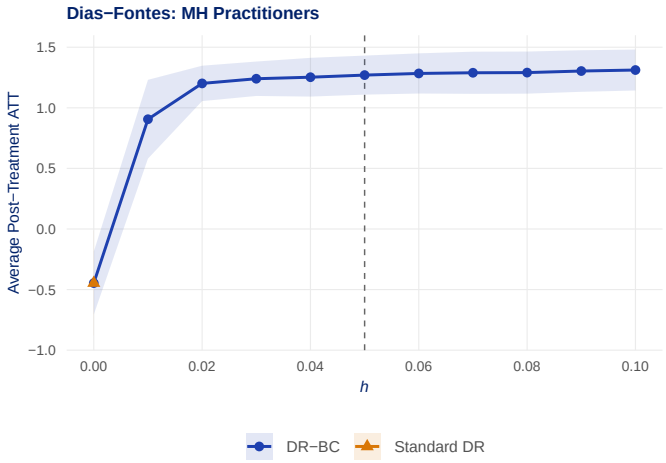
Notes. DR-BC (blue) and Standard DR (gold) estimates with 95% confidence intervals as a function of $h \in [0, 0.10]$ for the 401(k) application. Top row: LATE; bottom row: ITT. Left column: full sample (n = 9,910, positive income); right column: restricted sample (n = 9,275, income in [\$10k, \$200k]). The dashed vertical line marks $h = 0.05$. The Standard DR estimate is constant across h (it does not use trimming). DR-BC estimates stabilize for $h \geqslant 0.03$, with precision gains largest in the full sample where overlap is weakest.

Figure OA-8: Sensitivity to h: Medicaid Expansion on Mortality (Baker et al.)



Notes. Average post-treatment ATT for the crude mortality rate (ages 20–64, deaths per 100,000), computed as the mean of DR-BC event-study estimates across post-treatment event times. Never-treated counties serve as the comparison group, with population weights. The dashed vertical line marks $h = 0.05$. The Standard DR estimate (gold) is constant across h. DR-BC estimates are approximately stable across h, with tighter confidence intervals than Standard DR.

Figure OA-9: Sensitivity to h: Dias–Fontes DiD (Mental Health Practitioners, $\delta = 1$)



Notes. Average post-treatment ATT for mental health practitioners per 10,000 population with 1-year anticipation ($\delta = 1$), computed as the mean of DR-BC event-study estimates across post-treatment event times. The dashed vertical line marks $h = 0.05$. The Standard DR estimate (gold) is constant across h. DR-BC estimates are approximately stable for $h \geqslant 0.03$.

F Lower Bound on the Trimming Rate

Assumption 4(c) requires $\mathbb{E}\left[\omega_\ell(h,\gamma^*)^2\right] = o(n^{1/2})$. The following lemma gives a lower-level sufficient condition that yields the complementary rate condition $nh^4 \to \infty$.

Lemma F.1. *Let ℓ be any index with $1 \leqslant \ell \leqslant L$. Suppose $\mathbb{E}\left[B_\ell(\gamma^*)^2\right]$, $\xi_\ell^{(\kappa)}(0;\gamma^*)$ for each $\kappa = 1,\dots,k$, and $\mathbb{E}\left[\|\phi\|^2\right]$ are all bounded. If $nh^4 \to \infty$, $\|(\partial/\partial\gamma')\alpha_\ell(h,\gamma^*)\| = o(n^{1/4})$, and $\mathbb{E}\left[\psi_{\ell,\kappa}(\gamma^*)^2\right] = o(n^{1/2})$, then Assumption 4(c) holds.*

Proof. By the triangle inequality applied to the four-component decomposition (3.2):

$$\mathbb{E}\left[\omega_\ell(h,\gamma^*)^2\right]^{1/2} \leqslant h^{-1}\,\mathbb{E}\left[B_\ell(\gamma^*)^2\right]^{1/2} + \sum_{\kappa=1}^{k} \frac{h^{\kappa-1}}{\kappa!}\,|\xi_\ell^{(\kappa)}(0;\gamma^*)| + \sum_{\kappa=1}^{k} \frac{h^{\kappa-1}}{\kappa!}\,\mathbb{E}\left[\psi_{\ell,\kappa}(\gamma^*)^2\right]^{1/2} + \left\|\frac{\partial}{\partial\gamma'}\alpha_\ell(h,\gamma^*)\right\|\mathbb{E}\left[\|\phi\|^2\right]^{1/2}.$$

The first term is $O(h^{-1})$. Since $nh^4 \to \infty$ implies $h \to 0$ at most as fast as $n^{-1/4}$, we have $h^{-1} = o(n^{1/4})$. The remaining terms are $o(n^{1/4})$ by assumption. Hence $\mathbb{E}\left[\omega_\ell(h,\gamma^*)^2\right]^{1/2} = o(n^{1/4})$, which gives $\mathbb{E}\left[\omega_\ell(h,\gamma^*)^2\right] = o(n^{1/2})$. □

The condition $nh^4 \to \infty$ is a lower bound on h. Writing the bounds with their constants, the feasible range for $k = 1$ is $c_1\,n^{-1/4} < h < c_2\,n^{-1/2}$, where c_1 depends on the moment bounds of ω_ℓ and c_2 depends on the smoothness of ξ_ℓ. Since $n^{-1/4} \gg n^{-1/2}$ for large n, this range is asymptotically empty. However, for finite n the range may be non-empty if c_2/c_1 is sufficiently large: at $n = 10{,}000$, the range becomes $0.10\,c_1 < h < 0.01\,c_2$, which is non-empty whenever $c_2 > 10\,c_1$. The constants are DGP-dependent and not directly estimable with current methods, so a data-driven shrinking-h rule remains an open problem. For $k \geqslant 3$, the upper bound $h = O(n^{-1/(2k)})$ is less restrictive (e.g., $h = O(n^{-1/6})$ for $k = 3$), and the feasible range $c_1\,n^{-1/4} < h < c_2\,n^{-1/(2k)}$ is non-degenerate. In principle, this permits a vanishing-h analysis with $k \geqslant 3$, though in our simulations higher-order bias corrections degrade finite-sample performance because the required sieve derivatives $\xi_\ell^{(\kappa)}(0)$ for $\kappa \geqslant 3$ are imprecisely estimated with a fixed low-degree basis (Section 2.4).

G Proof of Theorem 3.1

We establish the four intermediate results:

$$\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = o(n^{-1/2}), \tag{G.1}$$

$$\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*) = (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\omega_\ell(h,\gamma^*)] + o_p(n^{-1/2}), \tag{G.2}$$

$$\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(0,\gamma^*) = o_p(n^{-1/4}), \tag{G.3}$$

$$\hat{\theta} - \theta_0 = (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\varphi] + o_p(n^{-1/2}). \tag{G.4}$$

Statement (i) of the Theorem 3.1 is (G.4). Statement (ii) follows immediately: by the Lyapunov CLT applied to $(\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\varphi]$ and the fact that $\mathbb{E}[\varphi^2]$ is bounded away from zero, combining with (G.4) gives $(\hat{\theta} - \theta_0)/\sqrt{\mathbb{E}[(\varphi - \mathbb{E}[\varphi])^2]/n} \xrightarrow{d} \mathcal{N}(0,1)$.

Step 1: Proof of (G.1).

By the definition of α_ℓ, the law of iterated expectations, and a kth-order Taylor expansion of $\xi_\ell(a;\gamma^*)$ around $a = 0$:

$$\begin{aligned}
&\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) \\
&= -\mathbb{E}\left[\frac{B_\ell(\gamma^*)}{A_\ell(\gamma^*)}\mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right] + \sum_{\kappa=1}^{k} \frac{\mathbb{E}\left[A_\ell(\gamma^*)^{\kappa-1}\mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right]}{\kappa!}\xi_\ell^{(\kappa)}(0;\gamma^*) \\
&= -\mathbb{E}\left[\frac{\xi_\ell(A_\ell(\gamma^*);\gamma^*)}{A_\ell(\gamma^*)}\mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right] + \sum_{\kappa=1}^{k} \frac{\mathbb{E}\left[A_\ell(\gamma^*)^{\kappa-1}\mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right]}{\kappa!}\xi_\ell^{(\kappa)}(0;\gamma^*) \\
&= -\frac{\mathbb{E}\left[A_\ell(\gamma^*)^k \int_0^1 (1-u)^k \xi_\ell^{(k+1)}(uA_\ell(\gamma^*);\gamma^*)\,du \cdot \mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right]}{k!},
\end{aligned}$$

where the second equality uses the law of iterated expectations $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot \mid A_\ell(\gamma^*)]]$, and the third uses Assumption 2. For residual moments (odd-indexed ℓ in the ATE and LATE designs, or the outcome-residual moments in DiD), $\xi_\ell(0;\gamma^*) = 0$ under either DR route, so the Taylor expansion around $a = 0$ eliminates all terms through order k and $\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = o(h^k)$. For normalization moments (even-indexed ℓ in ATE and LATE, or the comparison-group share in DiD), $\xi_\ell(0;\gamma^*) = 0$ holds under the propensity-score route, yielding the same $o(h^k)$ bound. Under the outcome regression route ($m^* = m$, $p^* \neq p$), however, $\xi_\ell(0;\gamma^*) \neq 0$, so $\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*)$ need not vanish. This does not compromise the estimand because the corresponding residual moments are identically zero along this route, making Λ insensitive to the normalization moments. It does, however, require the delta-method expansion in Step 4 to be centered at $\alpha(h,\gamma^*)$ rather than $\alpha(0,\gamma^*)$. Since ξ_ℓ is $(k+1)$-times differentiable near zero, the kth-order Taylor remainder with integral form is

$$\xi_\ell(a;\gamma^*) = \sum_{\kappa=1}^{k} \frac{a^\kappa}{\kappa!}\xi_\ell^{(\kappa)}(0;\gamma^*) + \frac{a^{k+1}}{k!}\int_0^1 (1-u)^k \xi_\ell^{(k+1)}(ua;\gamma^*)\,du.$$

Since $|A_\ell(\gamma^*)| \leqslant h$ on the indicator event and $\xi_\ell^{(k+1)}$ is bounded near zero,

$$\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = O\big(h^k\,\mathbb{E}[\mathbb{1}\{|A_\ell(\gamma^*)| < h\}]\big) = o(n^{-1/2}),$$

where the last step uses Assumption 4(e): $nh^{2k} = O(1)$ implies $h^k = O(n^{-1/2})$, and the

probability of the trimmed region goes to zero.

Step 2: Proof of (G.2).

By Assumption 4(d),

$$\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*) = \hat{\alpha}_\ell(h,\gamma^*) - \alpha_\ell(h,\gamma^*) + \alpha_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*) + o_p(n^{-1/2}).$$

By Assumptions 4(a) and (b):

$$\begin{aligned}\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*) &= \hat{\alpha}_\ell(h,\gamma^*) - \alpha_\ell(h,\gamma^*) + \frac{\partial}{\partial\gamma'}\alpha_\ell(h,\gamma^*)\cdot(\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\phi] + o_p(n^{-1/2}) \\ &= (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\omega_\ell(h,\gamma^*)] + o_p(n^{-1/2}),\end{aligned}$$

where the last equality collects the three stochastic terms from the trimmed ratio, bias correction, and sieve estimation parts of $\hat{\alpha}_\ell$ into $\omega_\ell(h,\gamma^*)$ as defined in (3.2).

Step 3: Proof of (G.3).

Combining (G.1) and (G.2):

$$\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(0,\gamma^*) = (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\omega_\ell(h,\gamma^*)] + o_p(n^{-1/2}).$$

By Assumption 4(c), $\mathbb{E}[\omega_\ell(h,\gamma^*)^2] = o(n^{1/2})$, so $(\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\omega_\ell(h,\gamma^*)] = o_p(n^{-1/4})$ by Markov's inequality. Hence $\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(0,\gamma^*) = o_p(n^{-1/4})$.

Step 4: Proof of (G.4).

By a first-order Taylor expansion of Λ around $(\alpha_1(h,\gamma^*),\ldots,\alpha_L(h,\gamma^*))$ under Assumption 3:

$$\begin{aligned}\hat{\theta} - \theta_h(\gamma^*) &= \Lambda(\hat{\alpha}_1(h,\hat{\gamma}),\ldots,\hat{\alpha}_L(h,\hat{\gamma})) - \Lambda(\alpha_1(h,\gamma^*),\ldots,\alpha_L(h,\gamma^*)) \\ &= \sum_{\ell=1}^{L}\Lambda_\ell(\alpha_1(h,\gamma^*),\ldots,\alpha_L(h,\gamma^*))\big(\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*)\big) + O_p\Big(\sum_{\ell=1}^{L}|\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*)|^2\Big).\end{aligned}$$

By (G.3), each $|\hat{\alpha}_\ell(h,\hat{\gamma}) - \alpha_\ell(h,\gamma^*)|^2 = o_p(n^{-1/2})$. By Assumption 1, $\theta_h(\gamma^*) = \theta_0 + o(h^k)$ and $nh^{2k} = O(1)$ gives the trimming bias as $o(n^{-1/2})$. Substituting (G.2) into the linear term yields

$$\begin{aligned}\hat{\theta} - \theta_0 &= \sum_{\ell=1}^{L}\Lambda_\ell(\alpha_1(h,\gamma^*),\ldots,\alpha_L(h,\gamma^*))\cdot(\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\omega_\ell(h,\gamma^*)] + o_p(n^{-1/2}) \\ &= (\mathbb{E}_n[\cdot] - \mathbb{E}[\cdot])[\varphi] + o_p(n^{-1/2}),\end{aligned}$$

where the last equality uses the definition of φ in (3.3). This completes the proof. □

H Proofs of Propositions 3.1–3.3

Each proposition simultaneously verifies Assumptions 1–3. This merges two steps that in earlier work (for a single-design estimator) appeared separately: (i) showing $\xi_\ell(0;\gamma^*) = 0$, which gives Assumption 2(i); and (ii) showing $\theta_h(\gamma^*) = \theta_0 + o(h^k)$ under either DR condition, which is Assumption 1. We establish both here for each design.

Proof of Proposition 3.1.

Part 1: Verifying Assumption 2.

For $\ell = 1$ ($A_1 = p_D^*(X)$, $B_1 = D(Y - m_Y^*(1,X))$), the law of iterated expectations gives

$$\xi_1(s;\gamma^*) = \mathbb{E}\left[D\big(Y - m_Y^*(1,X)\big) \mid p_D^*(X) = s\right].$$

Part (i) of the proposition requires $\xi_1(0;\gamma^*) = 0$ under either DR condition. When $p_D^* = p_D$: the boundary $p_D^*(X) = 0$ forces $D = 0$ a.s., so $B_1 = 0$ and $\xi_1(0;\gamma^*) = 0$. When $m_Y^* = m_Y$: the misspecification residual $m_Y(1,X) - m_Y^*(1,X) = 0$, so $\xi_1(0;\gamma^*) = s \cdot 0 = 0$. Condition (ii) implies $s \mapsto \mathbb{E}\left[p_D^*(X)(m_Y(1,X) - m_Y^*(1,X)) \mid p_D(X) = s\right]$ is $(k+1)$-times continuously differentiable near zero, which gives Assumption 2(ii) for $\ell = 1$. For $\ell = 2$ ($B_2 = D$, $A_2 = p_D^*(X)$), $\xi_2(0;\gamma^*) = \mathbb{E}\left[D \mid p_D^*(X) = 0\right] = 0$ when $p_D^* = p_D$ (since $D = 0$ a.s. at the boundary). When $m_Y^* = m_Y$, no separate verification for $\ell = 2$ is needed because the companion numerator moment $\ell = 1$ is already identically zero. The argument for $\ell = 3, 4$ is symmetric with $1 - p_D^*(X)$ replacing $p_D^*(X)$.

Part 2: Verifying Assumption 1.

From Step 1 of the proof of Theorem 3.1 (specifically (G.1)), we have for each ℓ:

$$\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = -\frac{\mathbb{E}\left[A_\ell(\gamma^*)^k \int_0^1 (1-u)^k \xi_\ell^{(k+1)}(uA_\ell(\gamma^*);\gamma^*)\,du \cdot \mathbb{1}\{|A_\ell(\gamma^*)| < h\}\right]}{k!}.$$

For the ATE with $\ell = 2$: when $p_D^* = p_D$, $\xi_2(s;\gamma^*) = \mathbb{E}\left[D \mid p_D(X) = s\right] = s$, which is exactly linear, so $\xi_2^{(\kappa)}(\cdot;\gamma^*) = 0$ for all $\kappa \geqslant 2$; the Taylor remainder therefore vanishes for any $k \geqslant 1$, and the expression above equals zero. When $m_Y^* = m_Y$, no separate Taylor argument is needed for $\ell = 2$, because the companion numerator moment $\ell = 1$ is already identically zero. For $\ell = 1$: when $p_D^* = p_D$, $\xi_1(s;\gamma^*) = s\,\mathbb{E}\left[m_Y(1,X) - m_Y^*(1,X) \mid p_D(X) = s\right]$ has a bounded $(k+1)$-th derivative, and the expression is $O(h^k \Pr(p_D(X) < h)) = o(h^k)$; when $m_Y^* = m_Y$, $\xi_1(\cdot;\gamma^*) = 0$ identically, so the remainder vanishes. Similarly for $\ell = 3, 4$. Under $p_D^* = p_D$: since $\Lambda(a_1,a_2,a_3,a_4) = \mathbb{E}[m_Y^*(1,X) - m_Y^*(0,X)] + a_1/a_2 - a_3/a_4$ is differentiable (Part 3) and each $\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = o(h^k)$, It follows that $\theta_h(\gamma^*) = \theta_0 + o(h^k)$. Under $m_Y^* = m_Y$: the residual moments $\ell = 1, 3$ are identically zero for all h, so the ATE representation is unchanged by trimming and $\theta_h(\gamma^*) = \theta_0$ exactly. Each route gives Assumption 1.

Part 3: Verifying Assumption 3. $\Lambda(a_1,a_2,a_3,a_4) = \mathbb{E}[m_Y^*(1,X) - m_Y^*(0,X)] + a_1/a_2 -$

a_3/a_4 is infinitely differentiable when the denominators $a_2 = \alpha_2(0, \gamma^*)$ and $a_4 = \alpha_4(0, \gamma^*)$ are bounded away from zero. Under either DR condition, these denominators are strictly positive: when $p_D^* = p_D$, $\alpha_2(0, \gamma^*) = \mathbb{E}[D/p_D(X)] = 1$; when $m_Y^* = m_Y$, $\alpha_2(0, \gamma^*) = \mathbb{E}[p_D(X)/p_D^*(X)] > 0$ since both propensity scores are strictly between 0 and 1. □

Proof of Proposition 3.2.

The LATE decomposition uses $L = 8$ moments: $\ell \in \{1, 2, 3, 4\}$ for the reduced-form Wald numerator and $\ell \in \{5, 6, 7, 8\}$ for the first-stage Wald denominator. In both blocks, $A_\ell \in \{p_Z^*(X), 1 - p_Z^*(X)\}$, mirroring the ATE structure with Z replacing D and $p_Z^*(X)$ replacing $p_D^*(X)$.

Part 1: Verifying Assumption 2.

For odd-indexed moments ($\ell = 1, 3$ in the reduced form, $\ell = 5, 7$ in the first stage), B_ℓ contains an outcome or treatment residual. Taking $\ell = 1$ ($B_1 = Z(Y - m_Y^{*\mathrm{LATE}}(1, X))$, $A_1 = p_Z^*(X)$) as representative:

$$\xi_1(s; \gamma^*) = \mathbb{E}\left[Z(m_Y^{\mathrm{LATE}}(1, X) - m_Y^{*\mathrm{LATE}}(1, X)) \mid p_Z^*(X) = s\right].$$

When $p_Z^* = p_Z$: the boundary $p_Z^*(X) = 0$ forces $Z = 0$ a.s., so $B_1 = 0$ and $\xi_1(0; \gamma^*) = 0$. When $m_Y^{*\mathrm{LATE}} = m_Y^{\mathrm{LATE}}$: the misspecification residual is zero, so $\xi_1(\cdot; \gamma^*) = 0$ identically. The argument for $\ell = 3, 5, 7$ is analogous (with $1 - p_Z^*(X)$ replacing $p_Z^*(X)$ for $\ell = 3, 7$, and D replacing Y for $\ell = 5, 7$). Condition (ii) of the proposition gives $(k + 1)$-fold differentiability for the odd-indexed moments.

For even-indexed moments ($\ell = 2, 4, 6, 8$), the identity $\xi_\ell(0; \gamma^*) = 0$ is invoked only along the correctly specified propensity score path. For example, if $\ell = 2$, ($B_2 = Z$, $A_2 = p_Z^*(X)$): $\xi_2(s; \gamma^*) = \mathbb{E}[Z \mid p_Z^*(X) = s] = s$, which is exactly linear, so $\xi_2(0; \gamma^*) = 0$ and $\xi_2^{(\kappa)}(0; \gamma^*) = 0$ for all $\kappa \geqslant 2$. Under the alternative double-robustness route, $m_R^{*\mathrm{LATE}} = m_R^{\mathrm{LATE}}$, the odd-indexed residual moments satisfy $\xi_\ell(a; \gamma^*) = 0$ for all a, so $\alpha_\ell(h, \gamma^*) = 0$ regardless of h. The even-indexed normalization moments need not equal zero when the propensity score is misspecified. The numerator moments of the Wald ratio are zero regardless of h, so the bias-correction terms vanish irrespective of the value of $\alpha_{\mathrm{even}}(h, \gamma^*)$. So no boundary expansion for the even-indexed moment is needed for the Wald-ratio limit. The argument for $\ell = 4, 6, 8$ is symmetric.

Part 2: Verifying Assumption 1. When $p_Z^* = p_Z$, the Taylor argument from Step 1 applies to both the reduced-form and first-stage moments, including their companion denominator moments, so $\theta_h^{\mathrm{num}}(\gamma^*) = \theta_0^{\mathrm{num}} + o(h^k)$ and $\theta_h^{\mathrm{den}}(\gamma^*) = \theta_0^{\mathrm{den}} + o(h^k)$. The Wald ratio $\theta_h(\gamma^*) = \theta_0 + o(h^k)$. When $m_Y^{*\mathrm{LATE}} = m_Y^{\mathrm{LATE}}$, the odd-indexed residual moments are identically zero. Hence no separate boundary expansion is needed for the even-indexed denominator moments, and the Wald ratio is unchanged by trimming. Therefore, the Wald

ratio $\theta_h(\gamma^*) = \theta_0$.

Part 3: Verifying Assumption 3. Λ = (numerator)/(denominator) is infinitely differentiable when the first-stage effect $\mathbb{E}[D(1) - D(0)] > 0$, which is condition (iii). □

Proof of Proposition 3.3.

For the DiD ATT(g,t), only $\ell = 2$ and $\ell = 4$ involve potentially weak overlap through $A_2 = A_4 = 1 - p^*_{g,t}(X)$; $\ell = 1, 3$ have $A_\ell = 1$, so $0 \notin \text{Supp}\,(A_\ell)$. These moments are unaffected by trimming.

Part 1: Verifying Assumption 2 for $\ell = 2$.

With $B_2 = p^*_{g,t}(X)C_{g,t}(Y_t - Y_{g-\delta-1} - m^*_{g,t}(X))$ and $A_2 = 1 - p^*_{g,t}(X)$, the law of iterated expectations gives

$$\xi_2(s;\gamma^*) = \mathbb{E}\left[p^*_{g,t}(X)C_{g,t}\big(Y_t - Y_{g-\delta-1} - m^*_{g,t}(X)\big) \mid 1 - p^*_{g,t}(X) = s\right].$$

Since $p^*_{g,t}(X) = 1 - s$ is fixed on the conditioning event, this equals

$$\xi_2(s;\gamma^*) = (1-s)\,\mathbb{E}\left[C_{g,t}\big(Y_t - Y_{g-\delta-1} - m^*_{g,t}(X)\big) \mid 1 - p^*_{g,t}(X) = s\right].$$

Condition (i) requires $\xi_2(0;\gamma^*) = 0$. When $p^*_{g,t} = p_{g,t}$: the boundary $p^*_{g,t}(X) = 1$ implies $C_{g,t} = 0$ a.s. (no comparison units at $p = 1$), so $B_2 = 0$ and $\xi_2(0;\gamma^*) = 0$. When $m^*_{g,t} = m_{g,t}$: the law of iterated expectations gives

$$\mathbb{E}\left[C_{g,t}\big(Y_t - Y_{g-\delta-1} - m_{g,t}(X)\big) \mid 1 - p^*_{g,t}(X) = s\right] = 0,$$

since $\mathbb{E}[Y_t - Y_{g-\delta-1} - m_{g,t}(X) \mid X, C_{g,t} = 1] = 0$ by definition of $m_{g,t}$. Condition (ii) gives $(k+1)$-fold differentiability. The argument for $\ell = 4$ is analogous when $p^*_{g,t} = p_{g,t}$; when $m^*_{g,t} = m_{g,t}$, no separate argument for $\ell = 4$ is needed because the companion residual moment $\alpha_2(h,\gamma^*)$ is already zero for every h.

Part 2: Verifying Assumption 1.

Under the correct propensity score route, the Taylor remainder argument of Step 1 applies to α_2 and α_4 separately, giving $\alpha_\ell(h,\gamma^*) - \alpha_\ell(0,\gamma^*) = o(h^k)$ for $\ell \in \{2,4\}$. Under the correct outcome regression route, $\alpha_2(h,\gamma^*) = 0$ for all h, so the ratio $\alpha_2/\alpha_4 = 0$ regardless of the value of $\alpha_4(h,\gamma^*)$, and Assumption 1 holds without requiring a Taylor argument for α_4.

Part 3: Verifying Assumption 3. Λ for ATT(g,t) is $\alpha_1/\alpha_3 - \alpha_2/\alpha_4$, differentiable when $\alpha_3 = \mathbb{E}[\mathbb{1}\{G = g\}] > 0$ (from $0 < \mathbb{E}[\mathbb{1}\{G = g\}] < 1$) and $\alpha_4 > 0$.

For the event-study aggregation $\text{ES}(e) = \sum_g w^{\text{es}}_{g,e} \cdot \text{ATT}(g, g+e)$, the result follows by a weighted-sum argument over each component ATT(g,t) for which the above holds. □

References

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato, "Some

new asymptotic theory for least squares series: pointwise and uniform results," *Journal of Econometrics*, 2015, *186* (2), 345–366.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Chen, Xiaohong and Timothy M Christensen, "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 2015, *188* (2), 447–465.

Dias, Mateus and Luiz Felipe Fontes, "The Effects of a Large-Scale Mental Health Reform: Evidence from Brazil," *American Economic Journal: Economic Policy*, 2024, *16* (3), 257–289.

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

Kang, Joseph D. Y. and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.," *Statistical Science*, 2007, *22* (4), 569–573.

Newey, Whitney K, "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 1997, *79* (1), 147–168.

Sant'Anna, Pedro HC and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.