econometrics

MDPI

Article

Joint Covariate Balancing for Difference-in-Differences with Missing Outcome Evolutions

Jun Luo [1] **and Junjie Li** [2,*]

[1] School of Business Administration and Customs Affairs, Shanghai Customs University, Shanghai 201204, China; luojun@shcc.edu.cn
[2] Department of Economics, Hitotsubashi University, Tokyo 186-8601, Japan
* Correspondence: junjieli.eco@gmail.com

Abstract

We propose a joint covariate balancing estimator for the average treatment effect on the treated (ATT) in a difference-in-differences (DID) design when the outcome evolution, rather than a specific outcome level, is missing at random (MAR). The proposed estimator combines treatment and response into one indicator and calibrates the two group-based density ratios directly. Point estimation therefore requires only two nuisance working models, and inference adds two linear projections, giving four in total, whereas a factorized implementation of the efficient influence function requires five: one treatment probability, two response probabilities, and two outcome-evolution models. The estimator is automatically self-normalized and attains the semiparametric efficiency bound for the coarsened data model when all working models are correctly specified. Furthermore, we show that the proposed estimator is doubly robust for both consistency and inference. The simulation results support componentwise robustness and confirm the efficiency result when all nuisance models are correctly specified, while highlighting substantial small-sample limitations under low response probabilities. An all-correct design further shows that the efficiency ordering relative to an estimator using the always-observed outcome level can reverse.

Keywords: covariate balancing; difference-in-differences; double robustness; missing at random; average treatment effect on the treated; semiparametric efficiency; density ratio

JEL Classification: C13; C14; C21; C23

1. Introduction

Semiparametric difference-in-differences (DID) estimators adjust for differences in pre-treatment covariates when parallel trends hold conditionally rather than unconditionally (Abadie, 2005; Sant'Anna & Zhao, 2020). In practice, however, the outcome is often missing for some units, and a simple complete-case analysis that discards all units with any missing outcome is inefficient, and is biased whenever response depends on covariates that also shift the outcome evolution. To address this issue, researchers have developed methods assuming that outcomes are missing at random (MAR) (Tang et al., 2024; Tsiatis, 2006; A. Zhao et al., 2024). In DID with MAR, Testa et al. (2026) propose a semiparametric DID estimator based on the efficient influence function (EIF) of the ATT to enhance efficiency. However, this construction introduces several nuisance models that must all be correctly specified to attain the semiparametric efficiency bound, which is challenging in practice. Reducing the number of nuisance models in the MAR DID setting is therefore desirable, though it is not obvious how to achieve this. The covariate balancing propensity score

Check for updates

Academic Editor: Vanessa Smith

Received: 19 July 2026
Revised: 7 October 2026
Accepted: 15 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

(CBPS) framework (Fan et al., 2023; Imai & Ratkovic, 2014; Li & Matsushita, 2025) provides this possibility, as it allows the composite of the treatment and response probabilities, namely, the group-based density ratio, to be calibrated directly in a single step.

In this paper, we propose a joint covariate balancing estimator for the ATT in a DID design with outcomes missing at random. The proof architecture for double robustness in DID follows Sant'Anna and Zhao (2020), and the calibration machinery follows the covariate balancing literature (Fan et al., 2023; Imai & Ratkovic, 2014) and its complete-data DID implementation (Li & Matsushita, 2025). Two elements are new. First, we coarsen the outcome to its evolution and combine treatment and response into a single indicator, so that the missingness restriction is placed on the outcome evolution rather than on a particular outcome level; the efficient influence function of the resulting observed-data model is derived in Proposition 1, and Remark 1 quantifies the efficiency consequence of the coarsening. Second, the calibration condition is joint across the two respondent groups, estimates the two group-based density ratios directly, and uses the treated sample as the common balancing target.

These two choices deliver the three properties that constitute our contribution. First, the proposed DID estimator reduces the number of nuisance models required, using only two working models for point estimation where a factorized implementation would require five, without losing local efficiency (i.e., it attains the semiparametric efficiency bound), and it is automatically self-normalized (i.e., it has the Hájek form). Inference additionally uses the two weighted least-squares projections, so that four working models enter in total, against five in the factorized implementation. Second, the proposed DID estimator is componentwise doubly robust for consistency: it is consistent if, for each respondent group, at least one of the density-ratio model and the outcome-evolution model is correctly specified. Third, the proposed DID estimator has a form-invariant asymptotic variance as long as, for each respondent group, at least one of the density-ratio model and the outcome-evolution model is correctly specified, a property known as double robustness for inference. Our numerical study examines small-sample inference, both favorable and unfavorable working-model configurations, and the efficiency comparison with an outcome-level estimator.

The estimator also sits within a broader literature on balancing and calibration weights. Inverse probability tilting (Graham et al., 2012) and entropy balancing (Hainmueller, 2012) construct weights by exponential tilting subject to moment constraints; Q. Zhao and Percival (2017) show that entropy balancing is doubly robust, and Chan et al. (2016) establish global efficiency for empirical balancing calibration weighting with a growing dictionary. Related constructions include stable balancing weights (Zubizarreta, 2015), kernel-based covariate functional balancing (Wong & Chan, 2018), and tailored loss functions (Q. Zhao, 2019). Our joint balancing condition inherits from this literature the exponential-tilting form of the weights, the exact finite-sample moment matching, and the convexity of the calibration problem. What is new is that the balancing target is the treated sample for both respondent groups simultaneously, and that each set of weights calibrates a composite of a treatment probability and a response probability rather than a single selection probability. On the missing-data side we rely on the coarsened data framework of Robins et al. (1994), Bang and Robins (2005) and Tsiatis (2006), and on the efficiency side on Hahn (1998) and Hirano et al. (2003). Finally, reducing the number of nuisance models is complementary to, rather than a substitute for, debiased machine learning (Chernozhukov et al., 2018).

The rest of the paper is organized as follows. Section 2 introduces the setup and assumptions. Section 3 derives the efficient influence function for the ATT and the associated componentwise doubly robust estimand. Section 4 presents the joint covariate balanc-

ing estimator. Section 5 establishes its local efficiency and its double robustness for both consistency and inference. Section 6 reports the Monte Carlo results. Section 7 concludes.

2. Setup

There are two periods, $t \in \{0,1\}$, and treatment starts only at $t = 1$. Let $D \in \{0,1\}$ indicate treatment, $Y_t(d)$ denote potential outcomes, and $\Delta Y = Y_1 - Y_0$. The indicator R equals one when both outcome levels needed for ΔY are observed. The coarsened data are i.i.d. copies of $O = (X, D, R, R\Delta Y)$, where X is a fully observed vector of pre-treatment covariates. The parameter of interest is the ATT, defined as

$$\tau = \mathbb{E}[Y_1(1) - Y_1(0) \mid D = 1]. \tag{1}$$

Define $p = \mathbb{P}(D = 1)$, $e^*(x) = \mathbb{P}(D = 1 \mid X = x)$, $r_d^*(x) = \mathbb{P}(R = 1 \mid D = d, X = x)$, $m_d^*(x) = \mathbb{E}(\Delta Y \mid D = d, X = x)$, $S_1 = DR$, and $S_0 = (1 - D)R$. Write $a_d^*(x) = \mathbb{P}(S_d = 1 \mid X = x)$, so that

$$a_1^*(x) = e^*(x)\, r_1^*(x), \qquad a_0^*(x) = \{1 - e^*(x)\}\, r_0^*(x). \tag{2}$$

The joint indicator is therefore multiplicative: each a_d^* is a product of two probabilities. Next, we introduce the identification conditions for the ATT and the MAR conditions for the outcome evolution.

Assumption 1 (Identification conditions)**.** *The following conditions hold:*

(i) $Y_1 = DY_1(1) + (1 - D)Y_1(0)$ *and* $Y_0 = Y_0(0)$ *almost surely;*
(ii) $\mathbb{E}[Y_1(0) - Y_0(0) \mid D = 1, X] = \mathbb{E}[Y_1(0) - Y_0(0) \mid D = 0, X]$ *almost surely;*
(iii) For some $\varepsilon > 0$, $p > \varepsilon$ *and* $e^*(X) \leq 1 - \varepsilon$ *almost surely.*

Assumption 1(i) imposes consistency and no anticipation: the observed post-treatment outcome agrees with the potential outcome under the realized treatment status, while the baseline outcome is untreated for every unit. Assumption 1(ii) is conditional parallel trends, which restricts the conditional mean of the untreated outcome evolution rather than the levels of the potential outcomes. Assumption 1(iii) is ATT overlap: treatment occurs with positive probability, and comparison units remain available over the covariate support relevant to the treated population. These conditions are standard in semiparametric DID analyses of the ATT (Abadie, 2005; Li & Matsushita, 2025; Sant'Anna & Zhao, 2020).

Assumption 2 (Missing at random and response overlap)**.** *For each* $d \in \{0,1\}$,

(i) $R \perp\!\!\!\perp \Delta Y \mid (D = d, X)$;
(ii) For some $\underline{r} > 0$, $r_d^*(X) \geq \underline{r}$ *almost surely on the covariate support of treated units.*

Assumption 2(i) is an MAR restriction on observability of the outcome evolution, not a pair of date-specific MAR restrictions as imposed in Testa et al. (2026). It allows $R = 0$ to arise from a missing baseline, a missing follow-up, or simultaneous missingness. If Y_0 is always observed, R is the post-treatment response indicator; if Y_1 is always observed, it is the pre-treatment response indicator. When either outcome may be missing, R is the complete-pair indicator, i.e., $R = R_0 R_1$, where $R_t = \mathbf{1}\{Y_t \text{ observed}\}$ for $t \in \{0,1\}$.

Example 1. *Assumption 2(i) restricts only outcome evolution, so it is useful to clarify the forms of dependence between R and the outcome levels* (Y_0, Y_1) *that it permits. For example, suppose that for some measurable h and some unobserved U,*

$$R = h(X, D, Y_0, U), \qquad U \perp\!\!\!\perp (Y_0, \Delta Y) \mid (D, X), \qquad Y_0 \perp\!\!\!\perp \Delta Y \mid (D, X). \tag{3}$$

Clearly, $(Y_0, U) \perp\!\!\!\perp \Delta Y \mid (D, X)$ *implies* $R \perp\!\!\!\perp \Delta Y \mid (D, X)$*, so Assumption 2(i) holds. However, since R is a nondegenerate function of the baseline outcome* Y_0 *and* $Y_1 = Y_0 + \Delta Y$*, neither of the date-specific conditions* $Y_0 \perp\!\!\!\perp R \mid (D, X)$ *and* $Y_1 \perp\!\!\!\perp R \mid (D, X)$ *need hold. An empirical example is missingness driven by the baseline outcome level: workers with low baseline earnings are more likely to be nonrespondents in a population where, conditional on age, education, and gender, baseline earnings carry no information about subsequent outcome evolution. In this setting, Assumption 2(i) accommodates mechanisms that are missing not at random (MNAR) with respect to the specific outcome levels, that is,* $Y_0 / \perp\!\!\!\perp R \mid (D, X)$ *and* $Y_1 / \perp\!\!\!\perp R \mid (D, X)$*.*

More importantly, under Assumption 2(i), one can use the observed data to identify the conditional mean of the outcome evolution for each treatment group, i.e., $\mathbb{E}(\Delta Y \mid D = d, R = 1, X) = \mathbb{E}(\Delta Y \mid D = d, X)$, $d \in \{0, 1\}$. Finally, under Assumptions 1–2, the ATT is identified by

$$\tau = \frac{1}{p}\mathbb{E}[\{S_1\omega_1^*(X) - S_0\omega_0^*(X)\}\Delta Y], \tag{4}$$

where

$$\omega_1^*(x) = \frac{e^*(x)}{a_1^*(x)} = \frac{1}{r_1^*(x)}, \qquad \omega_0^*(x) = \frac{e^*(x)}{a_0^*(x)} = \frac{e^*(x)}{\{1 - e^*(x)\}r_0^*(x)} \tag{5}$$

are the group-based density ratios. The role of Assumption 2(ii) is now transparent: by (2) the denominator of $\omega_d^* = e^*/a_d^*$ is a product of two probabilities, so overlap in treatment alone does not keep the weights bounded. Identification itself only requires $a_d^*(X) > 0$ on the covariate support of treated units, but Assumptions 1(iii) and 2(ii) are needed jointly to bound the weights. This identification result extends the inverse probability weighting (IPW) representation of Abadie (2005), in which only e^* enters, to outcome evolutions that are missing at random. When outcomes are fully observed, (4) reduces to that representation. The proof is given in Appendix A.1.

3. Efficient Influence Function and Componentwise Doubly Robust Estimation

Before introducing the proposed estimator, we characterize the efficiency benchmark under the outcome evolution MAR condition.

Proposition 1 (Efficient influence function)**.** *Under Assumptions 1–2, the efficient influence function for* τ *in the nonparametric model for* $O = (X, D, R, R\Delta Y)$ *is*

$$\begin{aligned}\varphi_{\text{eff}}(O) = \frac{1}{p}\Big[&S_1\omega_1^*(X)\{\Delta Y - m_1^*(X)\} - S_0\omega_0^*(X)\{\Delta Y - m_0^*(X)\} \\ &+ D\{m_1^*(X) - m_0^*(X) - \tau\}\Big],\end{aligned} \tag{6}$$

which satisfies $\mathbb{E}[\varphi_{\text{eff}}(O)] = 0$*, and* $V_{\text{eff}} = \mathbb{E}[\varphi_{\text{eff}}(O)^2]$ *is the semiparametric efficiency bound for regular estimators of* τ*. When the outcome evolution is fully observed, that is,* $r_1^*(X) = r_0^*(X) = 1$ *almost surely so that* $R = 1$ *for all units,* $\varphi_{\text{eff}}(O)$ *reduces to the efficient influence function for the ATT with fully observed panel data derived in Sant'Anna and Zhao (2020). The proof is given in Appendix A.2.*

Remark 1. *Proposition 2.8 of Testa et al. (2026) derives the efficient influence function for the ATT when the post-treatment outcome* Y_1 *is always observed and only the baseline outcome* Y_0 *may be missing. Their setting differs from ours in two respects. First, the MAR conditions differ. Their MAR condition is date-specific, requiring* $Y_0 \perp\!\!\!\perp R \mid (D, X)$ *(their Assumption 2.3), whereas Assumption 2(i) imposes a single restriction on the observability of the outcome evolution* ΔY *while allowing either outcome level, or both, to be missing. Hence, Assumption 2(i) is not nested within*

their date-specific restriction, nor is theirs nested within ours. Second, the observed data differ: in their setting, units with $R = 0$ still contribute their post-treatment outcomes, whereas under our coarsening $O = (X, D, R, R\Delta Y)$ they contribute only (X, D, R). To compare the efficiency bounds implied by the two MAR conditions, we work within the intersection of the two non-nested conditions by imposing the joint restriction

$$(Y_0, Y_1) \perp\!\!\!\perp R \mid (D, X), \tag{7}$$

which implies both their date-specific condition $Y_0 \perp\!\!\!\perp R \mid (D, X)$ and our Assumption 2(i). Specializing our setup to theirs, so that R indicates whether Y_0 is observed, let φ_T denote their efficient influence function and let $V_{\text{eff}}^T = \mathbb{E}(\varphi_T^2)$ denote the corresponding semiparametric efficiency bound for the observed data $O^T = (X, D, Y_1, R, RY_0)$. Under the intersection condition (7), their efficient influence function can then be written as $\varphi_T(O^T) = \varphi_{\text{eff}}(O) + \Delta(O, Y_1)$, where

$$\Delta(O, Y_1) = \frac{D}{p}\Big\{1 - \frac{R}{r_1^*(X)}\Big\}\{Y_1 - \nu_1^*(X)\} - \frac{(1-D)e^*(X)}{p\{1-e^*(X)\}}\Big\{1 - \frac{R}{r_0^*(X)}\Big\}\{Y_1 - \nu_0^*(X)\}, \tag{8}$$

is the adjustment that exploits the post-treatment outcomes of nonrespondents; here, $\nu_d^*(X) = \mathbb{E}(Y_1 \mid D = d, X)$, and $\Delta(O, Y_1)$ is not a function of O. Under the intersection condition (7), straightforward algebra yields the following expression for the efficiency comparison:

$$V_{\text{eff}}^T = V_{\text{eff}} - \frac{1}{p^2}\sum_{d=0}^{1}\mathbb{E}[q_d^*(X)\{\sigma_{\Delta,d}^{*2}(X) - \sigma_{0,d}^{*2}(X)\}], \tag{9}$$

where $\sigma_{\Delta,d}^{*2}(X) = \text{Var}(\Delta Y \mid D = d, X)$, $\sigma_{0,d}^{*2}(X) = \text{Var}(Y_0 \mid D = d, X)$, and the nonnegative weights are

$$q_1^*(X) = \frac{e^*(X)\{1 - r_1^*(X)\}}{r_1^*(X)}, \qquad q_0^*(X) = \frac{e^*(X)^2\{1 - r_0^*(X)\}}{\{1 - e^*(X)\}\, r_0^*(X)}.$$

The efficiency ranking can therefore go in either direction, depending on the serial dependence of the outcome. Writing $\sigma_{1,d}^{*2}(X) = \text{Var}(Y_1 \mid D = d, X)$ and $\sigma_{\Delta,d}^{*2} = \sigma_{0,d}^{*2} + \sigma_{1,d}^{*2} - 2\rho_d^*\sigma_{0,d}^*\sigma_{1,d}^*$ with ρ_d^* the conditional correlation between Y_0 and Y_1 given $(D = d, X)$, the coarsened bound V_{eff} is the smaller of the two whenever $\sigma_{\Delta,d}^{*2} < \sigma_{0,d}^{*2}$. For example, if $\sigma_{1,d}^* = \sigma_{0,d}^*$, then $\rho_d^* > 1/2$ implies $V_{\text{eff}} < V_{\text{eff}}^T$. In panel settings in which DID is typically applied, outcomes tend to exhibit strong serial correlation. Thus, coarsening the observed data to O and imposing Assumption 2(i) can be more, rather than less, informative than observing Y_1 for nonrespondents under the date-specific MAR condition. The ranking is reversed when the outcome evolution is more variable than the baseline outcome. The proof of (9) is provided in Appendix A.3.

In terms of the underlying conditional probabilities, $S_1\omega_1^*(X) = DR/r_1^*(X)$ and $S_0\omega_0^*(X) = (1-D)R\,e^*(X)/[\{1-e^*(X)\}r_0^*(X)]$, so (6) involves the three probability models e^*, r_1^*, and r_0^*, together with the two outcome-evolution models m_1^* and m_0^*. Motivated by Equation (6), let $\omega_d(X)$ and $m_d(X)$ denote working models for the group-based density ratio $\omega_d^*(X)$ and the outcome evolution $m_d^*(X)$, either of which may be misspecified, and define the augmented estimand

$$\tau_{\text{dr}} = \frac{1}{p}\,\mathbb{E}\Big[D\{m_1(X) - m_0(X)\} + S_1\omega_1(X)\{\Delta Y - m_1(X)\} - S_0\omega_0(X)\{\Delta Y - m_0(X)\}\Big]. \tag{10}$$

Proposition 2 (Componentwise double robustness)**.** *Under Assumptions 1–2, $\tau_{\text{dr}} = \tau$ if, for each $d \in \{0, 1\}$, either $\omega_d(X) = \omega_d^*(X)$ or $m_d(X) = m_d^*(X)$ almost surely.*

Proposition 2 shows that the augmented estimand equals the ATT as long as, for each respondent group, at least one of the two working models is correctly specified. Modeling the weights through the two density ratios rather than the three probabilities also weakens the requirements: ω_1 is correctly specified if and only if r_1 is, while ω_0 is correctly specified

whenever e and r_0 both are, but also under compensating misspecification of the two. Nevertheless, a factorized implementation of τ_{dr} would still fit the three probability models separately and combine them, so that five working models enter in total. The proof is given in Appendix A.4.

Recall that while the efficient influence function involves the outcome-evolution models, the identification result (4) involves only the weighting functions. It is therefore worth exploiting this lower-nuisance identification while retaining the double robustness property. The joint covariate balancing approach of the next section calibrates the two group-based density ratios in (5) directly and dispenses with the separate estimation of e^*, r_1^*, and r_0^*; in addition, no outcome-evolution model $m_d^*(X)$ needs to be estimated for point estimation. Furthermore, the resulting estimator retains componentwise double robustness.

4. Joint Covariate Balancing Estimator

Let $b(X) = (1, b_2(X), \ldots, b_K(X))^\top$ be a fixed dictionary. For each respondent group, consider a single-index working family generated by a scalar link function $g_d : \mathbb{R} \to (0, \infty)$ that is positive and twice continuously differentiable, with $\dot{g}_d(u) > 0$ for every $u \in \mathbb{R}$,

$$\omega_d(X; \lambda_d) = g_d\{b(X)^\top \lambda_d\}, \qquad \lambda_d \in \Lambda_d \subset \mathbb{R}^K, \quad d \in \{0, 1\},$$

these conditions on the link are also collected in Assumption 3(ii) below. To distinguish the three roles associated with ω_d, we use g_d for the scalar link function, $\omega_d(\cdot; \lambda_d)$ for the density-ratio working model induced by the link, and ω_d^* for the true group-specific density ratio. The derivative with respect to λ_d takes the convenient form $\partial g_d\{b(X)^\top \lambda_d\}/\partial \lambda_d = \dot{g}_d\{b(X)^\top \lambda_d\}\, b(X)$, where $\dot{g}_d(u) = dg_d(u)/du > 0$. The complete notation is summarized in Table 1 below.

Table 1. Notation for the working models.

Symbol	Meaning
$b(X)$	Fixed dictionary of K functions of the covariates, with $b_1 \equiv 1$
$g_d(\cdot), \dot{g}_d(\cdot)$	Scalar link function generating the working family for group d, and its derivative
$\omega_d^*(x) = e^*(x)/a_d^*(x)$	True group-based density ratio, defined in (5)
$\omega_d(x; \lambda_d) = g_d\{b(x)^\top \lambda_d\}$	Density-ratio working model at the parameter λ_d
$\omega_d^0 = \omega_d(\cdot; \lambda_d^0)$, $\widehat{\omega}_d = \omega_d(\cdot; \hat{\lambda}_d)$	Density-ratio working model at the pseudo-true and at the estimated parameter
$m_d^*(x) = \mathbb{E}(\Delta Y \mid D = d, X = x)$	True outcome evolution
$m_d(x)$	Generic outcome-evolution working model, used in (10) and Proposition 2
$\ell_d(x) = b(x)^\top \beta_d$	Linear outcome-evolution working model generated by the dictionary; ℓ_d^0 denotes it at the pseudo-true coefficient β_d^0 of (23)

For ω_d, one may, for example, use the exponential link $g_d(u) = \exp(u)$ to specify log-linear group-specific density ratios. Under this specification, correct specification requires $\log \omega_1^*(X) = -\log r_1^*(X) = b(X)^\top \lambda_1$ and $\log \omega_0^*(X) = \log\{e^*(X)/\{1 - e^*(X)\}\} - \log r_0^*(X) = b(X)^\top \lambda_0$ for some $\lambda_1, \lambda_0 \in \mathbb{R}^K$. Thus, the exponential specification is a direct working model for each composite density ratio; it is not generally implied by fitting separate logistic models for e and r_d. However, the comparison group specification can be restrictive because the true $\log \omega_0^*$ combines treatment odds and response selection; the robustness results below allow either working family to be misspecified.

4.1. Composite Versus Factorized Working Models

Correct specification of the composite model ω_0^* neither implies nor is implied by correct specification of the separate models for e^* and r_0^*. For example, consider the exponential link specification of ω_0^* mentioned above. In one direction, if e^* is logistic with index linear

in b and r_0^* is log-linear in b, then $\log \omega_0^* = \operatorname{logit} e^* - \log r_0^*$ is linear in b and therefore the composite model is correctly specified; if instead r_0^* is logistic, $\log \omega_0^*$ is not linear in b, so the composite model is misspecified under exponential link specification, even though both factors are correctly specified. In the other direction, the composite model is correct whenever the two factor misspecifications cancel in the difference $\operatorname{logit} e^* - \log r_0^*$; for example, if the two factors satisfy $\operatorname{logit} e^*(X) = b(X)^\top \alpha + h(X)$, $\log r_0^*(X) = b(X)^\top \gamma + h(X)$, for some $\alpha, \gamma \in \mathbb{R}^K$ and some bounded h, then neither the logistic model for e^* nor the log-linear model for r_0^* is correctly specified. However, $\log \omega_0^*(X) = b(X)^\top (\alpha - \gamma)$ is exactly linear in b and the composite model is correct. Thus, correctness of the composite can hold under compensating misspecification. The composite is the natural object when selection into treatment and nonresponse operate through the same latent index—for instance, when both are driven by a single measure of attachment to the program—since $\log \omega_0^*$ is then a single index even if neither factor belongs to a convenient parametric family; the factorized specification is preferable when assignment and response are governed by clearly distinct mechanisms with separately credible models. As a trade-off, the composite model is parsimonious and directly targeted, since ω_0^* is the only object that enters the identification result (4). However, it does not disentangle the two underlying mechanisms and therefore provides no mechanism-specific diagnostic, whereas the factorized model remains interpretable at the level of each mechanism.

4.2. Joint Covariate Balancing Condition

Let $\hat{\mathbb{E}}[h] = n^{-1} \sum_{i=1}^n h(O_i)$. We estimate $\lambda = (\lambda_1^\top, \lambda_0^\top)^\top$ from the joint covariate balancing condition

$$\hat{\mathbb{E}}\left[\begin{pmatrix} [S_1 g_1\{b(X)^\top \hat{\lambda}_1\} - D]b(X) \\ [S_0 g_0\{b(X)^\top \hat{\lambda}_0\} - D]b(X) \end{pmatrix}\right] = 0_{2K}. \tag{11}$$

This condition is valid because, at the true ratios,

$$\mathbb{E}[S_d \omega_d^*(X) b(X) \mid X] = a_d^*(X)\omega_d^*(X)b(X) = e^*(X)b(X) = \mathbb{E}[Db(X) \mid X].$$

Equation (11) extends the CBPS principle of Imai and Ratkovic (2014) and Fan et al. (2023) simultaneously to the two respondent groups. It targets balance rather than prediction of treatment or response, avoids separately fitting three probabilities, and exactly matches both groups, weighted by their group-based density ratios, to the treated sample. When $R = 1$ almost surely, the covariate balancing condition reduces to the complete-data CBPS DID condition of Li and Matsushita (2025). The link also governs whether the balancing condition can be solved in a given sample. Because g_d is strictly increasing, (11) is the first-order condition of the separable convex problem

$$\hat{\lambda}_d = \arg\min_{\lambda_d \in \Lambda_d} \hat{\mathbb{E}}[S_d G_d\{b(X)^\top \lambda_d\} - D\, b(X)^\top \lambda_d], \qquad G_d' = g_d, \tag{12}$$

whose Hessian matrix $\hat{\mathbb{E}}[S_d \dot{g}_d\{b(X)^\top \lambda_d\} b(X) b(X)^\top]$ is positive definite whenever $\hat{\mathbb{E}}[S_d b(X) b(X)^\top]$ is nonsingular, because $\dot{g}_d > 0$. Strict convexity delivers uniqueness, but not existence. We discuss this in Appendix A.6.

4.3. Estimation

Given a solution $\hat{\lambda}_d$ of the calibration problem (12), equivalently of the balancing condition (11), write $\widehat{\omega}_d(X) = g_d\{b(X)^\top \hat{\lambda}_d\}$ for the fitted group-based weights. The joint covariate balancing (JCB) DID estimator is

$$\widehat{\tau}_{\mathrm{JCB}} = \frac{1}{\hat{p}} \hat{\mathbb{E}}[\{S_1 \widehat{\omega}_1(X) - S_0 \widehat{\omega}_0(X)\}\Delta Y], \qquad \hat{p} = \hat{\mathbb{E}}[D]. \tag{13}$$

Balancing the intercept normalizes the group-based weights in each respondent group to the number treated, so that $\widehat{\tau}_{\mathrm{JCB}}$ is automatically self-normalized and takes the Hájek form (Hájek, 1971). Self-normalized weighting estimators are known to exhibit more stable finite-sample performance than their unnormalized counterparts (Busso et al., 2014). Moreover, the joint covariate balancing condition (11) implies, for any $\beta_1, \beta_0 \in \mathbb{R}^K$,

$$\widehat{\tau}_{\mathrm{JCB}} = \frac{1}{\hat{p}}\hat{\mathbb{E}}\Big[Db(X)^\top(\beta_1 - \beta_0) + S_1\widehat{\omega}_1(X)\{\Delta Y - b(X)^\top\beta_1\} - S_0\widehat{\omega}_0(X)\{\Delta Y - b(X)^\top\beta_0\}\Big], \tag{14}$$

which is similar in form to Equation (10) and hence doubly robust, as we will discuss in the next section. Unlike Equation (10), however, the JCB DID estimator does not require estimating the outcome-evolution models for point estimation.

For the choice of the dictionary $b(X)$, we recommend including an intercept and all covariates that enter the outcome-evolution model linearly, augmenting them with low-order transformations and interactions of variables that are likely to jointly affect treatment assignment and nonresponse, and keeping the dimension K fixed. For inference, we suggest choosing $\hat{\beta}_d$ to satisfy

$$\hat{\mathbb{E}}\big[S_d\,\dot{g}_d\{b(X)^\top\hat{\lambda}_d\}\,b(X)\{\Delta Y - b(X)^\top\hat{\beta}_d\}\big] = 0, \qquad d \in \{0,1\}. \tag{15}$$

Equation (15) is the normal equation of a weighted least-squares regression of ΔY on $b(X)$ with weights $S_d\dot{g}_d\{b(X)^\top\hat{\lambda}_d\}$ evaluated at the same $\hat{\lambda}_d$ that enters (13). Define the estimated influence function by

$$\widehat{\psi}_i = \frac{1}{\hat{p}}\Big[D_i\{b(X_i)^\top(\hat{\beta}_1 - \hat{\beta}_0) - \widehat{\tau}_{\mathrm{JCB}}\} + \sum_{d=0}^{1} s_d S_{di}\widehat{\omega}_d(X_i)\{\Delta Y_i - b(X_i)^\top\hat{\beta}_d\}\Big], \qquad s_1 = 1,\ s_0 = -1, \tag{16}$$

and estimate its variance by

$$\widehat{V} = \hat{\mathbb{E}}(\widehat{\psi}_i^2). \tag{17}$$

For $\alpha \in (0,1)$, an asymptotic level $1-\alpha$ confidence interval for τ is $\widehat{\tau}_{\mathrm{JCB}} \pm z_{1-\alpha/2}\sqrt{\widehat{V}/n}$. The same $\hat{\lambda}_d$ therefore enters (13), (15), and (16); Theorem 3 shows why this, and the choice of weights in (15), matters for the validity of the inference, and Section 6.1 gives the numerical details.

5. Theoretical Properties

This section establishes the asymptotic properties of the JCB DID estimator. We first state the regularity conditions and formalize the correctness conditions for the group-based density-ratio and outcome-evolution working models. We then show that the estimator is asymptotically normal and attains the semiparametric efficiency bound when all working models are correctly specified, and finally show that it remains consistent and that the variance estimator (17) delivers valid inference throughout the componentwise union model. Throughout this section, a superscript 0 denotes a pseudo-true parameter, defined as the solution to the corresponding population estimating equation, which is well defined even under misspecification; a superscript $*$ denotes the true quantity. Under correct specification the two coincide, for example, $\omega_d^0 = \omega_d^*$ and $\ell_d^0 = m_d^*$.

Assumption 3 (Regularity). *For $d \in \{0,1\}$,*

(i) The dictionary is bounded, $\|b(X)\| \le C_b < \infty$ almost surely; $\mathbb{E}(\Delta Y^4) < \infty$; and $\mathbb{E}[S_d b(X)b(X)^\top]$ is positive definite;

(ii) The link $g_d : \mathbb{R} \to (0,\infty)$ is twice continuously differentiable and satisfies $\dot{g}_d(u) > 0$ for every $u \in \mathbb{R}$;

(iii) The parameter space $\Lambda_d \subset \mathbb{R}^K$ *is compact and convex, and the population calibration equation* $\mathbb{E}[(S_d\, g_d\{b(X)^\top \lambda_d\} - D)\, b(X)] = 0$ *has a unique solution* λ_d^0*, which lies in the interior of* Λ_d*.*

Assumption 3 collects three primitive conditions. Part (i) restricts the dictionary: $b(X)$ is bounded, its second-moment matrix on the group-d respondents is nonsingular, and the outcome evolution has a finite fourth moment. Part (ii) requires a positive, twice continuously differentiable link with a strictly positive first derivative. Positivity makes $\omega_d(\cdot;\lambda_d)$ a candidate density ratio, and $\dot{g}_d > 0$ makes (11) the first-order condition of a strictly convex problem and $\dot{g}_d$ a positive weight in (15). The second derivative is needed because the projection estimating function already contains $\dot{g}_d$: differentiating it with respect to λ_d involves $\ddot{g}_d$. Together with the bounded dictionary and compact parameter space, continuity of $\ddot{g}_d$ gives the parameter-Lipschitz bounds used in Lemma A1. Part (iii) restricts the parameter space to be compact and convex and requires the population calibration equation to have a unique solution λ_d^0 in its interior; λ_d^0 is the pseudo-true calibration parameter, and its existence is an overlap statement, since it asks the treated distribution of $b(X)$ to be covered by that of the group-d respondents. Conditions of this kind are standard in the covariate balancing and calibration literature; see, for example, Chan et al. (2016), Q. Zhao (2019), and Imai and Ratkovic (2014).

Assumption 3 serves one purpose in what follows: it makes the two nuisance estimators $\hat{\lambda}_d$ and $\hat{\beta}_d$ root-n consistent and asymptotically linear, with a positive-definite Jacobian and square-integrable envelopes. This is exactly what the uniform laws of large numbers, mean-value expansions, and central limit theorems behind Theorems 1–3 require. The statement is given in Appendix A.5.

Write $\omega_d^0(X) = g_d\{b(X)^\top \lambda_d^0\}$, $\dot{g}_d^0(X) = \dot{g}_d\{b(X)^\top \lambda_d^0\}$, and $\ell_d^0 = b(X)^\top \beta_d^0$. For each respondent group define the group-based density-ratio and outcome-model correctness conditions:

$$\mathcal{W}_d : \ \omega_d^*(X) = g_d\{b(X)^\top \lambda_d^0\}, \qquad \mathcal{M}_d : \ m_d^*(X) = b(X)^\top \beta_d^0, \tag{18}$$

that is, the working family is correctly specified at the pseudo-true parameter. By the uniqueness of pseudo-true parameters, $\mathcal{W}_d$ (resp. $\mathcal{M}_d$) is equivalent to the existence of some $\lambda_d \in \mathbb{R}^K$ (resp. $\beta_d \in \mathbb{R}^K$) for which the working model equals the true function, since any such parameter solves the corresponding population equation and must therefore coincide with the pseudo-true value. In particular, $\omega_d^0 = \omega_d^*$ under $\mathcal{W}_d$ and $\ell_d^0 = m_d^*$ under $\mathcal{M}_d$.

5.1. Local Efficiency

We first consider the benchmark case in which all working models are correctly specified.

Theorem 1 (Local efficiency)**.** *Under Assumptions 1–3, if* $\mathcal{W}_d$ *and* $\mathcal{M}_d$ *hold for both* $d \in \{0,1\}$*, then*

$$\sqrt{n}(\widehat{\tau}_{\mathrm{JCB}} - \tau) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{eff}}), \tag{19}$$

where V_{eff} *is the semiparametric efficiency bound of Proposition 1.*

Theorem 1 shows that the JCB DID estimator is locally efficient: it attains the semiparametric efficiency bound without estimating the three probability models e^*, r_1^*, and r_0^* separately. The proof is given in Appendix A.7. Local efficiency is a property at the intersection of correctly specified working models; the next two subsections show that consistency and valid inference are preserved on the much larger componentwise union model.

5.2. Double Robustness for Consistency

The relevant union model is componentwise:

$$(\mathcal{W}_1 \text{ or } \mathcal{M}_1) \quad \text{and} \quad (\mathcal{W}_0 \text{ or } \mathcal{M}_0). \tag{20}$$

Theorem 2 (Double robustness for consistency)**.** *Under Assumptions 1–3, (20) implies* $\widehat{\tau}_{\mathrm{JCB}} \xrightarrow{p} \tau$.

Theorem 2 shows that the JCB DID estimator is doubly robust for consistency in a componentwise sense: for each respondent group d, it suffices that either the group-based density-ratio model or the outcome-evolution model is correctly specified, and the two groups may rely on different correctly specified components. The probability-limit bias of $\widehat{\tau}_{\mathrm{JCB}}$ is

$$\operatorname{plim} \widehat{\tau}_{\mathrm{JCB}} - \tau = \frac{1}{p}\Big[\mathbb{E}[(a_1^* \omega_1^0 - e^*)(m_1^* - \ell_1^0)] - \mathbb{E}[(a_0^* \omega_0^0 - e^*)(m_0^* - \ell_0^0)]\Big]. \tag{21}$$

For component d, $\mathcal{W}_d$ makes $a_d^* \omega_d^0 = e^*$, whereas $\mathcal{M}_d$ makes $m_d^* - \ell_d^0 = 0$. Thus, either model removes that component's bias, independently across the two respondent groups. It should be stressed that $\mathcal{M}_d$ requires the true outcome evolution m_d^* to be linear in the same dictionary $b(X)$ that is used for calibration. The proof is given in Appendix A.8.

5.3. Double Robustness for Inference

Consistency alone does not guarantee valid inference: the limiting distribution of $\sqrt{n}(\widehat{\tau}_{\mathrm{JCB}} - \tau)$ could in principle depend on the estimation error of $\hat{\lambda}_d$, and this contribution varies with which working models are correctly specified. This subsection shows that the weighted least-squares choice of $\hat{\beta}_d$ in (15) removes this first-order effect, so that the variance estimator (17) delivers valid inference throughout the union model (20).

Define

$$\psi_i^0 = \frac{1}{p}\Big[D_i\{b(X_i)^\top(\beta_1^0 - \beta_0^0) - \tau\} + \sum_{d=0}^{1} s_d S_{di} \omega_d^0(X_i)\{\Delta Y_i - b(X_i)^\top \beta_d^0\}\Big]. \tag{22}$$

The population projection equation selects

$$\beta_d^0 = \{\mathbb{E}[S_d \dot{g}_d^0(X) b(X) b(X)^\top]\}^{-1} \mathbb{E}[S_d \dot{g}_d^0(X) b(X) \Delta Y]. \tag{23}$$

Under (20), a first-order Taylor expansion of (14), evaluated at $\beta^0 = (\beta_1^{0\top}, \beta_0^{0\top})^\top$ and expanded in $\hat{\lambda}$ around $\lambda^0 = (\lambda_1^{0\top}, \lambda_0^{0\top})^\top$, gives

$$\begin{aligned} \sqrt{n}(\widehat{\tau}_{\mathrm{JCB}} - \tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^0 + \frac{1}{p} \sum_{d=0}^{1} s_d \kappa_d(\beta_d^0)^\top \sqrt{n}(\hat{\lambda}_d - \lambda_d^0) + o_p(1), \\ \kappa_d(\beta_d) &= \mathbb{E}\big[S_d \dot{g}_d^0(X)\, b(X)\{\Delta Y - b(X)^\top \beta_d\}\big]. \end{aligned} \tag{24}$$

Exact calibration makes the derivative with respect to β_d vanish. Moreover, $\kappa_d(\beta_d) = 0$ is precisely the population projection equation that defines β_d^0 in (23), so $\kappa_d(\beta_d^0) = 0$ by construction, even when the outcome model is misspecified; for any other choice of β_d, the corresponding term $\kappa_d(\beta_d)^\top \sqrt{n}(\hat{\lambda}_d - \lambda_d^0)$ would not vanish and the estimation error of $\hat{\lambda}_d$ would enter the limiting distribution. This choice removes the first-order effect of estimating λ_d; it does not remove the fixed misspecification bias in (21), which is why the union condition remains necessary. This projection argument parallels the complete-data constructions for doubly robust inference in Sant'Anna and Zhao (2020) and Li and Matsushita (2025).

Theorem 3 (Double robustness for inference). *Let ψ_i^0 be defined by (22) with β^0 given by the weighted least-squares projection (23). Under Assumptions 1–3 and (20),*

$$\sqrt{n}(\widehat{\tau}_{\mathrm{JCB}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^0 + o_p(1) \xrightarrow{d} \mathcal{N}(0, V^0), \qquad V^0 = \mathbb{E}[(\psi_i^0)^2], \tag{25}$$

and $\widehat{V} \xrightarrow{p} V^0$.

Theorem 3 shows that inference based on (17) is valid throughout the componentwise union model, with the same variance formula regardless of which working models are correctly specified. If both the group-based density-ratio and outcome-evolution models are correctly specified for both groups, ψ_i^0 equals the efficient influence function $\varphi_{\mathrm{eff}}(O_i)$ of Proposition 1 and $V^0 = V_{\mathrm{eff}}$, recovering the local efficiency stated in Theorem 1. The proof is given in Appendix A.9.

Remark 2. *It is worth being precise about what Theorem 3 delivers outside the intersection of correctly specified working models, since V^0 and V_{eff} are then different objects. Because $\widehat{V} \xrightarrow{p} V^0$ and V^0 is the actual asymptotic variance of $\widehat{\tau}_{\mathrm{JCB}}$ under whichever part of the union model (20) holds, the confidence intervals are asymptotically exact rather than conservative: their coverage converges to the nominal level throughout the union model. What is given up outside that intersection is efficiency, not validity, since V^0 need not equal the bound V_{eff} attained in Theorem 1. The consistency of $\widehat{V}$ in (25) is not automatic. In particular, the weights in (15) must be $S_d \dot{g}_d\{b(X)^\top \hat{\lambda}_d\}$, evaluated at the estimated calibration parameter $\hat{\lambda}_d$, and that same $\hat{\lambda}_d$ must be used in (13), (15), and (16).*

6. Monte Carlo Study

This section examines the finite-sample behavior of $\widehat{\tau}_{\mathrm{JCB}}$ along four dimensions: componentwise robustness and the accuracy of the associated inference throughout the union model (20); local efficiency at the intersection of correctly specified working models, together with a direct comparison with the estimator of Testa et al. (2026); what is given up when the factorized working models are correct and the composite ones are not; and the sensitivity of precision to the choice of dictionary. Appendix B.1 adds a weak-response design in which the calibration itself can fail.

6.1. Baseline Designs

We first consider four baseline designs with $n \in \{250, 500, 1000, 2000, 5000\}$, while the additional designs use selected sample sizes specified below. Within each design and sample size, all estimators are evaluated on the same simulated datasets. We fix the number of attempts per design and sample size, and record method-specific failures without replacing unsuccessful replications.

The raw covariates are independent standard normal variables, $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Define $Z_j = 2\Phi(X_j) - 1$, $j \in \{1,2\}$, where Φ is the standard normal distribution function. Thus, Z_1 and Z_2 are independent Uniform$[-1,1]$ variables, while the underlying covariates remain normally distributed. All estimators use the same working dictionary $b(X) = (1, Z_1, Z_2)^\top$. Let $Q = Z_1^2 - 1/3$, which is omitted from the dictionary. Treatment is generated from

$$\operatorname{logit} e^*(X) = -0.20 + 0.80\left(Z_2^2 - \frac{1}{3}\right).$$

The true group-based density ratios are

$$\log \omega_1^*(X) = \begin{cases} 1 + 0.75Z_1, & \mathcal{W}_1 \text{ correct}, \\ 1 + 0.75Z_1 + 0.50Z_1^2, & \mathcal{W}_1 \text{ incorrect}, \end{cases}$$

$$\log \omega_0^*(X) = \begin{cases} 1.30 - 0.65Z_1 + 0.10Z_2, & \mathcal{W}_0 \text{ correct}, \\ 1.30 - 0.65Z_1 + 0.10Z_2 + 0.60Z_1^2, & \mathcal{W}_0 \text{ incorrect}. \end{cases}$$

Thus correctness means that the log density ratio is linear in $b(X)$; misspecification is created by adding Z_1^2. The response probabilities are

$$r_1^*(X) = \exp\{-\log \omega_1^*(X)\}, \qquad r_0^*(X) = \exp\{\operatorname{logit} e^*(X) - \log \omega_0^*(X)\}.$$

The separate logistic working models for e^*, r_1^*, and r_0^* are therefore generally misspecified even when a composite density-ratio model $\mathcal{W}_d$ is correct. Let $g(X) = 1 + Z_1 - 0.70Z_2$. The true outcome-evolution models are

$$m_0^*(X) = \begin{cases} g(X), & \mathcal{M}_0 \text{ correct}, \\ g(X) - 4Q, & \mathcal{M}_0 \text{ incorrect}, \end{cases} \qquad m_1^*(X) = \begin{cases} g(X) + 2, & \mathcal{M}_1 \text{ correct}, \\ g(X) + 2 + 2.5Q, & \mathcal{M}_1 \text{ incorrect}. \end{cases}$$

Hence, a correct outcome model is linear in $b(X)$, whereas an incorrect model contains the omitted quadratic component Q. We generate $\Delta Y = Dm_1^*(X) + (1 - D)m_0^*(X) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 1)$, and draw $R \mid (D = d, X) \sim \text{Bernoulli}\{r_d^*(X)\}$ independently of ε. Consequently, $R \perp\!\!\!\perp \Delta Y \mid (D, X)$, so Assumption 2(i) holds by construction. Because Z_1 and Z_2 are bounded, the response probabilities are also bounded away from zero, so Assumption 2(ii) holds.

The four designs are DGP1: $\mathcal{W}_1$ and $\mathcal{W}_0$ correct, with $\mathcal{M}_1$ and $\mathcal{M}_0$ incorrect; DGP2: $\mathcal{W}_1$ and $\mathcal{M}_0$ correct, with $\mathcal{M}_1$ and $\mathcal{W}_0$ incorrect; DGP3: $\mathcal{M}_1$ and $\mathcal{W}_0$ correct, with $\mathcal{W}_1$ and $\mathcal{M}_0$ incorrect; and DGP4: $\mathcal{M}_1$ and $\mathcal{M}_0$ correct, with $\mathcal{W}_1$ and $\mathcal{W}_0$ incorrect. Each design satisfies (20). Since treatment depends on Z_2 but not on Z_1, $\mathbb{E}(Q \mid D = 1) = 0$. The true ATT is $\tau = 2$ in every design.

We consider three estimators: the naive estimator, the proposed JCB estimator, and the plug-in estimator. The naive estimator is the difference in mean outcome evolution between the two respondent groups. The plug-in estimator uses (10), with three logistic probability regressions and two ordinary least-squares outcome-evolution regressions on the same dictionary. The JCB estimator is defined in Section 4.3. For the plug-in estimator and, in DGP5 below, the Testa estimator, standard errors are computed from joint sandwich estimators that include all fitted nuisance scores and their cross-covariances; JCB uses the influence-function variance estimator in (17). The calibration problem (12) is solved by Newton–Raphson with step halving, started at $\lambda_d = 0$, which under the exponential link corresponds to the unweighted respondent sample, and stopped when the largest absolute element of the gradient falls below 10^{-8} within 200 iterations. The projection (15) is an ordinary weighted least-squares fit. All computations are carried out in R version 4.6.0; the code and the simulation results are provided in the Supplementary Materials.

Remark 3. *Equation (5) implies $\omega_1^*(x) = 1/r_1^*(x) \geq 1$, a restriction the exponential link does not impose, so $\widehat{\omega}_1(x) < 1$ can occur in finite samples. A reasonable alternative is the shifted link $g_1(u) = 1 + \exp(u)$, which enforces $\omega_1 > 1$ and has $\dot{g}_1(u) = \exp(u) > 0$, so it satisfies Assumption 3(ii) and every result above applies to it unchanged. We nevertheless work with the unconstrained exponential link, for two reasons. It makes the calibration problem an exponential tilting, and hence directly comparable to entropy balancing (Hainmueller, 2012; Q. Zhao & Percival, 2017); and $\omega_1^* \geq 1$ is a property of the truth rather than a requirement for any of Theorems 1–3, so imposing it is unnecessary for consistency, efficiency or valid inference. We therefore report the share of treated respondents with $\widehat{\omega}_1 < 1$ as a diagnostic, and the constrained link is the natural choice when that share is large.*

The results of the four baseline designs are summarized in Table 2 at $n = 5000$ and, across sample sizes, in Figure 1. Bias is the average deviation of the replicated estimates

from the true ATT; SD is the empirical standard deviation of the replicated estimates and SE is the average of their estimated standard errors; RMSE is the root mean squared error; and coverage is the proportion of nominal 95% Wald intervals containing the true ATT. Parentheses give Monte Carlo standard errors (MCSEs), computed as in Table 6 of Morris et al. (2019): SD/$\sqrt{m}$ for the bias and $\{\hat{c}(1-\hat{c})/m\}^{1/2}$ for the coverage $\hat{c}$, where m is the number of successful fits in the cell, equal to 2000 throughout Table 2.

Table 2. Baseline results.

Design	Estimator	Bias (MCSE)	SD	SE	RMSE	Cover. (MCSE)
DGP1	Naive	−0.3485 (0.0017)	0.076	0.076	0.357	0.0035 (0.0013)
DGP1	Plug-in	0.0751 (0.0019)	0.086	0.086	0.114	0.8580 (0.0078)
DGP1	JCB	−0.0041 (0.0018)	0.081	0.080	0.081	0.9455 (0.0051)
DGP2	Naive	−0.3867 (0.0016)	0.070	0.068	0.393	0.0000 (0.0000)
DGP2	Plug-in	0.0542 (0.0015)	0.069	0.067	0.088	0.8625 (0.0077)
DGP2	JCB	0.0016 (0.0015)	0.067	0.065	0.067	0.9380 (0.0054)
DGP3	Naive	−0.3775 (0.0017)	0.075	0.076	0.385	0.0005 (0.0005)
DGP3	Plug-in	0.0243 (0.0017)	0.078	0.080	0.082	0.9480 (0.0050)
DGP3	JCB	−0.0026 (0.0017)	0.074	0.076	0.074	0.9545 (0.0047)
DGP4	Naive	−0.4216 (0.0015)	0.066	0.067	0.427	0.0000 (0.0000)
DGP4	Plug-in	−0.0021 (0.0014)	0.061	0.060	0.061	0.9465 (0.0050)
DGP4	JCB	−0.0021 (0.0014)	0.061	0.060	0.061	0.9455 (0.0051)

Notes: Each cell uses 2000 attempted datasets. Parentheses give Monte Carlo standard errors (MCSEs) of bias and coverage. SD is the Monte Carlo standard deviation; SE is the mean estimated standard error. Intervals are normal 95% Wald intervals.

At $n = 5000$, JCB has absolute bias below 0.005 in each baseline design, whereas the plug-in biases in DGP1–DGP3 are 0.075, 0.054, and 0.024. Measured against the Monte Carlo standard errors in Table 2, these plug-in departures are 39, 35, and 14 MCSEs from zero, and the associated coverages of 0.8580 in DGP1 and 0.8625 in DGP2 lie about twelve MCSEs below nominal, so they reflect misspecification rather than simulation noise. The JCB departures are smaller by an order of magnitude: only DGP1 exceeds two MCSEs, at −0.0041 against an MCSE of 0.0018, and JCB coverage is within 2.3 MCSEs of 0.95 in every design. In DGP4, where both outcome models are correct, JCB and the plug-in estimator have essentially identical RMSE (0.061); their biases agree to four decimals and both lie 1.5 MCSEs from zero, so this design does not separate them. These comparisons illustrate correctness of the composite working models as a distinct source of robustness, rather than a general advantage over factorized models. The results above are all at $n = 5000$, whereas Theorems 2 and 3 are asymptotic and the samples available in applied DID studies are often far smaller. We therefore turn to the behavior of the estimator and of its inference as n falls.

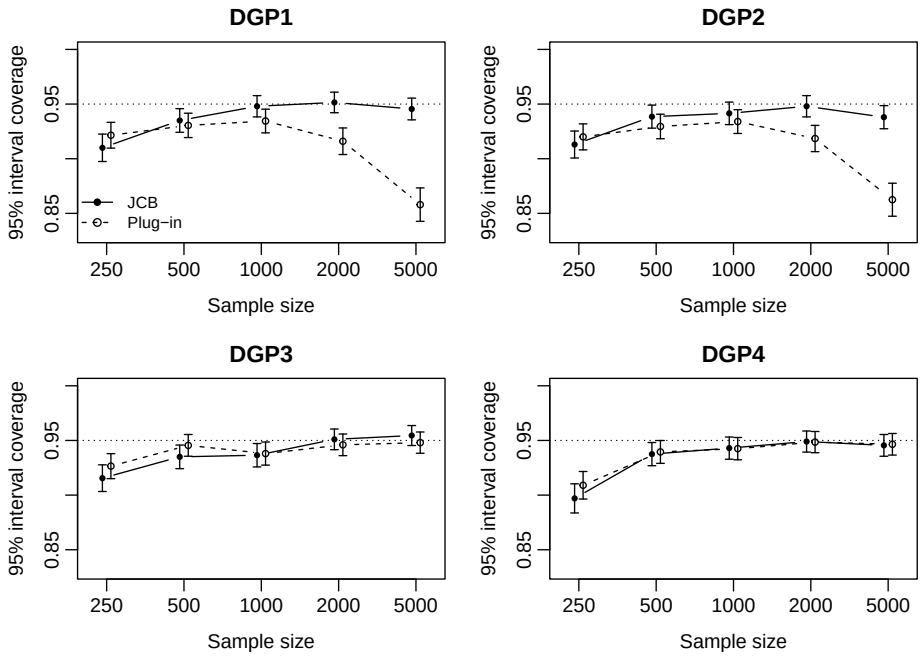


Figure 1. Coverage against sample size in DGP1–DGP4. Bars are ±1.96 Monte Carlo standard errors; the horizontal dotted line marks 0.95. Filled circles and solid lines denote JCB; open circles and dashed lines denote the plug-in estimator.

Figure 1 plots coverage against sample size for both estimators in the four baseline designs, with bars of ±1.96 Monte Carlo standard errors. Two features stand out. First, at $n = 250$ neither estimator attains nominal coverage: JCB ranges from 0.897 to 0.916, and its mean estimated standard error is only 0.87 to 0.90 times the Monte Carlo standard deviation, so the asymptotic approximation understates the sampling variability. The gap closes as n grows; by $n = 5000$, JCB coverage ranges from 0.938 to 0.955 and that ratio lies between 0.97 and 1.03. Second, the two estimators diverge in DGP1 and DGP2: the plug-in curve peaks near 0.934 at $n = 1000$ and then falls to 0.858 and 0.863 at $n = 5000$, while JCB stays close to nominal. The plug-in bias settles near 0.07 and 0.05 from $n = 500$ onwards, while its standard error shrinks at the rate $n^{-1/2}$, so the ratio of the two grows and the intervals miss increasingly often; this is the finite-sample signature of an inconsistent estimator. In DGP3 and DGP4 the two curves stay close, the smaller plug-in bias in DGP3 not yet being visible in coverage at these sample sizes. Detailed results are reported in Table A2 of Appendix B. We also examine the behavior of the JCB estimator when respondents are scarce, a setting in which its calibration can fail outright; that experiment is reported in Appendix B.1 and Table A4.

6.2. Efficiency Comparison

To illustrate the non-monotonic efficiency ranking discussed in Remark 1, we consider DGP5. Let $e^* = p = 0.4$ and $r_1^* = r_0^* = r = 0.6$. Define $\mu(X) = 0.5Z_1 + 0.5Z_2$ and generate $Y_0 = \mu(X) + \varepsilon_0$, $Y_1 = \mu(X) + g(X) + 2D + \varepsilon_1$, where $(\varepsilon_0, \varepsilon_1)$ is mean-zero bivariate normal with unit marginal variances and correlation $\rho \in \{0.2, 0.8\}$, independent of (X, D, R). We retain Y_1 for every unit and reveal Y_0 only when $R = 1$. Both MAR restrictions in Remark 1 hold, the ATT is 2, and all probability, density-ratio, and outcome working models on $b^{(3)} = (1, Z_1, Z_2)^\top$ are correctly specified. The sample sizes are 500, 1000, and 5000.

We directly implement the estimator under Assumption 2.3 of Testa et al. (2026), using low-dimensional logistic and linear regressions. This is a parametric implementation

of their estimating equation, not a replication of their random-forest procedure. Write $\mu_d^*(X) = \mathbb{E}(Y_0 \mid D = d, X)$, $\nu_0^*(X) = \mathbb{E}(Y_1 \mid D = 0, X)$ and $m_0^* = \nu_0^* - \mu_0^*$. The estimator is

$$\widehat{\tau}_T = \frac{1}{n_1} \sum_{i=1}^{n} \left[D_i \{ \widehat{U}_{1i} - \widehat{m}_0(X_i) \} - (1 - D_i) \frac{\widehat{e}(X_i)}{1 - \widehat{e}(X_i)} \{ \widehat{U}_{0i} - \widehat{m}_0(X_i) \} \right], \tag{26}$$

where $\widehat{m}_0(X) = \widehat{\nu}_0(X) - \widehat{\mu}_0(X)$, $n_1 = \sum_{i=1}^{n} D_i$ is the number of treated units and $\widehat{U}_{di} = Y_{1i} - \widehat{\mu}_d(X_i) - \frac{R_i}{\widehat{r}_d(X_i)} \{ Y_{0i} - \widehat{\mu}_d(X_i) \}$. Write $V_J = V_{\text{eff}}$ for the semiparametric efficiency bound of the coarsened model in Proposition 1 and $V_T = V_{\text{eff}}^T$ for the bound of the outcome-level model in Remark 1, both evaluated at the generating distribution; the same notation is used for the other designs in Appendix B. In the present design, Proposition 1 and the comparison identity (9) give

$$V_J = 2(1 - \rho) \left\{ \frac{1}{pr} + \frac{1}{(1 - p)r} \right\}, \quad V_T = \left(1 + \frac{1}{r} - 2\rho \right) \left\{ \frac{1}{p} + \frac{1}{1 - p} \right\}.$$

For $\rho = 0.2$, $(V_J, V_T) = (11.1111, 9.4444)$; for $\rho = 0.8$, they are $(2.7778, 4.4444)$. Since e^* and r_d^* are constant here and $m_1^* - m_0^* = \tau$, both bounds reduce to functions of ρ alone, and because $\sigma_{0,d}^{*2} = \sigma_{1,d}^{*2}$, Remark 1 places the crossing point at $\rho = 1/2$: the two values chosen are symmetric about it, so the ranking reverses, with $V_J > V_T$ at $\rho = 0.2$ and $V_J < V_T$ at $\rho = 0.8$.

Table 3 confirms this prediction in the simulated data. Both estimators are essentially unbiased, and $\overline{\widehat{V}}$ matches the corresponding bound. The reversal is therefore visible in the estimated variances and not only in the efficiency bounds: $\overline{\widehat{V}} = 11.1267$ for JCB against $\overline{\widehat{V}}_T = 9.4559$ for Testa at $\rho = 0.2$, and 2.7765 against 4.4475 at $\rho = 0.8$. Because all four working models are correctly specified in DGP5, this finding also directly verifies Theorem 1.

Table 3. Efficiency comparison under intersection MAR.

ρ	Estimator	Bias (MCSE)	SD	SE	V	$\overline{\widehat{V}}$	Cover. (MCSE)
0.2	JCB	−0.0012 (0.0011)	0.0471	0.0472	11.1111	11.1267	0.9470 (0.0050)
0.2	Testa	−0.0002 (0.0010)	0.0440	0.0435	9.4444	9.4559	0.9405 (0.0053)
0.8	JCB	0.0007 (0.0005)	0.0236	0.0236	2.7778	2.7765	0.9510 (0.0048)
0.8	Testa	0.0004 (0.0007)	0.0296	0.0298	4.4444	4.4475	0.9580 (0.0045)

Notes: V is the semiparametric efficiency bound of the model the estimator is built for, V_J for JCB and V_T for Testa, and $\overline{\widehat{V}}$ is the Monte Carlo average of $n\widehat{\text{SE}}^2$. The reported rows use $n = 5000$ and 2000 attempted datasets per cell. All other columns are defined as in Table 2. Both estimators are evaluated on the same simulated datasets.

6.3. Correct Factorized Models and Dictionary Sensitivity

Designs DGP6-N and DGP6-L reverse the roles of the two implementations: the probability models are genuinely logistic, so the factorized plug-in estimator is correctly specified, and we can quantify what JCB gives up in that case. They use

$$\begin{aligned} \operatorname{logit} e^* &= -1 + 1.5Z_1 + 0.3Z_2, \\ \operatorname{logit} r_1^* &= 0.3 - 0.7Z_1 + 0.2Z_2, \\ \operatorname{logit} r_0^* &= -0.2 + 0.6Z_1 - 0.2Z_2. \end{aligned}$$

These models belong to the factorized working families on $b^{(3)}$, whereas the composite density ratios generally do not belong to the exponential working family. DGP6-N uses $m_0^* = g - 4Q$ and $m_1^* = g + 2 + 2.5Q$, so both linear outcome models are misspecified. Numerical integration gives $\mathbb{E}(Q \mid D = 1) = 0.0242$ and $\tau = 2.1572$. DGP6-L uses $m_0^* = g$, $m_1^* = g + 2$ and $\tau = 2$. Both designs use $n \in \{500, 1000\}$ and unit-variance normal outcome-evolution noise.

Design DGP7 turns to the choice of dictionary. It retains the two correct composite ratios and two nonlinear outcome models from DGP1, but changes treatment assignment to

$$\operatorname{logit} e^* = -0.5 + 0.8(Z_2^2 - 1/3) + 0.6Q.$$

Now, $\mathbb{E}(Q \mid D = 1) = 0.0329$ and $\tau = 2 + 6.5\mathbb{E}(Q \mid D = 1) = 2.2137$; the omitted term therefore contributes to the ATT. We use $n \in \{500, 1000\}$ and compare JCB with $b^{(2)} = (1, Z_1)^\top$, $b^{(3)}$, and $b^{(4)} = (1, Z_1, Z_2, Q)^\top$ on the same data. The control-group robustness condition fails for $b^{(2)}$, both density ratios are correct for $b^{(3)}$, and all four JCB working models are correct for $b^{(4)}$. The results of DGP6-N, DGP6-L, and DGP7 are summarized in Table 4.

Table 4. Factorized correctness and dictionary sensitivity at $n = 1000$.

Design	Estimator	Bias (MCSE)	SD	SE	RMSE	Cover. (MCSE)
DGP6-N	Plug-in	−0.0014 (0.0036)	0.159	0.155	0.159	0.9425 (0.0052)
DGP6-N	JCB	−0.0249 (0.0034)	0.151	0.147	0.153	0.9350 (0.0055)
DGP6-L	Plug-in	0.0022 (0.0024)	0.109	0.107	0.109	0.9445 (0.0051)
DGP6-L	JCB	0.0023 (0.0024)	0.109	0.107	0.109	0.9425 (0.0052)
DGP7	JCB, $K = 2$	−0.0337 (0.0048)	0.214	0.206	0.217	0.9390 (0.0054)
DGP7	JCB	−0.0106 (0.0046)	0.207	0.198	0.207	0.9415 (0.0052)
DGP7	JCB, $K = 4$	−0.0022 (0.0040)	0.181	0.169	0.181	0.9315 (0.0056)

Notes: Each cell uses 2000 attempted datasets. Parentheses give Monte Carlo standard errors (MCSEs) of bias and coverage. SD is the Monte Carlo standard deviation; SE is the mean estimated standard error. Intervals are normal 95% Wald intervals. Unqualified estimators use $K = 3$. The plug-in estimator is shown only for DGP6, where the factorized models are correct; its DGP7 results are in Table A2. True ATTs: DGP6-N = 2.1572; DGP6-L = 2.0000; DGP7 = 2.2137.

At $n = 1000$ in DGP6-N, JCB bias is −0.0249 (MCSE: 0.0034), compared with −0.0014 (MCSE: 0.0036) for the plug-in estimator. JCB is outside its componentwise union model here. Its somewhat smaller RMSE, 0.153 versus 0.159, reflects a finite-sample variance trade-off and does not establish consistency. In DGP6-L both outcome models are correct, and the two estimators have almost identical bias and RMSE.

At $n = 1000$, enriching $b^{(3)}$ to $b^{(4)}$ reduces JCB bias from −0.0106 to −0.0022 and RMSE from 0.207 to 0.181; it does not improve coverage (0.9415 versus 0.9315). The $K = 4$ calibration fails once in 2000 attempts at $n = 500$. Dictionary enrichment can reduce approximation error and variance while making calibration more demanding; its effect on finite-sample coverage is not monotone.

7. Conclusions

This paper studies the estimation of the ATT in a DID design when the outcome evolution is missing at random. We derive the efficient influence function for the coarsened observed-data model and propose a joint covariate balancing estimator that combines treatment and response into a single indicator and calibrates the two group-based density ratios directly, thereby dispensing with the separate estimation of the treatment and response probabilities. The estimator requires only two working models for point estimation, is automatically self-normalized, and attains the semiparametric efficiency bound when all working models are correctly specified. Beyond local efficiency, it is componentwise doubly robust: for each respondent group, correct specification of either the density-ratio model or the outcome-evolution model suffices for consistency, and the weighted least-squares projection makes the same influence-function-based variance estimator valid throughout the union model, so that the resulting inference is doubly robust. The simulation results support the componentwise robustness of the proposed estimator and confirm the efficiency result when all nuisance models are correctly specified, while also highlighting substantial

small-sample limitations under low response probabilities. Designs with correctly specified factorized probability models further show that the JCB estimator does not enjoy a universal advantage under model misspecification.

Several extensions merit further study. First, the two-period design studied here is the building block of the group-time average treatment effects of Callaway and Sant'Anna (2021), surveyed together with the recent DID literature in Roth et al. (2023); the joint balancing approach could be extended to multiple periods and staggered adoption, where group-time ATTs would require a family of calibration conditions. Second, the fixed dictionary $b(X)$ could be allowed to grow with the sample size, connecting the joint balancing approach to sieve-based automatic estimation of the density ratios. Third, relaxing the MAR restriction on the outcome evolution toward instrument- or shadow-variable-based identification under missingness not at random is a natural next step.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/econometrics14040054/s1.

Author Contributions: Conceptualization, J.L. (Jun Luo); methodology, J.L. (Jun Luo) and J.L. (Junjie Li); software, J.L. (Junjie Li); writing—original draft preparation, J.L. (Junjie Li); writing—review and editing, J.L. (Jun Luo). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 2026 Annual Project of the National Social Science Fund of China (NSSFC).

Data Availability Statement: All numerical results are generated by simulation. The R code and the simulation results are provided in the Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proofs

Throughout, $\mathbb{E}$ denotes expectation under the true data-generating distribution of $O = (X, D, R, R\Delta Y)$, and we use the factorization $a_1^* = e^* r_1^*$, $a_0^* = (1 - e^*) r_0^*$ recorded in (2). Assumption 2(i) and iterated expectations imply, for any measurable g with the requisite moments and $d \in \{0, 1\}$,

$$\mathbb{E}[S_d\, g(X)] = \mathbb{E}[a_d^*(X) g(X)], \qquad \mathbb{E}[S_d\, g(X) \Delta Y] = \mathbb{E}[a_d^*(X) g(X) m_d^*(X)]. \tag{A1}$$

Appendix A.1. Proof of the Identification Result (4)

Proof. Applying the second identity in (A1) with $g = \omega_d^*$ and using $a_d^* \omega_d^* = e^*$ from (5),

$$\frac{1}{p}\mathbb{E}[\{S_1 \omega_1^*(X) - S_0 \omega_0^*(X)\}\Delta Y] = \frac{1}{p}\mathbb{E}[e^*(X) m_1^*(X)] - \frac{1}{p}\mathbb{E}[e^*(X) m_0^*(X)].$$

By Assumption 1(i), $m_1^*(X) = \mathbb{E}[Y_1(1) - Y_0(0) \mid D = 1, X]$, and by conditional parallel trends (Assumption 1(ii)), $m_0^*(X) = \mathbb{E}[Y_1(0) - Y_0(0) \mid D = 0, X] = \mathbb{E}[Y_1(0) - Y_0(0) \mid D = 1, X]$ on the treated support. Hence, the display equals $\mathbb{E}[\mathbb{E}[Y_1(1) - Y_1(0) \mid D = 1, X] \mid D = 1] = \tau$. □

Appendix A.2. Proof of Proposition 1

Proof. The proof follows the structure of Hahn (1998) and Sant'Anna and Zhao (2020). The observed-data density with respect to a dominating measure factorizes as

$$f(O) = f_X(x)\, e^*(x)^D \{1 - e^*(x)\}^{1-D}\, r_D^*(x)^R \{1 - r_D^*(x)\}^{1-R} f_1(\Delta y \mid x)^{DR} f_0(\Delta y \mid x)^{(1-D)R},$$

where $r_D^*(x) = D\,r_1^*(x) + (1-D)\,r_0^*(x)$ and $f_d(\Delta y \mid x)$ denotes the conditional density of ΔY given $(D=d, R=1, X=x)$; by Assumption 2(i), $m_d^*(x) = \int \Delta y f_d(\Delta y \mid x)\, d\Delta y$. For a regular parametric submodel through the truth, the score takes the form

$$\begin{aligned} S(O) = {} & u(X)\{D - e^*(X)\} + v(D,X)\{R - r_D^*(X)\} \\ & + DR\,s_1(\Delta Y \mid X) + (1-D)R\,s_0(\Delta Y \mid X) + t(X), \end{aligned}$$

with $\int s_d(\Delta y \mid x) f_d(\Delta y \mid x)\, d\Delta y = 0$ for all x and $\mathbb{E}[t(X)] = 0$.

We justify the claim that these elements generate the tangent space. The five factors f_X, e^*, r_D^*, f_1, and f_0 in the displayed factorization are variation-independent, and the observed-data model places no restriction on any of them: Assumption 2(i) is a restriction on the joint law of $(R, \Delta Y)$ given (D, X), which is not identified from O because ΔY is recorded only when $R = 1$; it therefore restricts the full-data law but leaves the model for O nonparametric. Consequently, the tangent space is the orthogonal direct sum

$$\mathcal{T} = \mathcal{T}_X \oplus \mathcal{T}_D \oplus \mathcal{T}_R \oplus \mathcal{T}_1 \oplus \mathcal{T}_0,$$

where $\mathcal{T}_X = \{t(X) : \mathbb{E}t = 0,\ \mathbb{E}t^2 < \infty\}$, $\mathcal{T}_D = \{u(X)\{D - e^*(X)\}\}$, $\mathcal{T}_R = \{v(D,X)\{R - r_D^*(X)\}\}$, and $\mathcal{T}_d = \{S_d\, s_d(\Delta Y \mid X) : \mathbb{E}[s_d \mid D = d, R = 1, X] = 0\}$, each intersected with L_2. Orthogonality of the five summands follows from iterated expectations, using $\mathbb{E}[D - e^* \mid X] = 0$, $\mathbb{E}[R - r_D^* \mid D, X] = 0$, and $\mathbb{E}[s_d \mid D = d, R = 1, X] = 0$. That each summand is generated by regular parametric submodels is standard: for a bounded h in the relevant subspace, the one-dimensional submodel obtained by tilting the corresponding factor by $1 + \theta h$ is regular with score h at $\theta = 0$, and bounded functions are dense in each summand, so the mean-square closure of the scores is the whole summand. This is precisely the coarsened data construction of Tsiatis (2006, Chapters 7–10), specialized to a two-stage coarsening by D and R; see also Hahn (1998) for the complete-data case. Because the observed-data model is nonparametric, $\mathcal{T}$ is the entire space of mean-zero square-integrable functions of O, so an influence function for τ is unique and is automatically efficient once it is shown to be an influence function; the verification below therefore establishes both claims at once. Along the submodel, (4) gives

$$\tau(\theta) = \frac{\int e_\theta(x)\{m_{1,\theta}(x) - m_{0,\theta}(x)\} f_{X,\theta}(x)\, dx}{\int e_\theta(x) f_{X,\theta}(x)\, dx}, \qquad m_{d,\theta}(x) = \int \Delta y\, f_{d,\theta}(\Delta y \mid x)\, d\Delta y,$$

whose pathwise derivative at the truth is

$$\begin{aligned} \partial_\theta \tau = \frac{1}{p}\Big[& \mathbb{E}\Big[e^*(X) \int \Delta y\, s_1(\Delta y \mid X) f_1(\Delta y \mid X)\, d\Delta y\Big] \\ & - \mathbb{E}\Big[e^*(X) \int \Delta y\, s_0(\Delta y \mid X) f_0(\Delta y \mid X)\, d\Delta y\Big] \\ & + \mathbb{E}\big[\{m_1^*(X) - m_0^*(X) - \tau\}\, \partial_\theta e(X)\big] \\ & + \mathbb{E}\big[\{m_1^*(X) - m_0^*(X) - \tau\} e^*(X)\, t(X)\big]\Big], \end{aligned} \tag{A2}$$

where $\partial_\theta e(X) = \partial e_\theta(X)/\partial\theta\big|_{\theta=\theta^*}$ denotes the pathwise derivative of the treatment probability along the submodel; matching the Bernoulli score component $u(X)\{D - e^*(X)\}$ gives $\partial_\theta e(X) = u(X) e^*(X)\{1 - e^*(X)\}$. Decompose

$$\varphi_{\text{eff}}(O) = \bar{u}(X)\{D - e^*(X)\} + DR\,\bar{s}_1(\Delta Y \mid X) + (1-D)R\,\bar{s}_0(\Delta Y \mid X) + \bar{t}(X),$$

with $\bar{u} = \{m_1^* - m_0^* - \tau\}/p$, $\bar{s}_1 = \omega_1^*(X)\{\Delta Y - m_1^*(X)\}/p$, $\bar{s}_0 = -\omega_0^*(X)\{\Delta Y - m_0^*(X)\}/p$, and $\bar{t} = e^*(X)\{m_1^* - m_0^* - \tau\}/p$. Each $\bar{s}_d$ integrates to zero against f_d, and $\mathbb{E}[\bar{t}(X)] = 0$ because $p\tau = \mathbb{E}[e^*(m_1^* - m_0^*)]$; hence, $\varphi_{\text{eff}} \in \mathcal{T}$, and $\mathbb{E}[\varphi_{\text{eff}}(O)] = 0$ follows from the same identities. Pairing φ_{eff} with each score component via iterated expectations, using

$\mathbb{E}[D\{D-e^*(X)\} \mid X] = e^*(X)\{1-e^*(X)\}, \mathbb{E}[R-r_D^*(X) \mid D, X] = 0, r_1^*\omega_1^* = 1, (1-e^*)r_0^*\omega_0^* = e^*$, and the zero conditional means of the residuals $\Delta Y - m_d^*(X)$ given $(D = d, R = 1, X)$, all cross-terms vanish and

$$\mathbb{E}[\varphi_{\text{eff}}(O)S(O)] = \partial_\theta \tau$$

for every regular submodel, with the four surviving pairings matching the four terms of (A2). Therefore, φ_{eff} is an influence function for τ; being an element of $\overline{\mathcal{T}}$, it is the efficient influence function, and $V_{\text{eff}} = \mathbb{E}[\varphi_{\text{eff}}^2]$ is the semiparametric efficiency bound. □

Appendix A.3. Proof of the Efficiency Comparison in Remark 1

Proof. Throughout this subsection we specialize to the design of Testa et al. (2026) under their Assumption 2.3: the post-treatment outcome Y_1 is always observed, R indicates that the baseline outcome Y_0 is observed, and the observed data are $O^T = (X, D, Y_1, R, RY_0)$, of which $O = (X, D, R, R\Delta Y)$ is a coarsening. We impose the intersection condition (7), that is, $(Y_0, Y_1) \perp\!\!\!\perp R \mid (D, X)$, under which both MAR conditions are satisfied. The two bounds are defined in their respective observed-data models: V_{eff} in the nonparametric model for O under Assumption 2(i), and V_{eff}^T in the nonparametric model for O^T under their Assumption 2.3. Write $\nu_d^*(X) = \mathbb{E}(Y_1 \mid D = d, X)$, $\mu_d^*(X) = \mathbb{E}(Y_0 \mid D = d, X)$, so that $m_d^* = \nu_d^* - \mu_d^*$, and set

$$\sigma_{1,d}^{*2}(X) = \text{Var}(Y_1 \mid D{=}d, X),\ \sigma_{0,d}^{*2}(X) = \text{Var}(Y_0 \mid D{=}d, X),\ \sigma_{\Delta,d}^{*2}(X) = \text{Var}(\Delta Y \mid D{=}d, X).$$

Step 1 (Comparison of the two efficient influence functions). Based on Equation (5) in Proposition 2.8 of Testa et al. (2026), the efficient influence function for τ in the nonparametric model for O^T is given by

$$\varphi_T = \frac{1}{p}\Big[D\Big\{(Y_1 - \nu_1^*) - \frac{R}{r_1^*}(Y_0 - \mu_1^*)\Big\} - \frac{(1-D)e^*}{1-e^*}\Big\{(Y_1 - \nu_0^*) - \frac{R}{r_0^*}(Y_0 - \mu_0^*)\Big\} + D\{m_1^* - m_0^* - \tau\}\Big],$$

where the dependence on X is suppressed for notational simplicity. Substituting $\Delta Y - m_d^* = (Y_1 - \nu_d^*) - (Y_0 - \mu_d^*)$ into (6), we obtain $\varphi_T - \varphi_{\text{eff}} = \Delta(O, Y_1)$, where Δ is defined in (8). Under (7), an application of the law of iterated expectations yields $\mathbb{E}[\Delta(O, Y_1)] = 0$.

Step 2 (Second and cross-moments). Because $r_d^*(X) = \mathbb{P}(R = 1 \mid D = d, X)$ and $R^2 = R$, for $d \in \{0, 1\}$,

$$\mathbb{E}\Big[1 - \frac{R}{r_d^*} \,\Big|\, D = d, X\Big] = 0, \qquad \mathbb{E}\Big[\Big(1 - \frac{R}{r_d^*}\Big)^2 \,\Big|\, D = d, X\Big] = \frac{1 - r_d^*}{r_d^*},$$

$$\mathbb{E}\Big[\frac{R}{r_d^*}\Big(1 - \frac{R}{r_d^*}\Big) \,\Big|\, D = d, X\Big] = -\frac{1 - r_d^*}{r_d^*},$$

and, by (7), each of them factorizes from any function of (Y_0, Y_1) given (D, X). The two terms of Δ are supported on $D = 1$ and $D = 0$, respectively, so their cross-product vanishes because $D(1 - D) = 0$. Applying the law of iterated expectations conditional on (D, X) together with the second identity above, the second moment of Δ is

$$p^2\mathbb{E}(\Delta^2) = \mathbb{E}\Big[e^*\frac{1 - r_1^*}{r_1^*}\sigma_{1,1}^{*2}\Big] + \mathbb{E}\Big[(1 - e^*)\Big(\frac{e^*}{1 - e^*}\Big)^2\frac{1 - r_0^*}{r_0^*}\sigma_{1,0}^{*2}\Big] = \sum_{d=0}^{1}\mathbb{E}[q_d^*\,\sigma_{1,d}^{*2}],$$

where $\mathbb{E}[(Y_1 - \nu_d^*)^2 \mid D = d, X] = \sigma_{1,d}^{*2}$, $q_1^* = \frac{e^*\{1-r_1^*\}}{r_1^*}$, and $q_0^* = \frac{(e^*)^2\{1-r_0^*\}}{\{1-e^*\}\,r_0^*}$. By a similar argument, the cross-moment of φ_{eff} and Δ can be evaluated component by component. The component $D\{m_1^* - m_0^* - \tau\}$ of φ_{eff} is a function of (D, X); it is orthogonal to the first term of Δ by the first identity above, and to the second term because $D(1 - D) = 0$. For the

remaining components, $S_1\omega_1^* = DR/r_1^*$ and $S_0\omega_0^* = (1-D)Re^*/\{(1-e^*)r_0^*\}$, so the third identity gives

$$p^2\,\mathbb{E}(\varphi_{\text{eff}}\Delta) = -\mathbb{E}\Big[e^*\frac{1-r_1^*}{r_1^*}c_1^*\Big] - \mathbb{E}\Big[\frac{(e^*)^2}{1-e^*}\frac{1-r_0^*}{r_0^*}c_0^*\Big] = -\sum_{d=0}^{1}\mathbb{E}[q_d^*\,c_d^*],$$

where $c_d^* = \mathrm{Cov}(\Delta Y, Y_1 \mid D=d, X)$.

Step 3 (Conclusion). Since $\varphi_T = \varphi_{\text{eff}} + \Delta$ and both terms have mean zero, $V_{\text{eff}}^T = \mathbb{E}(\varphi_T^2) = V_{\text{eff}} + 2\mathbb{E}(\varphi_{\text{eff}}\Delta) + \mathbb{E}(\Delta^2)$. Combining the two displays in Step 2,

$$p^2\big(V_{\text{eff}}^T - V_{\text{eff}}\big) = p^2\{2\mathbb{E}(\varphi_{\text{eff}}\Delta) + \mathbb{E}(\Delta^2)\} = \sum_{d=0}^{1}\mathbb{E}[q_d^*\big(\sigma_{1,d}^{*2} - 2c_d^*\big)].$$

Since $c_d^* = \mathrm{Cov}(Y_1 - Y_0, Y_1 \mid D=d, X) = \sigma_{1,d}^{*2} - \mathrm{Cov}(Y_0, Y_1 \mid D=d, X)$,

$$\sigma_{1,d}^{*2} - 2c_d^* = 2\,\mathrm{Cov}(Y_0, Y_1 \mid D=d, X) - \sigma_{1,d}^{*2} = -\{\sigma_{\Delta,d}^{*2} - \sigma_{0,d}^{*2}\},$$

which is (9). □

Appendix A.4. Proof of Proposition 2

Proof. Applying the identities in (A1) to each term of (10),

$$p\,\tau_{\text{dr}} = \mathbb{E}[e^*(m_1 - m_0)] + \mathbb{E}[a_1^*\omega_1(m_1^* - m_1)] - \mathbb{E}[a_0^*\omega_0(m_0^* - m_0)],$$

where all functions are evaluated at X. Subtracting $p\tau = \mathbb{E}[e^*(m_1^* - m_0^*)]$ and rearranging,

$$p\,(\tau_{\text{dr}} - \tau) = \mathbb{E}[(a_1^*\omega_1 - e^*)(m_1^* - m_1)] - \mathbb{E}[(a_0^*\omega_0 - e^*)(m_0^* - m_0)]. \tag{A3}$$

Each term in (A3) is the product of a weighting error and an outcome error for one respondent group. If $\omega_d = \omega_d^*$, then $a_d^*\omega_d = e^*$ by (5); if $m_d = m_d^*$, then $m_d^* - m_d = 0$. Hence, under the componentwise condition of Proposition 2, both terms vanish and $\tau_{\text{dr}} = \tau$. □

Appendix A.5. Auxiliary Lemmas

Lemma A1 (Primitive conditions imply the asymptotic regularity used in Section 5)**.** *Let Assumptions 1–3 hold and let $d \in \{0,1\}$. Then,*

(a) The population calibration objective $\lambda_d \mapsto \mathbb{E}[S_d G_d\{b(X)^\top\lambda_d\} - D\,b(X)^\top\lambda_d]$, with $G_d' = g_d$, is strictly convex on Λ_d, so the interior solution λ_d^0 of Assumption 3(iii) is its global minimizer, and the Jacobian $J_d = \mathbb{E}[S_d\dot{g}_d\{b(X)^\top\lambda_d^0\}b(X)b(X)^\top]$ is finite and positive definite;

(b) The population projection equation has the unique solution β_d^0 of (23), whose design matrix is exactly the calibration Jacobian J_d, and the projection is a weighted least-squares fit with weights $S_d\dot{g}_d^0(X)$;

(c) With probability approaching one, the sample problem (12) has a unique minimizer $\hat{\lambda}_d$ in the interior of Λ_d, and $\hat{\lambda}_d \xrightarrow{p} \lambda_d^0$ and $\hat{\beta}_d \xrightarrow{p} \beta_d^0$;

(d) $(\hat{\lambda}_d, \hat{\beta}_d)$ is root-n consistent and asymptotically linear, with

$$\sqrt{n}(\hat{\lambda}_d - \lambda_d^0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J_d^{-1}\{D_i - S_{di}\omega_d^0(X_i)\}b(X_i) + o_p(1),$$

and

$$\sqrt{n}(\hat{\beta}_d - \beta_d^0) = J_d^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[\zeta_d^0(O_i) + H_d J_d^{-1}\{D_i - S_{di}\omega_d^0(X_i)\}b(X_i)\big] + o_p(1),$$

where

$$\zeta_d^0(O) = S_d\dot{g}_d^0(X)b(X)\{\Delta Y - b(X)^\top\beta_d^0\},$$
$$H_d = \mathbb{E}\big[S_d\ddot{g}_d\{b(X)^\top\lambda_d^0\}b(X)b(X)^\top\{\Delta Y - b(X)^\top\beta_d^0\}\big];$$

(e) *The calibration and projection estimating functions, indexed by* (λ_d, β_d) *in a neighborhood of* (λ_d^0, β_d^0)*, are continuously differentiable in the parameters; the functions and their first derivatives have square-integrable envelopes, and the functions satisfy a parameter-Lipschitz bound with a square-integrable random coefficient.*

Lemma A2 (Population calibration and projection identities)**.** *Let Assumptions 1–3 hold. For* $d \in \{0, 1\}$*,*

(a) $\mathbb{E}[\{a_d^*(X)\omega_d^0(X) - e^*(X)\}b(X)] = 0$*;*
(b) *If* $\omega_d^*(X) = g_d\{b(X)^\top \bar{\lambda}_d\}$ *for some* $\bar{\lambda}_d \in \mathbb{R}^K$*, then* $\bar{\lambda}_d = \lambda_d^0$*, so that* $\mathcal{W}_d$ *holds and* $a_d^*\omega_d^0 = e^*$ *almost surely;*
(c) *If* $m_d^*(X) = b(X)^\top \bar{\beta}_d$ *for some* $\bar{\beta}_d \in \mathbb{R}^K$*, then* $\bar{\beta}_d = \beta_d^0$*, so that* $\mathcal{M}_d$ *holds and* $\ell_d^0 = m_d^*$ *almost surely.*

The two lemmas play different roles. Lemma A1 converts Assumption 3 into the asymptotic regularity that the later proofs use: a strictly convex population calibration objective with a finite, positive-definite Jacobian, a unique population projection coefficient whose design matrix is that same Jacobian, consistency of $(\hat{\lambda}_d, \hat{\beta}_d)$, their root-$n$ asymptotic linearity, and square-integrable envelopes for the estimating functions. Parts (c)–(e) are what justify the uniform laws of large numbers, mean-value expansions and central limit theorems invoked in the proofs of Theorems 2 and 3.

Lemma A2 records the population identities behind the double-robustness arguments. Part (a) is the population counterpart of the sample balancing condition (11); it holds at the pseudo-true parameter whether or not the working models are correctly specified, and it is what removes each respondent group's bias term. Parts (b)–(c) state that if the true density ratio ω_d^* or the true outcome evolution m_d^* belongs to the corresponding working family, then the pseudo-true parameter recovers it, so that $\mathcal{W}_d$ holds at λ_d^0 or $\mathcal{M}_d$ holds at β_d^0. The lemma is used in the proofs of Theorems 1–3.

Proof of Lemma A1. Fix $d \in \{0, 1\}$ and let $C_\Lambda = \sup_{\lambda_d \in \Lambda_d} \|\lambda_d\| < \infty$, which is finite by Assumption 3(iii). By Assumption 3(i), $|b(X)^\top \lambda_d| \le C_b C_\Lambda =: C$ almost surely, uniformly in $\lambda_d \in \Lambda_d$. Because g_d is continuous and strictly positive and $\dot{g}_d$ is continuous and strictly positive by Assumption 3(ii), the constants

$$0 < c_1 = \min_{|u| \le C} \dot{g}_d(u) \le \max_{|u| \le C} \dot{g}_d(u) = c_2 < \infty,$$
$$0 < c_3 = \min_{|u| \le C} g_d(u) \le \max_{|u| \le C} g_d(u) = c_4 < \infty,$$

are well defined, since a continuous positive function attains its bounds on a compact interval. By Assumption 3(ii), $c_5 := \max_{|u| \le C} |\ddot{g}_d(u)| < \infty$. Hence, g_d and $\dot{g}_d$ are Lipschitz on $[-C, C]$, with constants c_2 and c_5, respectively.

(a) The population objective $Q_d(\lambda_d) = \mathbb{E}[S_d G_d\{b(X)^\top \lambda_d\} - D\, b(X)^\top \lambda_d]$ is finite on Λ_d and twice continuously differentiable, with $\nabla^2 Q_d(\lambda_d) = \mathbb{E}[S_d \dot{g}_d\{b(X)^\top \lambda_d\} b(X) b(X)^\top] \succeq c_1 \mathbb{E}[S_d b(X) b(X)^\top] \succ 0$ by Assumption 3(i). Hence, Q_d is strictly convex on the convex set Λ_d, and the interior stationary point λ_d^0 of Assumption 3(iii) is its global minimizer. The same display at λ_d^0 shows that J_d is positive definite, and $J_d \preceq c_2 C_b^2 I_K$ shows that it is finite.

(b) The population projection equation is linear in β_d with coefficient matrix $\mathbb{E}[S_d \dot{g}_d^0(X) b(X) b(X)^\top] = J_d$, which by (a) is nonsingular; the unique solution is therefore (23), and it is finite because $\|b\| \le C_b$, $\dot{g}_d^0 \le c_2$ and $\mathbb{E}(\Delta Y^2) < \infty$. Since $\dot{g}_d^0 > 0$, the equation is the normal equation of a genuine weighted least-squares fit.

(c) The class $\{S_d G_d\{b(X)^\top \lambda_d\} - D\, b(X)^\top \lambda_d : \lambda_d \in \Lambda_d\}$ is uniformly bounded by $\max_{|u| \le C} |G_d(u)| + C$ and is Lipschitz in λ_d with constant $(c_4 + 1) C_b$, hence Glivenko–Cantelli; therefore, $\sup_{\lambda_d \in \Lambda_d} |\widehat{Q}_d(\lambda_d) - Q_d(\lambda_d)| \xrightarrow{p} 0$, where $\widehat{Q}_d$ is the sample objective in (12). Both

objectives are convex, and Q_d has a unique minimizer in the interior of Λ_d, so the standard argument for convex M-estimation (Newey & McFadden, 1994) gives $\hat{\lambda}_d \xrightarrow{p} \lambda_d^0$ and, with probability approaching one, $\hat{\lambda}_d$ lies in the interior of Λ_d and solves (11). Uniqueness holds with probability approaching one because $\nabla^2 \widehat{Q}_d \geq c_1 \hat{\mathbb{E}}[S_d bb^\top] \xrightarrow{p} \mathbb{E}[S_d bb^\top] \succ 0$. For β_d, $\hat{\beta}_d$ solves a linear system whose matrix $\hat{\mathbb{E}}[S_d \dot{g}_d\{b(X)^\top \hat{\lambda}_d\} bb^\top]$ converges in probability to J_d by the same uniform law and the continuity of $\dot{g}_d$, and whose right-hand side converges to $\mathbb{E}[S_d \dot{g}_d^0 b \Delta Y]$; hence, $\hat{\beta}_d \xrightarrow{p} \beta_d^0$.

(d) On the event that $\hat{\lambda}_d$ is an interior solution, the integral form of the mean-value expansion of the sample calibration equation gives

$$0 = \hat{\mathbb{E}}[\{S_d \omega_d^0(X) - D\} b(X)] + \bar{J}_d(\hat{\lambda}_d - \lambda_d^0),$$

where

$$\bar{J}_d = \int_0^1 \hat{\mathbb{E}}\big[S_d \dot{g}_d\{b(X)^\top [\lambda_d^0 + t(\hat{\lambda}_d - \lambda_d^0)]\} b(X) b(X)^\top\big] dt.$$

The integrand has a bounded envelope and is Lipschitz in the calibration parameter, with constant $c_5 C_b^3$. A uniform law of large numbers, consistency from part (c), and continuity therefore give $\bar{J}_d \xrightarrow{p} J_d \succ 0$. Rearranging and applying the central limit theorem to $\{D_i - S_{di}\omega_d^0(X_i)\} b(X_i)$, which has mean zero and a finite second moment, yields the stated asymptotic expansion for $\hat{\lambda}_d$.

For the projection coefficient, use ζ_d^0 and H_d as defined in part (d) of the lemma. The derivatives of the projection equation with respect to $\lambda_d^\top$ and $\beta_d^\top$ have population limits H_d and $-J_d$, respectively. They are continuous in the parameters and dominated by an integrable envelope, as shown in part (e). Their uniform laws of large numbers, consistency from part (c), and the integral mean-value expansion therefore give

$$\{J_d + o_p(1)\}\sqrt{n}(\hat{\beta}_d - \beta_d^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_d^0(O_i) + \{H_d + o_p(1)\}\sqrt{n}(\hat{\lambda}_d - \lambda_d^0).$$

Since J_d is nonsingular, $\sqrt{n}(\hat{\lambda}_d - \lambda_d^0) = O_p(1)$, and the sum on the right is $O_p(1)$, this first establishes $\sqrt{n}(\hat{\beta}_d - \beta_d^0) = O_p(1)$ and then yields

$$J_d \sqrt{n}(\hat{\beta}_d - \beta_d^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_d^0(O_i) + H_d \sqrt{n}(\hat{\lambda}_d - \lambda_d^0) + o_p(1).$$

Here, $\mathbb{E}[\zeta_d^0(O)] = 0$ by the population projection equation, and ζ_d^0 has a finite second moment. Substituting the expansion for $\hat{\lambda}_d$ gives the stated expansion for $\hat{\beta}_d$. Thus, both nuisance estimators are root-n consistent and asymptotically linear.

(e) Write $b = b(X)$, $u = b^\top \lambda_d$, and $r_\beta = \Delta Y - b^\top \beta_d$. The two functions used in Appendices A.8 and A.9 are

$$h_d(O; \lambda_d, \beta_d) = S_d g_d(u) r_\beta, \qquad \zeta_d(O; \lambda_d, \beta_d) = S_d \dot{g}_d(u) b\, r_\beta,$$

or products of them with D or components of b. Their parameter derivatives are

$$\nabla_{\lambda_d} h_d = S_d \dot{g}_d(u) b\, r_\beta, \qquad \nabla_{\beta_d} h_d = -S_d g_d(u) b,$$
$$\frac{\partial \zeta_d}{\partial \lambda_d^\top} = S_d \ddot{g}_d(u) bb^\top r_\beta, \qquad \frac{\partial \zeta_d}{\partial \beta_d^\top} = -S_d \dot{g}_d(u) bb^\top.$$

These derivatives exist and are continuous by Assumption 3(ii). The calibration estimating function $\{S_d g_d(u) - D\} b$ is also continuously differentiable, with derivative $S_d \dot{g}_d(u) bb^\top$.

Let $\mathcal{K}$ be a compact convex neighborhood of β_d^0 and $B_\beta = \sup_{\beta_d \in \mathcal{K}} \|\beta_d\|$. Since $|r_\beta| \leq |\Delta Y| + C_b B_\beta$, the functions above and their first derivatives are bounded in norm, uniformly on $\Lambda_d \times \mathcal{K}$, by

$$F_d(O) = A_d(1 + |\Delta Y|),$$

where $A_d < \infty$ depends only on C_b, C_Λ, c_2, c_4, c_5, and $\mathcal{K}$. Products with D or components of b obey a bound of the same form. Integrating the first derivatives along the segment between any two parameter values $\theta = (\lambda_d^\top, \beta_d^\top)^\top$ and $\tilde{\theta}$ gives

$$|h_d(O;\theta) - h_d(O;\tilde{\theta})| + \|\zeta_d(O;\theta) - \zeta_d(O;\tilde{\theta})\| \le F_d(O)\|\theta - \tilde{\theta}\|,$$

after increasing A_d if necessary. By Assumption 3(i), $\mathbb{E}[F_d(O)^4] < \infty$, so this is a square-integrable random Lipschitz coefficient and envelope, and the squared functions have integrable envelopes.

On the compact parameter neighborhood, the estimating functions, their first derivatives, and the squared estimating functions are continuous in the parameters and have integrable envelopes. Hence, uniform laws of large numbers apply. Together with consistency, these justify the convergence of the sample Jacobians in part (d) and the consistency of the variance estimator. At the limiting parameter values, square integrability ensures the ordinary central limit theorems used below. □

Proof of Lemma A2. (a) Applying the first identity in (A1) coordinatewise, together with $\mathbb{E}[Db(X)] = \mathbb{E}[e^*(X)b(X)]$, the population calibration equation in Assumption 3(iii) becomes $\mathbb{E}[\{a_d^*\omega_d^0 - e^*\}b] = 0$. (b) By (5), $a_d^*\omega_d^* = e^*$, so $\bar{\lambda}_d$ solves the population calibration equation; uniqueness in Assumption 3(iii) gives $\bar{\lambda}_d = \lambda_d^0$, and $\mathcal{U}_d$ holds with $a_d^*\omega_d^0 = a_d^*\omega_d^* = e^*$. (c) If $m_d^* = b^\top\bar{\beta}_d$, the second identity in (A1) gives

$$\mathbb{E}[S_d\dot{g}_d^0(X)b(X)\{\Delta Y - b(X)^\top\bar{\beta}_d\}] = \mathbb{E}[a_d^*(X)\dot{g}_d^0(X)b(X)\{m_d^*(X) - b(X)^\top\bar{\beta}_d\}] = 0,$$

so $\bar{\beta}_d$ solves the population projection equation; uniqueness (Lemma A1(b)) gives $\bar{\beta}_d = \beta_d^0$, and $\mathcal{M}_d$ holds with $\ell_d^0 = b(X)^\top\beta_d^0 = m_d^*$. □

Appendix A.6. Existence of a Finite Calibration Solution Under the Exponential Link

Here, we discuss the existence of a finite solution to the sample calibration problem in (12) under the exponential link; i.e., $g_d(\cdot) = \exp(\cdot)$, for which $G_d(\cdot) = \exp(\cdot)$. Under this link, (12) is an exponential tilting, or entropy balancing, problem (Graham et al., 2012; Hainmueller, 2012). Recall that the first element of $b(X)$ is the constant $b_1 \equiv 1$, and suppose $\hat{\mathbb{E}}[S_d b(X)b(X)^\top]$ is nonsingular. Then, the sample objective in (12), minimized over $\mathbb{R}^K$, has a minimizer only if the support condition

$$\left(\sum_{i=1}^n D_i\right)^{-1}\sum_{i=1}^n D_i\, b(X_i) \;\in\; \text{ri conv}\,\{b(X_i) : S_{di} = 1\}, \tag{A4}$$

where ri denotes the relative interior. Equivalently, $\bar{b}_T \in \text{ri conv}\,\mathcal{B}_d$, where $\bar{b}_T = n_1^{-1}\sum_{i=1}^n D_i b(X_i)$, $n_1 = \sum_{i=1}^n D_i$, and $\mathcal{B}_d = \{b(X_i) : S_{di} = 1\}$. The minimizer, when it exists, is unique because the objective is strictly convex.

Two consequences are worth recording. First, the population analogue of (A4) is implied by Assumptions 1(iii) and 2(ii) with the factorization (2), since these guarantee $a_d^*(X) > 0$ on the covariate support of the treated, so that the treated distribution of $b(X)$ is dominated by the group-d respondent distribution; the sample condition then holds with probability approaching one. Second, condition (A4) shows exactly how the risk of failure scales: (A4) is a statement in $\mathbb{R}^K$ about $\sum_i S_{di}$ points, so it becomes more demanding as K increases or as the number of respondents in the smaller group falls, which is the formal counterpart of separation in logistic regression and of the convex-hull condition in empirical likelihood. Condition (A4) therefore fails in exactly the situations practitioners worry about: a comparison group with few respondents; a covariate region populated by treated units but not by group-d respondents; or a dictionary rich enough that the treated mean escapes the K-dimensional hull.

Proof. Write $nL(\lambda) = \sum_{i:S_{di}=1} \exp\{b(X_i)^\top \lambda\} - n_1 \bar{b}_T^\top \lambda$ for n times the objective in (12). Its Hessian is $\sum_{i:S_{di}=1} \exp\{b(X_i)^\top \lambda\} b(X_i) b(X_i)^\top$, which is positive definite because $\hat{\mathbb{E}}[S_d bb^\top]$ is nonsingular and the exponential weights are strictly positive; hence, L is strictly convex. At a minimizer $\hat{\lambda}_d$ the first-order condition is $\sum_{i:S_{di}=1} \exp\{b(X_i)^\top \hat{\lambda}_d\} b(X_i) = n_1 \bar{b}_T$. Its first coordinate, which exists because $b_1 \equiv 1$, gives $\sum_{i:S_{di}=1} \exp\{b(X_i)^\top \hat{\lambda}_d\} = n_1$. Hence, $p_i = \exp\{b(X_i)^\top \hat{\lambda}_d\}/n_1$ are strictly positive, sum to one, and satisfy $\sum_i p_i b(X_i) = \bar{b}_T$, which exhibits $\bar{b}_T$ as a convex combination of $\mathcal{B}_d$ with all weights strictly positive, that is, $\bar{b}_T \in \operatorname{ri} \operatorname{conv} \mathcal{B}_d$. □

Appendix A.7. Proof of Theorem 1

Proof. If $\mathcal{W}_d$ and $\mathcal{M}_d$ hold for both $d \in \{0,1\}$, then the union condition (20) also holds. Hence, Equation (25) in Theorem 3 yields

$$\sqrt{n}(\widehat{\tau}_{\mathrm{JCB}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^0 + o_p(1) \xrightarrow{d} \mathcal{N}(0, V^0), \qquad V^0 = \mathbb{E}[(\psi_i^0)^2].$$

The proof of Equation (25) is provided in Appendix A.9. Obviously, when $\omega_d^0 = \omega_d^*$ and $\ell_d^0 = m_d^*$ for both d, we have $\psi_i^0 = \varphi_{\mathrm{eff}}(O_i)$ in (A10), and hence $V^0 = \mathbb{E}[\varphi_{\mathrm{eff}}(O)^2] = V_{\mathrm{eff}}$. □

Appendix A.8. Proof of Theorem 2

Proof. By Lemma A1(c), $\hat{\lambda}_d \xrightarrow{p} \lambda_d^0$ and the law of large numbers gives $\hat{p} \xrightarrow{p} p > 0$. The envelope conditions supplied by Lemma A1(e) justify a uniform law of large numbers for $\lambda_d \mapsto \hat{\mathbb{E}}[S_d g_d\{b(X)^\top \lambda_d\} \Delta Y]$ in a neighborhood of λ_d^0, so by the continuous mapping theorem,

$$\widehat{\tau}_{\mathrm{JCB}} \xrightarrow{p} \frac{1}{p} \mathbb{E}[\{S_1 \omega_1^0(X) - S_0 \omega_0^0(X)\} \Delta Y] = \frac{1}{p}[\mathbb{E}[a_1^* \omega_1^0 m_1^*] - \mathbb{E}[a_0^* \omega_0^0 m_0^*]], \tag{A5}$$

where the equality uses (A1). Combining (4) with (A1) evaluated at ω_d^* and using $a_d^* \omega_d^* = e^*$ gives $p\tau = \mathbb{E}[e^*(m_1^* - m_0^*)]$. Hence,

$$p\,(\operatorname{plim} \widehat{\tau}_{\mathrm{JCB}} - \tau) = \mathbb{E}[(a_1^* \omega_1^0 - e^*) m_1^*] - \mathbb{E}[(a_0^* \omega_0^0 - e^*) m_0^*]. \tag{A6}$$

Because $\ell_d^0 = b(X)^\top \beta_d^0$, Lemma A2(a) yields $\mathbb{E}[(a_d^* \omega_d^0 - e^*) \ell_d^0] = 0$; subtracting these zero terms from (A6) gives the product-bias formula (21). Consider component d. Under $\mathcal{W}_d$, Lemma A2(b) gives $a_d^* \omega_d^0 - e^* = 0$ almost surely. Under $\mathcal{M}_d$, $m_d^* = b^\top \beta_d^0$ is linear in b, so $\mathbb{E}[(a_d^* \omega_d^0 - e^*) m_d^*] = 0$ by Lemma A2(a). Under the union condition (20), both terms in (A6) therefore vanish, and $\widehat{\tau}_{\mathrm{JCB}} \xrightarrow{p} \tau$. □

Appendix A.9. Proof of Theorem 3

Proof. Let $\beta^0 = (\beta_1^{0\top}, \beta_0^{0\top})^\top$ be defined by (23). The proof proceeds in six steps.

Step 1 (Augmented representation). The sample calibration condition (11) implies $\hat{\mathbb{E}}[S_d \widehat{\omega}_d(X) b(X)^\top \beta_d] = \hat{\mathbb{E}}[D b(X)^\top \beta_d]$ for every $\beta_d \in \mathbb{R}^K$; adding and subtracting these quantities in (13) yields (14). Evaluating (14) at $\beta = \beta^0$ and subtracting τ,

$$\hat{p}\,(\widehat{\tau}_{\mathrm{JCB}} - \tau) = \hat{\mathbb{E}}\Big[D\{b(X)^\top(\beta_1^0 - \beta_0^0) - \tau\} + \sum_{d=0}^{1} s_d S_d g_d\{b(X)^\top \hat{\lambda}_d\}\{\Delta Y - b(X)^\top \beta_d^0\}\Big]. \tag{A7}$$

Step 2 (Expansion in $\hat{\lambda}$). Lemma A1(d)–(e) permits a mean-value expansion of $\lambda_d \mapsto \hat{\mathbb{E}}[S_d g_d\{b(X)^\top \lambda_d\}\{\Delta Y - b(X)^\top \beta_d^0\}]$ around λ_d^0; the uniform convergence of the sample derivative to $\kappa_d(\beta_d^0)$ in shrinking neighborhoods of λ_d^0, combined with $\hat{\lambda}_d - \lambda_d^0 = O_p(n^{-1/2})$, gives

$$\hat{p}\,(\widehat{\tau}_{\mathrm{JCB}} - \tau) = p\,\hat{\mathbb{E}}(\psi_i^0) + \sum_{d=0}^{1} s_d\, \kappa_d(\beta_d^0)^\top (\hat{\lambda}_d - \lambda_d^0) + o_p(n^{-1/2}), \tag{A8}$$

which yields (24) after dividing by $\hat{p}$ and using $\hat{p} \xrightarrow{p} p$. By the definition of β_d^0 in (23), $\kappa_d(\beta_d^0) = 0$, so the second term on the right-hand side of (A8) vanishes.

Step 3 (Mean zero under the union model). By (A1) and rearrangement,

$$p\,\mathbb{E}(\psi_i^0) = [\mathbb{E}[a_1^* \omega_1^0 m_1^*] - \mathbb{E}[a_0^* \omega_0^0 m_0^*] - p\tau] - \sum_{d=0}^{1} s_d\,\mathbb{E}[(a_d^* \omega_d^0 - e^*) b(X)^\top \beta_d^0]. \tag{A9}$$

The bracketed term equals the right-hand side of (A6) and vanishes under (20) by the proof of Theorem 2; each summand in the second term vanishes by Lemma A2(a), regardless of the value of β_d^0. Hence, $\mathbb{E}(\psi_i^0) = 0$.

Step 4 (Asymptotic normality). Assumption 3(i) and Lemma A1(e) ensure $V^0 = \mathbb{E}[(\psi_i^0)^2] < \infty$. By Steps 1–3 and the central limit theorem, $\sqrt{n}\,\hat{\mathbb{E}}(\psi_i^0) \xrightarrow{d} \mathcal{N}(0, V^0)$. Since $\hat{p} \xrightarrow{p} p > 0$, Slutsky's theorem applied to $\sqrt{n}(\hat{\tau}_{\mathrm{JCB}} - \tau) = (p/\hat{p})\,\sqrt{n}\,\hat{\mathbb{E}}(\psi_i^0) + o_p(1)$ yields (25).

Step 5 (Consistency of the variance estimator). The quantity $\widehat{\psi}_i$ in (16) equals ψ_i^0 with $(\lambda^0, \beta^0, \tau, p)$ replaced by $(\hat{\lambda}, \hat{\beta}, \hat{\tau}_{\mathrm{JCB}}, \hat{p})$. All replaced quantities are consistent by Lemma A1(c)–(d) and Steps 1–4, the estimating function is continuous in the parameters in a neighborhood of their limits, and Assumption 3(i) with Lemma A1(e) supplies square-integrable envelopes; a standard uniform law of large numbers then gives $\widehat{V} = \hat{\mathbb{E}}(\widehat{\psi}_i^2) \xrightarrow{p} \mathbb{E}[(\psi_i^0)^2] = V^0$.

Step 6 (Local efficiency). Suppose $\mathcal{U}_d$ and $\mathcal{M}_d$ hold for both $d \in \{0,1\}$. Lemma A2(b)–(c) give $\omega_d^0 = \omega_d^*$ and $\ell_d^0 = m_d^*$, so

$$\begin{aligned} \psi_i^0 = \frac{1}{p}\Big[& D_i\{m_1^*(X_i) - m_0^*(X_i) - \tau\} \\ & + S_{1i}\omega_1^*(X_i)\{\Delta Y_i - m_1^*(X_i)\} - S_{0i}\omega_0^*(X_i)\{\Delta Y_i - m_0^*(X_i)\}\Big]. \end{aligned} \tag{A10}$$

The right-hand side of (A10) coincides with the efficient influence function $\varphi_{\mathrm{eff}}(O_i)$ derived in Proposition 1. Hence, $V^0 = V_{\mathrm{eff}}$ equals the semiparametric efficiency bound for the coarsened observed-data model. □

Appendix B. Additional Numerical Results

This appendix collects the full numerical output of the Monte Carlo study of Section 6. Table A1 records the population quantities that define each design, and Tables A2 and A3 report performance and weight diagnostics for every design, estimator and sample size. Appendix B.1 then presents an additional design in which respondents are scarce.

Table A1. Probability bounds and population group response rates.

Design	e^* Range	r_1^* Range	r_0^* Range	p	Response 1/0
DGP1	[0.3854, 0.5826]	[0.1738, 0.7788]	[0.0889, 0.8052]	0.4507	0.4034/0.2401
DGP2	[0.3854, 0.5826]	[0.1738, 0.7788]	[0.0488, 0.5013]	0.4507	0.4034/0.1976
DGP3	[0.3854, 0.5826]	[0.1054, 0.4874]	[0.0889, 0.8052]	0.4507	0.3412/0.2401
DGP4	[0.3854, 0.5826]	[0.1054, 0.4874]	[0.0488, 0.5013]	0.4507	0.3412/0.1976
DGP5, $\rho = 0.2$	[0.4000, 0.4000]	[0.6000, 0.6000]	[0.6000, 0.6000]	0.4000	0.6000/0.6000
DGP5, $\rho = 0.8$	[0.4000, 0.4000]	[0.6000, 0.6000]	[0.6000, 0.6000]	0.4000	0.6000/0.6000
DGP6-N	[0.0573, 0.6900]	[0.3543, 0.7685]	[0.2689, 0.6457]	0.2992	0.5218/0.4335
DGP6-L	[0.0573, 0.6900]	[0.3543, 0.7685]	[0.2689, 0.6457]	0.2992	0.5218/0.4335
DGP7	[0.2755, 0.6067]	[0.1738, 0.7788]	[0.0867, 0.8899]	0.3798	0.4069/0.1805
DGP8	[0.3854, 0.5826]	[0.0579, 0.2596]	[0.0296, 0.2684]	0.4507	0.1345/0.0800

Notes: The e^*, r_1^*, and r_0^* range columns give the smallest and largest values of the treatment probability and of the two group response probabilities over the closed support $[-1,1]^2$ of (Z_1, Z_2), obtained by exact optimization of the indices rather than by evaluating a finite grid; they are constant in DGP5. $p = \mathbb{E}\{e^*(X)\}$ is the population treated share. Response 1/0 gives $\mathbb{P}(R = 1 \mid D = 1)/\mathbb{P}(R = 1 \mid D = 0)$, the population response rate within each treatment arm, which differs from the unconditional average $\mathbb{E}\{r_d^*(X)\}$.

Table A2. Complete Monte Carlo performance results (2000 attempts per cell).

Design, n	Method	Bias (MCSE)	SD	SE	RMSE	Cover. (MCSE)	Success
DGP1, 250	JCB	−0.0464 (0.0087)	0.389	0.346	0.392	0.9100 (0.0064)	2000
DGP1, 250	Naive	−0.3493 (0.0077)	0.344	0.337	0.490	0.8020 (0.0089)	2000
DGP1, 250	Plug-in	0.0314 (0.0095)	0.427	0.388	0.428	0.9215 (0.0060)	2000
DGP1, 500	JCB	−0.0104 (0.0059)	0.265	0.251	0.265	0.9350 (0.0055)	2000
DGP1, 500	Naive	−0.3353 (0.0053)	0.238	0.239	0.411	0.7015 (0.0102)	2000
DGP1, 500	Plug-in	0.0713 (0.0065)	0.291	0.274	0.299	0.9305 (0.0057)	2000
DGP1, 1000	JCB	−0.0100 (0.0041)	0.183	0.179	0.183	0.9480 (0.0050)	2000
DGP1, 1000	Naive	−0.3492 (0.0039)	0.173	0.170	0.390	0.4590 (0.0111)	2000
DGP1, 1000	Plug-in	0.0714 (0.0044)	0.196	0.193	0.209	0.9345 (0.0055)	2000
DGP1, 2000	JCB	−0.0077 (0.0028)	0.125	0.127	0.125	0.9515 (0.0048)	2000
DGP1, 2000	Naive	−0.3507 (0.0027)	0.120	0.120	0.371	0.1690 (0.0084)	2000
DGP1, 2000	Plug-in	0.0718 (0.0030)	0.134	0.136	0.152	0.9160 (0.0062)	2000
DGP1, 5000	JCB	−0.0041 (0.0018)	0.081	0.080	0.081	0.9455 (0.0051)	2000
DGP1, 5000	Naive	−0.3485 (0.0017)	0.076	0.076	0.357	0.0035 (0.0013)	2000
DGP1, 5000	Plug-in	0.0751 (0.0019)	0.086	0.086	0.114	0.8580 (0.0078)	2000
DGP2, 250	JCB	−0.0103 (0.0070)	0.312	0.279	0.312	0.9130 (0.0063)	2000
DGP2, 250	Naive	−0.3841 (0.0069)	0.310	0.301	0.494	0.7320 (0.0099)	2000
DGP2, 250	Plug-in	0.0432 (0.0073)	0.326	0.298	0.329	0.9200 (0.0061)	2000
DGP2, 500	JCB	−0.0050 (0.0047)	0.211	0.203	0.211	0.9385 (0.0054)	2000
DGP2, 500	Naive	−0.3895 (0.0049)	0.219	0.214	0.447	0.5600 (0.0111)	2000
DGP2, 500	Plug-in	0.0486 (0.0049)	0.218	0.212	0.223	0.9295 (0.0057)	2000
DGP2, 1000	JCB	−0.0043 (0.0032)	0.144	0.144	0.144	0.9415 (0.0052)	2000
DGP2, 1000	Naive	−0.3885 (0.0034)	0.151	0.151	0.417	0.2690 (0.0099)	2000
DGP2, 1000	Plug-in	0.0478 (0.0033)	0.148	0.149	0.156	0.9340 (0.0056)	2000
DGP2, 2000	JCB	0.0013 (0.0023)	0.102	0.102	0.102	0.9480 (0.0050)	2000
DGP2, 2000	Naive	−0.3853 (0.0024)	0.106	0.107	0.399	0.0450 (0.0046)	2000
DGP2, 2000	Plug-in	0.0541 (0.0024)	0.106	0.105	0.119	0.9185 (0.0061)	2000
DGP2, 5000	JCB	0.0016 (0.0015)	0.067	0.065	0.067	0.9380 (0.0054)	2000
DGP2, 5000	Naive	−0.3867 (0.0016)	0.070	0.068	0.393	0.0000 (0.0000)	2000
DGP2, 5000	Plug-in	0.0542 (0.0015)	0.069	0.067	0.088	0.8625 (0.0077)	2000
DGP3, 250	JCB	−0.0244 (0.0081)	0.363	0.327	0.363	0.9155 (0.0062)	2000
DGP3, 250	Naive	−0.3729 (0.0076)	0.339	0.336	0.504	0.7850 (0.0092)	2000
DGP3, 250	Plug-in	0.0060 (0.0089)	0.396	0.363	0.396	0.9265 (0.0058)	2000
DGP3, 500	JCB	−0.0159 (0.0056)	0.249	0.240	0.249	0.9350 (0.0055)	2000
DGP3, 500	Naive	−0.3789 (0.0053)	0.239	0.238	0.448	0.6290 (0.0108)	2000
DGP3, 500	Plug-in	0.0111 (0.0059)	0.264	0.257	0.264	0.9455 (0.0051)	2000
DGP3, 1000	JCB	−0.0097 (0.0039)	0.175	0.171	0.175	0.9365 (0.0055)	2000
DGP3, 1000	Naive	−0.3778 (0.0038)	0.171	0.169	0.415	0.3860 (0.0109)	2000
DGP3, 1000	Plug-in	0.0167 (0.0041)	0.184	0.181	0.185	0.9380 (0.0054)	2000
DGP3, 2000	JCB	−0.0034 (0.0027)	0.120	0.121	0.120	0.9510 (0.0048)	2000
DGP3, 2000	Naive	−0.3759 (0.0027)	0.120	0.120	0.394	0.1240 (0.0074)	2000
DGP3, 2000	Plug-in	0.0238 (0.0028)	0.127	0.127	0.129	0.9460 (0.0051)	2000
DGP3, 5000	JCB	−0.0026 (0.0017)	0.074	0.076	0.074	0.9545 (0.0047)	2000
DGP3, 5000	Naive	−0.3775 (0.0017)	0.075	0.076	0.385	0.0005 (0.0005)	2000
DGP3, 5000	Plug-in	0.0243 (0.0017)	0.078	0.080	0.082	0.9480 (0.0050)	2000
DGP4, 250	JCB	0.0106 (0.0066)	0.294	0.257	0.294	0.8970 (0.0068)	2000
DGP4, 250	Naive	−0.4070 (0.0070)	0.313	0.298	0.514	0.7075 (0.0102)	2000
DGP4, 250	Plug-in	0.0109 (0.0065)	0.293	0.263	0.293	0.9090 (0.0064)	2000
DGP4, 500	JCB	−0.0044 (0.0043)	0.194	0.188	0.194	0.9375 (0.0054)	2000
DGP4, 500	Naive	−0.4204 (0.0048)	0.214	0.213	0.472	0.4950 (0.0112)	2000
DGP4, 500	Plug-in	−0.0043 (0.0043)	0.194	0.189	0.194	0.9395 (0.0053)	2000
DGP4, 1000	JCB	0.0034 (0.0031)	0.138	0.134	0.138	0.9430 (0.0052)	2000
DGP4, 1000	Naive	−0.4137 (0.0034)	0.150	0.150	0.440	0.2115 (0.0091)	2000
DGP4, 1000	Plug-in	0.0034 (0.0031)	0.138	0.134	0.138	0.9425 (0.0052)	2000
DGP4, 2000	JCB	0.0017 (0.0021)	0.095	0.095	0.095	0.9490 (0.0049)	2000
DGP4, 2000	Naive	−0.4157 (0.0024)	0.108	0.107	0.430	0.0260 (0.0036)	2000
DGP4, 2000	Plug-in	0.0017 (0.0021)	0.095	0.095	0.095	0.9485 (0.0049)	2000
DGP4, 5000	JCB	−0.0021 (0.0014)	0.061	0.060	0.061	0.9455 (0.0051)	2000

Table A2. *Cont.*

Design, n	Method	Bias (MCSE)	SD	SE	RMSE	Cover. (MCSE)	Success
DGP4, 5000	Naive	−0.4216 (0.0015)	0.066	0.067	0.427	0.0000 (0.0000)	2000
DGP4, 5000	Plug-in	−0.0021 (0.0014)	0.061	0.060	0.061	0.9465 (0.0050)	2000
DGP5, $\rho = 0.2$, 500	JCB	0.0004 (0.0034)	0.150	0.149	0.150	0.9460 (0.0051)	2000
DGP5, $\rho = 0.2$, 500	Naive	−0.0010 (0.0039)	0.174	0.170	0.174	0.9365 (0.0055)	2000
DGP5, $\rho = 0.2$, 500	Plug-in	0.0004 (0.0034)	0.150	0.149	0.150	0.9460 (0.0051)	2000
DGP5, $\rho = 0.2$, 500	Testa	−0.0019 (0.0032)	0.141	0.138	0.141	0.9415 (0.0052)	2000
DGP5, $\rho = 0.2$, 1000	JCB	−0.0002 (0.0024)	0.109	0.106	0.109	0.9460 (0.0051)	2000
DGP5, $\rho = 0.2$, 1000	Naive	0.0000 (0.0027)	0.122	0.120	0.122	0.9520 (0.0048)	2000
DGP5, $\rho = 0.2$, 1000	Plug-in	−0.0002 (0.0024)	0.109	0.106	0.109	0.9460 (0.0051)	2000
DGP5, $\rho = 0.2$, 1000	Testa	−0.0004 (0.0023)	0.101	0.097	0.101	0.9425 (0.0052)	2000
DGP5, $\rho = 0.2$, 5000	JCB	−0.0012 (0.0011)	0.047	0.047	0.047	0.9470 (0.0050)	2000
DGP5, $\rho = 0.2$, 5000	Naive	−0.0009 (0.0012)	0.054	0.054	0.054	0.9495 (0.0049)	2000
DGP5, $\rho = 0.2$, 5000	Plug-in	−0.0012 (0.0011)	0.047	0.047	0.047	0.9470 (0.0050)	2000
DGP5, $\rho = 0.2$, 5000	Testa	−0.0002 (0.0010)	0.044	0.043	0.044	0.9405 (0.0053)	2000
DGP5, $\rho = 0.8$, 500	JCB	−0.0029 (0.0017)	0.074	0.075	0.074	0.9535 (0.0047)	2000
DGP5, $\rho = 0.8$, 500	Naive	−0.0025 (0.0024)	0.110	0.111	0.110	0.9525 (0.0048)	2000
DGP5, $\rho = 0.8$, 500	Plug-in	−0.0029 (0.0017)	0.074	0.075	0.074	0.9535 (0.0047)	2000
DGP5, $\rho = 0.8$, 500	Testa	−0.0027 (0.0021)	0.095	0.095	0.095	0.9520 (0.0048)	2000
DGP5, $\rho = 0.8$, 1000	JCB	−0.0005 (0.0012)	0.052	0.053	0.052	0.9535 (0.0047)	2000
DGP5, $\rho = 0.8$, 1000	Naive	−0.0010 (0.0017)	0.077	0.079	0.077	0.9575 (0.0045)	2000
DGP5, $\rho = 0.8$, 1000	Plug-in	−0.0005 (0.0012)	0.052	0.053	0.052	0.9535 (0.0047)	2000
DGP5, $\rho = 0.8$, 1000	Testa	0.0015 (0.0015)	0.067	0.067	0.067	0.9545 (0.0047)	2000
DGP5, $\rho = 0.8$, 5000	JCB	0.0007 (0.0005)	0.024	0.024	0.024	0.9510 (0.0048)	2000
DGP5, $\rho = 0.8$, 5000	Naive	0.0005 (0.0008)	0.035	0.035	0.035	0.9485 (0.0049)	2000
DGP5, $\rho = 0.8$, 5000	Plug-in	0.0007 (0.0005)	0.024	0.024	0.024	0.9510 (0.0048)	2000
DGP5, $\rho = 0.8$, 5000	Testa	0.0004 (0.0007)	0.030	0.030	0.030	0.9580 (0.0045)	2000
DGP6-N, 500	JCB	−0.0289 (0.0050)	0.222	0.208	0.224	0.9250 (0.0059)	2000
DGP6-N, 500	Naive	−0.0962 (0.0051)	0.229	0.220	0.248	0.9135 (0.0063)	2000
DGP6-N, 500	Plug-in	−0.0033 (0.0052)	0.233	0.219	0.233	0.9350 (0.0055)	2000
DGP6-N, 1000	JCB	−0.0249 (0.0034)	0.151	0.147	0.153	0.9350 (0.0055)	2000
DGP6-N, 1000	Naive	−0.0972 (0.0035)	0.158	0.155	0.185	0.8920 (0.0069)	2000
DGP6-N, 1000	Plug-in	−0.0014 (0.0036)	0.159	0.155	0.159	0.9425 (0.0052)	2000
DGP6-L, 500	JCB	0.0017 (0.0034)	0.153	0.151	0.153	0.9500 (0.0049)	2000
DGP6-L, 500	Naive	0.1537 (0.0038)	0.169	0.167	0.228	0.8490 (0.0080)	2000
DGP6-L, 500	Plug-in	0.0014 (0.0034)	0.153	0.151	0.153	0.9515 (0.0048)	2000
DGP6-L, 1000	JCB	0.0023 (0.0024)	0.109	0.107	0.109	0.9425 (0.0052)	2000
DGP6-L, 1000	Naive	0.1510 (0.0027)	0.119	0.119	0.192	0.7525 (0.0096)	2000
DGP6-L, 1000	Plug-in	0.0022 (0.0024)	0.109	0.107	0.109	0.9445 (0.0051)	2000
DGP7, 500	JCB	−0.0233 (0.0066)	0.295	0.278	0.296	0.9250 (0.0059)	2000
DGP7, 500	JCB, $K = 2$	−0.0445 (0.0069)	0.307	0.291	0.310	0.9285 (0.0058)	2000
DGP7, 500	JCB, $K = 4$	−0.0047 (0.0057)	0.253	0.237	0.253	0.9300 (0.0057)	1999
DGP7, 500	Naive	−0.3799 (0.0060)	0.269	0.264	0.465	0.6930 (0.0103)	2000
DGP7, 500	Plug-in	0.0994 (0.0073)	0.327	0.313	0.342	0.9225 (0.0060)	2000
DGP7, 1000	JCB	−0.0106 (0.0046)	0.207	0.198	0.207	0.9415 (0.0052)	2000
DGP7, 1000	JCB, $K = 2$	−0.0337 (0.0048)	0.214	0.206	0.217	0.9390 (0.0054)	2000
DGP7, 1000	JCB, $K = 4$	−0.0022 (0.0040)	0.181	0.169	0.181	0.9315 (0.0056)	2000
DGP7, 1000	Naive	−0.3787 (0.0042)	0.188	0.187	0.423	0.4690 (0.0112)	2000
DGP7, 1000	Plug-in	0.1093 (0.0051)	0.226	0.218	0.251	0.9155 (0.0062)	2000
DGP8, 250	JCB	−0.1889 (0.0177)	0.747	0.479	0.770	0.7668 (0.0100)	1784
DGP8, 250	Naive	−0.3469 (0.0137)	0.611	0.577	0.703	0.8705 (0.0075)	2000
DGP8, 250	Plug-in	−0.2645 (0.0209)	0.933	0.583	0.970	0.7863 (0.0092)	1998
DGP8, 250	Plug-in (unclipped)	−0.2595 (0.0211)	0.943	0.594	0.978	0.7878 (0.0091)	1998
DGP8, 500	JCB	−0.1022 (0.0110)	0.491	0.397	0.501	0.8763 (0.0074)	1989
DGP8, 500	Naive	−0.3575 (0.0094)	0.421	0.409	0.552	0.8420 (0.0082)	2000
DGP8, 500	Plug-in	−0.0968 (0.0123)	0.551	0.458	0.560	0.8910 (0.0070)	2000
DGP8, 500	Plug-in (unclipped)	−0.0925 (0.0127)	0.566	0.462	0.573	0.8910 (0.0070)	2000
DGP8, 1000	JCB	−0.0347 (0.0073)	0.325	0.303	0.327	0.9250 (0.0059)	2000
DGP8, 1000	Naive	−0.3475 (0.0066)	0.295	0.293	0.456	0.7605 (0.0095)	2000

Table A2. *Cont.*

Design, n	Method	Bias (MCSE)	SD	SE	RMSE	Cover. (MCSE)	Success
DGP8, 1000	Plug-in	−0.0208 (0.0079)	0.353	0.333	0.354	0.9285 (0.0058)	2000
DGP8, 1000	Plug-in (unclipped)	−0.0205 (0.0079)	0.354	0.333	0.354	0.9280 (0.0058)	2000

Notes: "Design, n" identifies the design and the sample size, and "Method" the estimator, with "Plug-in (unclipped)" the plug-in variant that leaves fitted probabilities unclipped. "Bias" is the average deviation of the replicated estimates from the design-specific true ATT; "SD" is their sample standard deviation; "SE" is the average estimated standard error; and "RMSE" is the square root of the mean squared deviation from that true ATT. "Cover." is the proportion of nominal 95% Wald intervals that contain the true ATT. "Success" counts the replications with a valid standard error out of 2000 attempts; every statistic conditions on method-specific success.

Table A3. JCB response, weight, and relative-variance diagnostics.

Design, n	$\hat{p}$	Resp. 1/0	ESS 1/0	Max w, 1/0	P95 Max w_0	$w_1 < 1$ (%)	$\overline{\widehat{V}}/V_J$
DGP1, 250	0.451	0.404/0.240	36.5/26.3	5.9/7.8	16.5	2.84	1.286
DGP1, 500	0.452	0.403/0.240	74.2/55.2	6.0/7.7	13.6	0.89	1.345
DGP1, 1000	0.451	0.404/0.240	150.2/112.4	5.9/7.6	11.8	0.15	1.356
DGP1, 2000	0.451	0.403/0.239	300.7/225.3	5.9/7.7	10.5	0.01	1.363
DGP1, 5000	0.451	0.404/0.240	756.3/568.7	5.9/7.6	9.3	0.00	1.358
DGP2, 250	0.451	0.404/0.197	36.4/21.8	5.9/9.2	21.0	3.00	1.057
DGP2, 500	0.450	0.403/0.198	74.1/46.1	5.9/8.9	16.3	0.81	1.104
DGP2, 1000	0.451	0.403/0.197	150.3/93.9	5.9/8.9	14.3	0.12	1.099
DGP2, 2000	0.451	0.404/0.197	301.2/189.7	5.9/9.0	12.6	0.01	1.106
DGP2, 5000	0.451	0.403/0.198	755.1/478.7	5.9/9.0	11.2	0.00	1.107
DGP3, 250	0.451	0.340/0.241	31.1/26.6	6.8/7.5	16.7	1.67	1.379
DGP3, 500	0.451	0.342/0.240	64.5/54.8	6.7/7.7	13.7	0.32	1.462
DGP3, 1000	0.451	0.341/0.240	129.8/111.9	6.8/7.7	11.5	0.02	1.473
DGP3, 2000	0.451	0.342/0.240	261.2/226.7	6.8/7.5	10.3	0.00	1.472
DGP3, 5000	0.451	0.341/0.240	654.0/568.4	6.8/7.6	9.3	0.00	1.472
DGP4, 250	0.451	0.343/0.197	31.6/21.8	6.7/9.3	20.2	1.57	0.891
DGP4, 500	0.451	0.340/0.197	64.2/45.6	6.8/9.1	17.0	0.32	0.940
DGP4, 1000	0.450	0.340/0.198	128.8/94.4	6.8/9.0	14.0	0.02	0.946
DGP4, 2000	0.451	0.341/0.197	260.4/190.1	6.9/9.0	12.7	0.00	0.948
DGP4, 5000	0.451	0.341/0.197	651.8/478.6	6.9/9.0	11.5	0.00	0.951
DGP5, $\rho = 0.2$, 500	0.400	0.600/0.601	119.6/176.3	1.9/1.4	2.0	0.00	1.004
DGP5, $\rho = 0.2$, 1000	0.400	0.600/0.600	239.3/355.8	1.9/1.3	1.7	0.00	1.003
DGP5, $\rho = 0.2$, 5000	0.400	0.600/0.600	1198.0/1796.2	1.7/1.2	1.3	0.00	1.001
DGP5, $\rho = 0.8$, 500	0.400	0.601/0.601	119.6/176.5	1.9/1.4	2.0	0.00	1.004
DGP5, $\rho = 0.8$, 1000	0.400	0.599/0.600	239.1/355.8	1.8/1.3	1.7	0.00	1.002
DGP5, $\rho = 0.8$, 5000	0.400	0.600/0.600	1200.8/1797.0	1.7/1.2	1.3	0.00	1.000
DGP6-N, 500	0.299	0.521/0.435	74.5/100.5	2.7/3.4	5.6	0.49	0.882
DGP6-N, 1000	0.299	0.521/0.434	150.9/202.7	2.6/3.6	5.2	0.13	0.885
DGP6-L, 500	0.299	0.521/0.434	74.7/99.8	2.7/3.4	5.5	0.46	0.995
DGP6-L, 1000	0.299	0.522/0.433	150.9/201.7	2.7/3.6	5.1	0.16	1.001
DGP7, 500	0.380	0.405/0.180	61.7/45.7	5.9/7.8	14.6	1.19	1.356
DGP7, 500 (K2)	0.380	0.405/0.180	62.4/47.1	5.6/6.8	12.5	0.41	1.487
DGP7, 500 (K4)	0.380	0.405/0.180	60.9/44.3	6.1/8.3	17.9	1.20	0.982
DGP7, 1000	0.379	0.407/0.180	125.1/93.2	5.9/7.7	12.2	0.23	1.365
DGP7, 1000 (K2)	0.379	0.407/0.180	125.6/94.9	5.7/7.0	10.7	0.04	1.485
DGP7, 1000 (K4)	0.379	0.407/0.180	124.4/91.8	6.0/8.0	14.7	0.30	0.998
DGP8, 250	0.450	0.133/0.079	11.0/8.0	17.7/22.4	54.4	2.60	1.153
DGP8, 500	0.451	0.134/0.080	23.6/16.9	17.8/23.9	55.0	0.25	1.494
DGP8, 1000	0.450	0.135/0.080	48.9/36.4	17.7/22.5	45.1	0.02	1.710

Notes: "$\hat{p}$" is the replication mean of the realized treated share. "Resp. 1/0" gives the replication means of the realized response rates among treated and control units. "ESS 1/0" gives the replication medians, over successful fits, of the effective sample size $(\sum_i S_{di}\widehat{\omega}_d)^2/\sum_i S_{di}\widehat{\omega}_d^2$ for the two respondent groups. "Max w, 1/0" gives the replication medians of the largest calibrated weight in each group, and "P95 max w_0" the 95th percentile across replications of that largest control weight. "$w_1 < 1$ (%)" is the average share of treated respondents whose fitted weight falls below one, a region the true density ratio $\omega_1^* \geq 1$ excludes. "$\overline{\widehat{V}}/V_J$" is the ratio of $\overline{\widehat{V}}$, the Monte Carlo average of $n\widehat{\mathrm{SE}}^2$, to V_J, the semiparametric efficiency bound of the coarsened model in Proposition 1 at each design's generating distribution.

Appendix B.1. Weak Response

Design DGP8 divides both response probabilities of DGP1 by three, preserving correctness of both exponential density-ratio models by adding $\log 3$ to their intercepts. The control-group response probability can be as low as 0.0296, and the population control-group response rate is approximately 0.0800. We use $n \in \{250, 500, 1000\}$.

Table A4. Weak response: Failures and conditional coverage.

n	Estimator	Success	Fail (%)	Bias (MCSE)	Cover. (MCSE)	Joint
250	JCB	1784	10.80	−0.1889 (0.0177)	0.7668 (0.0100)	0.684
250	Plug-in	1998	0.10	−0.2645 (0.0209)	0.7863 (0.0092)	0.785
250	Plug-in (unclipped)	1998	0.10	−0.2595 (0.0211)	0.7878 (0.0091)	0.787
500	JCB	1989	0.55	−0.1022 (0.0110)	0.8763 (0.0074)	0.872
500	Plug-in	2000	0.00	−0.0968 (0.0123)	0.8910 (0.0070)	0.891
500	Plug-in (unclipped)	2000	0.00	−0.0925 (0.0127)	0.8910 (0.0070)	0.891
1000	JCB	2000	0.00	−0.0347 (0.0073)	0.9250 (0.0059)	0.925
1000	Plug-in	2000	0.00	−0.0208 (0.0079)	0.9285 (0.0058)	0.928
1000	Plug-in (unclipped)	2000	0.00	−0.0205 (0.0079)	0.9280 (0.0058)	0.928

Notes: There are 2000 attempts per row. Success counts valid estimates with valid SEs. Bias conditions on point-estimate success, and coverage on valid-SE success. Joint is the proportion of all attempts that both succeed and cover the true ATT. The default plug-in clips fitted probabilities to $[0.01, 0.99]$; the unclipped version uses the same generated data.

DGP8 exposes an important finite-sample limitation. At $n = 250$, JCB fails in 216 of 2000 attempts under the stated implementation; conditional coverage is 0.7668 and the proportion of all attempts yielding a covering interval is only 0.6840. Table A3 shows what drives this. The median effective sample sizes among successful fits are 11.0 for the treated respondents and 8.0 for the controls, while the median largest calibrated weight is 17.7 and 22.4 in the two groups and the 95th percentile of the largest control weight is 54.4, so a handful of respondents carries the estimate; 2.60% of treated respondents even receive a fitted weight below one, which $\omega_1^* \geq 1$ excludes. Both the failures and the undercoverage recede as the sample grows: at $n = 1000$ every fit succeeds, the effective sample sizes rise to 48.9 and 36.4, and coverage reaches 0.9250. None of this is an artifact of trimming. The calibrated weights of JCB are never trimmed, and the $[0.01, 0.99]$ clipping that the plug-in applies to its fitted probabilities is immaterial here: removing it leaves that estimator's bias and coverage essentially unchanged, so the difficulty lies in the scarcity of respondents rather than in any stabilizing device.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, *72*(1), 1–19. [CrossRef]

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. [CrossRef]

Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, *96*(5), 885–897. [CrossRef]

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

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B*, *78*(3), 673–700. [CrossRef]

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

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y., & Yang, X. (2023). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics*, *41*(1), 97–110. [CrossRef]

Graham, B. S., Pinto, C. C. d. X., & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, *79*(3), 1053–1079. [CrossRef]

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

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46. [CrossRef]

Hájek, J. (1971). Discussion of "An essay on the logical foundations of survey sampling, Part I," by D. Basu. In V. P. Godambe, & D. A. Sprott (Eds.), *Foundations of statistical inference*. Holt, Rinehart and Winston.

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

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, *76*(1), 243–263. [CrossRef]

Li, J., & Matsushita, Y. (2025). A difference-in-differences estimator by covariate balancing propensity score. *arXiv*, arXiv:2508.02097. [CrossRef]

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. [CrossRef]

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle, & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. 4, pp. 2111–2245). Elsevier.

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

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244. [CrossRef]

Sant'Anna, P. H. C., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, *219*(1), 101–122. [CrossRef]

Tang, S., Zhan, M., Jiang, Q., & Zhang, T. (2024). Efficient covariate balancing for the average treatment effect with missing outcome. *Economics Letters*, *244*, 111961. [CrossRef]

Testa, L., Kennedy, E. H., & Reimherr, M. (2026). Efficient difference-in-differences estimation when outcomes are missing at random. *arXiv*, arXiv:2509.25009v2. Available online: https://arxiv.org/abs/2509.25009 (accessed on 14 September 2026). [CrossRef]

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

Wong, R. K. W., & Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, *105*(1), 199–213. [CrossRef]

Zhao, A., Ding, P., & Li, F. (2024). Covariate adjustment in randomized experiments with missing outcomes and covariates. *Biometrika*, *111*(4), 1413–1420. [CrossRef]

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, *47*(2), 965–993. [CrossRef]

Zhao, Q., & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, *5*(1), 20160010. [CrossRef]

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*(511), 910–922. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.