Design-Based Uncertainty for Quasi-Experiments*

Ashesh Rambachan† Jonathan Roth‡

November 21, 2022

Abstract

This paper develops a design-based theory of uncertainty that is suitable for analyzing quasi-experimental settings, such as difference-in-differences (DiD). A key feature of our framework is that each unit has an idiosyncratic treatment probability that is unknown to the researcher and may be related to the potential outcomes. We derive formulas for the bias of common estimators (including DiD), and provide conditions under which they are unbiased for an intrepretable causal estimand (e.g., analogs to the ATE or ATT). We further show that when the finite population is large, conventional standard errors are valid but typically conservative estimates of the variance of the estimator over the randomization distribution. An interesting feature of our framework is that conventional standard errors tend to become more conservative when treatment probabilities vary across units. This conservativeness helps to (partially) mitigate the undercoverage of confidence intervals when the estimator is biased. Thus, for example, confidence intervals for the DiD estimator can have correct coverage for the average treatment effect on the treated even if the parallel trends assumption does not hold exactly. We show that these dynamics can be important in simulations calibrated to real labor-market data. Our results also have implications for the appropriate level to cluster standard errors, and for the analysis of linear covariate adjustment and instrumental variables.

*We thank Alberto Abadie, Isaiah Andrews, Josh Angrist, Iavor Bojinov, Kirill Borusyak, Kevin Chen, Peng Ding, Avi Feller, Peter Hull, Chuck Manski, Evan Rose, Pedro Sant'Anna, Yotam Shem-Tov, Neil Shephard, Tymon Słoczyński, Chris Walker, Ruonan Xu, Davide Viviano, and seminar/conference participants at Brown, Ohio State, Cornell, Northwestern, Sciences Po, Toulouse School of Economics, SEA, NASMES, and the CEME Young Econometrician's Conference for helpful comments and suggestions. Rambachan gratefully acknowledges support from the NSF Graduate Research Fellowship under Grant DGE1745303.

†Microsoft Research New England. Email: arambachan@microsoft.com

‡Brown University. Email: jonathanroth@brown.edu

1 Introduction

Standard econometric approaches to inference are based on repeated sampling from an infinite super-population, yet this perspective may be unnatural in settings when the entire population of interest is observed, such as when the researcher has data on aggregate outcomes for all 50 U.S. states (Manski and Pepper, 2018). In such settings, it may be attractive to instead view uncertainty in the data as *design-based*, i.e. arising solely from the stochastic assignment of units to treatment (e.g. Abadie, Athey, Imbens and Wooldridge, 2020). A celebrated literature on design-based inference, dating to Neyman (1923) and Fisher (1935), has therefore received substantial recent attention in both statistics and econometrics.[1]

Existing work in the design-based literature has focused mainly on settings where treatment probabilities are known to the researcher, as in a randomized experiment (Imbens and Rubin, 2015), or where treatment probabilities are determined independently of potential outcomes, possibly conditional on some observable characteristics (Abadie et al., 2020; Xu, 2021; Abadie et al., 2022).[2] In these settings, it is typically possible to obtain unbiased estimates and (asymptotically) valid confidence intervals for causal estimands, such as the average treatment effect (ATE).

In practice, however, social scientists often study settings where the exact treatment assignment mechanism is unknown, and the assumption that treatment probabilities depend only on observable characteristics may be questionable. In these settings, it is common to instead adopt alternative, quasi-experimental strategies such as difference-in-differences (DiD) or instrumental variables (IV). From the super-population perspective, much attention has been paid to the conditions under which these estimators and their confidence intervals are valid for causal estimands — such as the average treatment effect on the treated (ATT) or local average treatment effect (LATE) — and the bias and undercoverage that result if these assumptions are violated.

In this paper, we develop a theory of uncertainty to analyze the properties of these quasi-experimental estimators from a design-based perspective. We introduce a design-based data-generating process where the treatment assignment is stochastic, but each unit has an idiosyncratic marginal probability π_i of receiving the treatment. This reflects the fact that in quasi-experimental contexts, researchers often argue that treatment status is determined by idiosyncratic factors — e.g. delays in the court system (Jackson, Johnson and Persico,

[1] See, for example, Imbens and Rubin (2015); Aronow and Middleton (2015); Li and Ding (2017); Savje and Delevoye (2020) in statistics and Abadie et al. (2020); Xu (2021); Bojinov, Rambachan and Shephard (2021); Roth and Sant'Anna (2021); Abadie, Athey, Imbens and Wooldridge (2022) in econometrics, among many others.

[2] See, also, Borusyak and Hull (2020), who study a setting where treatment is determined by non-experimentally-assigned shocks with a known distribution.

2016; Lafortune, Rothstein and Schanzenbach, 2018) or fluctuations in local weather patterns (Deryugina, Heutel, Miller, Molitor and Reif, 2019) — that might reasonably be thought of as stochastic. The repeated sampling thought experiment in our framework is then over the realization of these idiosyncratic factors that determine treatment, holding fixed the units in the population and their potential outcomes. We do not assume that the treatment probabilities π_i are known to the econometrician, and allow them to be related to the potential outcomes. This reflects that, for example, the researcher may not know the exact distribution of possible court delays, and the propensity to realize a court delay may be related to a state's potential outcomes.

We begin by analyzing the properties of the simple difference-in-means (SDIM) estimator under this DGP. This allows us to directly connect our results to existing work in the design-based literature, which has often focused on this estimator. Our results for the SDIM are also a useful building block for analyzing other estimators. Indeed, the DiD estimator can be viewed as an SDIM for a first-differenced outcome, and thus our results for the SDIM are immediately applicable to the DiD estimator. In Section 6, we show that our results for the SDIM can also be extended to regression adjustment with covariates – which can be viewed as an SDIM with a covariate-adjusted outcome – as well as the IV estimator, which can be viewed as the ratio of two SDIMs (for the reduced-form and first-stage).

In Section 3, we derive design-based analogs to the familiar omitted variables bias formula for the SDIM. These formulas imply that the SDIM will be unbiased for a design-based analog to the ATT if the idiosyncratic treatment probabilities π_i are orthogonal to the untreated potential outcomes $Y_i(0)$. The SDIM is further unbiased for the average treatment effect (ATE) if the π_i are orthogonal to $Y_i(1)$ as well. Since the DiD estimator can be viewed as an SDIM estimator for a first-differenced outcome, our results imply that the DiD estimator is unbiased for a design-based analog to the ATT under a design-based analog to parallel trends, which imposes that the π_i are orthogonal to trends in the untreated potential outcomes.

We then analyze the distribution of the SDIM and the properties of conventional CIs in Section 4. Our results imply that when the finite population is large, the SDIM is approximately normally distributed with a particular variance that depends on the finite-population variances of the potential outcomes and the treatment effects. More formally, we establish the approximate normality of the SDIM under finite-population asymptotics similar to those in Li and Ding (2017) and Abadie et al. (2020, 2022). We also establish a Berry-Esseen type result that bounds how far the distribution can deviate from normality in a finite population of fixed size. We then show that the usual heteroskedasticity-robust variance estimator for the SDIM is consistent for an upper bound on the variance of the SDIM estimator.

When the finite population is large, the usual t-based confidence intervals therefore yield

valid but potentially conservative inference for the expectation of the SDIM estimator (which corresponds with a causal estimand under the orthogonality conditions described above). Applying these results to the DiD estimator, we obtain that the usual cluster-robust standard errors (Bertrand, Duflo and Mullainathan, 2004) yield valid but potentially conservative inference for a design-based analog to the ATT under a design-based analog to the parallel trends assumption.

An interesting feature of our setting is that conventional standard errors tend to overstate the variance of the SDIM estimator when the treatment probabilities π_i are heterogeneous across units. For example, when treatment effects are constant, the conventional standard errors are strictly conservative when the π_i differ across units, except in knife-edge cases (see Corollary 4.1). Thus, conventional standard errors may overstate the variance of the SDIM from the design-based perspective even under constant treatment effects. This contrasts with the well-known result from Neyman (1923) for completely randomized experiments that conventional standard errors are strictly conservative if and only if treatment effects are heterogeneous.

An important implication of this variance conservativeness result is that conventional confidence intervals for the ATT or ATE need not necessarily undercover even when the SDIM is biased. Rather, there is a tradeoff between two forces: as the treatment probabilities π_i differ across units, this (i) may induce bias if the π_i covary with the potential outcomes, but (ii) induces the usual standard errors to become more conservative. Depending on which effect dominates, coverage of conventional confidence intervals can be either above or below the nominal level even when the estimator is biased for the ATT or ATE (see Proposition 4.5). Thus, for example, conventional confidence intervals for the DiD estimator can have correct coverage for the ATT under certain violations of the design-based analog to the parallel trends assumption.

We highlight these tradeoffs in Monte Carlo simulations in Section 5, where we consider DiD analyses of simulated treatments using state-level data from the LEHD, and allow the state-level π_i to depend on a state's voting results in the 2016 presidential election. Remarkably, for log employment as the outcome, we find that strengthening the relationship between the π_i and state-level voting patterns *increases* the coverage rate of conventional confidence intervals, despite the fact that doing so leads to more severe violations of the design-based analog to the parallel trends assumption. This is because the induced conservativeness of conventional confidence intervals dominates the bias from the violation of parallel trends. By contrast, when log earnings is the outcome, the bias effect dominates, and so the comparative static is reversed, although undercoverage is substantially less severe than it would be with a consistent estimate of the variance (e.g. coverage of 89% vs. 52% in one specification).

In Section 6, we present three extensions that illustrate the usefulness of our design-based framework. We first extend our results on inference to the setting of clustered treatment assignment where, for example, we observe individual-level data but treatment is determined at the regional level (e.g. states or counties). We show that when the number of regions in the finite population is large, the cluster-robust variance estimator at the region level is valid but potentially conservative from the design-based perspective. These results provide formal justification for the heuristic that in quasi-experimental settings, standard errors should be clustered at the level at which treatment assignment is determined. This echoes the recommendation in Abadie et al. (2022), although the variance calculations in that paper are for a setting in which units have the same probability of receiving treatment marginalized over a two-stage assignment process; thus treatment probabilities in Abadie et al. (2022) are not related to potential outcomes, and so their calculations are not directly applicable to quasi-experimental settings like DiD, where treatment probabilities may differ across units and the target parameter is typically the ATT rather than ATE.[3]

We next extend our results to analyze the properties of the ordinary least squares estimator that linearly adjusts for observed, pretreatment covariates. We show that the estimand of the OLS estimator equals a design-based analog to the ATT plus a bias term that now depends on the finite-population covariance between the treatment probabilities and a covariate-adjusted untreated potential outcome. We also characterize the estimand of the OLS estimator in terms of the relationship between the treatment probabilities and the covariates, nesting as a special case the known result that the OLS estimand provides a variance-weighted average of treatment effects when the propensity is linear in the covariates (see, e.g., Angrist (1998) in a super-population setting and Abadie et al. (2020) in a design-based setting). As before, conventional standard errors are valid but potentially conservative estimates of the variance of the estimator.

Finally, we show that our results can be used to analyze instrumental variables (IV) estimators from a design-based perspective, where the stochastic nature of the data arises from the assignment of the instrument. We derive an intuitive expression for the IV estimand allowing for an arbitrary relationship between the probability that $Z_i = 1$ and the potential outcomes. In the case where the instrument is completely randomly assigned, our expression reduces to a local average treatment effect (LATE), as in Angrist and Imbens (1994) and Angrist, Imbens and Rubin (1996) from the sampling perspective, and Kang, Peck and Keele (2018) from the design-based view. Our results imply, however, that the IV estimand has an interpretation as an instrument-propensity reweighted LATE under weaker orthogonality

[3]In contrast to our results, however, Abadie et al. (2022) allow for both sampling- and design-based uncertainty simultaneously, whereas we only consider the case of design-based uncertainty.

conditions that do not impose that the instrument be completely randomly assigned. Our results also imply that standard inference methods yield asymptotically conservative inference for this estimand in large finite populations where the instrument is strong, provided that standard errors are clustered at the level at which the instrument is determined.

2 Data-generating process

There is a finite population of N units. Each unit is associated with potential outcomes $Y_i(\cdot) := (Y_i(0), Y_i(1))$, which correspond with their outcomes under the control and treatment conditions. The observed outcome is $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, where $D_i \in \{0, 1\}$ denotes the treatment status of unit i. Both the N units and their collection of potential outcomes $Y(\cdot) := \{Y_i(\cdot) \colon i = 1, \ldots, N\}$ are fixed (or conditioned on).

The stochastic nature of the data arises from the vector of treatment assignments, $D = (D_1, ..., D_N)'$. To build intuition for the general assignment mechanism we analyze, first consider an important special case from Neyman (1923). Neyman considered a completely randomized experiment, where the number of treated units ($N_1 := \sum_i D_i$) is fixed, and any treatment assignment with N_1 treated units is assumed to be equally likely. If each unit is independently assigned to treatment according to $D_i \sim Bernoulli(\bar{p})$ for some constant $\bar{p} \in (0, 1)$, then the assignment mechanism studied by Neyman corresponds with the distribution of D conditional on N_1 units being treated.[4]

We consider a generalization, where each unit is independently assigned to treatment according to $D_i \sim Bernoulli(p_i)$, where p_i is an individual-specific probability of treatment that can potentially be arbitrarily related to the potential outcomes or other fixed covariates. This nests the special case where $p_i = \bar{p}$ for all i. We then analyze the distribution of D conditional on the number of treated units N_1 and the potential outcomes, yielding the following data-generating process:

$$\mathbb{P}\left(D = d \,\middle|\, \sum_i D_i = N_1, Y(\cdot)\right) \propto \prod_i p_i^{d_i}(1 - p_i)^{1-d_i} \tag{1}$$

for all $d \in \{0, 1\}^N$ such that $\sum_i d_i = N_1$, and zero otherwise.

The stochastic assignment of treatment in our model reflects the fact that in "quasi-experimental" settings, researchers often argue that treatment status is determined by idiosyncratic factors, e.g. delays in the court system. If we view these factors as stochastic, then each unit has some probability of being treated based on the realization of these id-

[4]Note that once we condition on N_1, D_i and D_j are no longer independent for $i \neq j$. The same is true under our generalization of Neyman's model described below.

iosyncratic factors. The individual-specific probabilities p_i corresponds with the probability that the idiosyncratic factors are such that unit i is treated (before conditioning on N_1).

We place no restrictions on how the p_i are related to the potential outcomes (or other fixed covariates), allowing for very rich forms of selection. For example, our data-generating process nests a Heckman (1976)-style selection model in which

$$D_i = 1\left[g(W_i, Y_i(1), Y_i(0)) + \epsilon_i \geqslant 0\right],$$

where W_i are fixed individual characteristics and $g(\cdot)$ is a possibly unknown link function. The random variable ϵ_i is a stochastic idiosyncratic error (independent across i) that could correspond with preference shocks, expectational errors, or other exogenous choice shifters.[5] We would then have that $p_i = P(\epsilon_i \geqslant -g(W_i, Y_i(1), Y_i(0)) \mid Y_i(\cdot), W_i)$.

Remark 1 (Comparison to other design-based models)**.** The existing design-based literature has mainly focused on settings where the marginal probabilities of treatment are known to the researcher, as in randomized experiments (e.g., Imbens and Rubin, 2015; Li and Ding, 2017), or settings in which treatment probabilities are determined independently of potential outcomes (possibly conditional on some observable characteristics). For example, Section 3 of Abadie et al. (2020) provides asymptotic results for large finite populations in a setting where treatment probabilities can differ across units; however, for parameters with a causal interpretation, Abadie et al. (2020) require that treatment probabilities are a linear function of observable characteristics, whereas we allow the p_i to be arbitrary. See, also, Xu (2021) for an extension of these results to non-linear estimators.[6] ■

Remark 2 (Connection to sampling literature)**.** Hajek (1964) studied the problem of drawing a sample of size N_1 from a finite population of size N with unequal probabilities. He considered a data-generating process where D follows (1), where in his model $D_i = 1$ corresponds with the event that unit i is included in the sample (rather than treated). Hajek (1964) referred to this scheme as *rejective sampling*, and so by analogy we refer to treatment assignment following (1) as a *rejective assignment* mechanism. Many of our technical results exploit connections between rejective sampling and rejective assignment. ■

[5] The ϵ_i could also be thought of as a "trembling," as in the game theory literature on quantal response equilibrium (McKelvey and Palfrey, 1995). We thank Chuck Manski for noting this connection.

[6] The results in Abadie et al. (2020) and Xu (2021) allow for both sampling- and design-based uncertainty, whereas we focus on design-based uncertainty only. The setting in Section 3 of Abadie et al. (2020) also differs from ours in that it does not condition on the number of treated units (although Section 2 of Abadie et al. (2020) does condition on N_1). Abadie et al. (2022) consider a two-step process where cluster-level treatment probabilities are initially drawn from a distribution (independent of potential outcomes) and units are then assigned to treatment based on the probability for their cluster; see Section 6 for additional discussion.

Remark 3 (Conditioning on N_1)**.** We follow standard practice in the design-based literature in statistics and conduct our analysis conditional on the number of treated units N_1. As described in Pashley, Basse and Miratrix (2021), it is often desirable to conduct inference as-if the number of treated units is fixed (i.e. conditional on N_1) even if this is not guaranteed by the assignment mechanism. Intuitively, the precision of treatment effect estimates typically depends on the number of treated units — for example, we would expect less precise estimates if we have 1 treated unit and 99 untreated units than if there are 50 treated units and 50 controls — and so conditioning on N_1 yields inference more relevant to the observed data.[7] Conducting inference conditional on ancillary statistics has a long history in statistics and econometrics dating to at least Fisher (1959), who argued that doing so guarantees valid inference on "recognizable subset[s]" of the sample space.[8] ■

Notation. We refer to the distribution of D given in (1) as the "randomization distribution", and denote probabilities of random variables (i.e. functions of D) over the randomization distribution by $\mathbb{P}_R(\cdot) := \mathbb{P}(\cdot \mid \sum_i D_i = N_1, Y(\cdot))$. We define expectations and variances, $\mathbb{E}_R[\cdot]$ and $\mathbb{V}_R[\cdot]$, analogously. A particularly important definition will be

$$\pi_i := \mathbb{P}_R(D_i = 1),$$

which is the idiosyncratic (marginal) probability that $D_i = 1$ conditional on N_1 units being treated. We note that the π_i are a function of the unconditional probabilities p_i introduced above. When the finite population is large, Hajek (1964, Theorem 5) showed that the π_i are approximately equal to the p_i; for our results, however, it will typically be more useful to work with the idiosyncratic (marginal) probabilities π_i directly rather than the p_i.[9]

For non-stochastic weights w_i and a non-stochastic attribute X_i (such as a potential outcome), we define

$$\mathbb{E}_w[X_i] := \frac{1}{\sum_i w_i}\sum_i w_i X_i \text{ and } \mathbb{V}\text{ar}_w[X_i] := \frac{1}{\sum_i w_i}\sum_i w_i (X_i - \mathbb{E}_w[X_i])^2$$

to be the finite-population weighted expectation and variance respectively. The finite-population weighted covariance $\mathbb{C}\text{ov}_w[\cdot,\cdot]$ is defined analogously. It is worth emphasizing

[7] Pashley et al. (2021) show, for example, that a confidence interval that is valid unconditionally, but not conditional on N_1, will fail to be "bet-proof" in the sense considered by, e.g., Buehler (1959) and Müller and Norets (2016).

[8] In a Bernoulli randomized experiment with equal probabilities, the statistic N_1 is ancillary. In our generalization, N_1 is a *specific ancillary* (Basu, 1977), in the sense that its distribution depends on the p_i but not on the potential outcomes.

[9] Hajek imposes a normalization so that $\sum_i p_i = N_1$.

that $\mathbb{E}_w[X_i]$ is not the expectation of a random variable, but rather a weighted average of a fixed attribute over the finite population. So, for example, $\mathbb{E}_1[Y_i(0)] = \frac{1}{N}\sum_i Y_i(0)$ is the finite-population average untreated potential outcome. Finally, we denote by $N_0 := \sum_i (1 - D_i)$ the number of untreated units.

3 The SDIM estimator and its expectation

We begin by analyzing the properties of the simple difference in means (SDIM) estimator over the randomization distribution. By the SDIM, we mean

$$\hat{\tau} := \frac{1}{N_1}\sum_i D_i Y_i - \frac{1}{N_0}\sum_i (1 - D_i) Y_i, \tag{2}$$

which compares the mean outcome for the treated and untreated units in the data. The SDIM has been studied in detail in the context of completely randomized experiments, beginning with Neyman (1923), and so by focusing on the SDIM we nest many existing results in the design-based literature as special cases of our results. Our results are also relevant for other estimators commonly used in quasi-experimental contexts.

Special case: difference-in-differences. The commonly-used difference-in-differences (DiD) estimator can be cast as an SDIM for a first-differenced outcome. Specifically, suppose we have balanced panel data for two periods $t \in \{1, 2\}$.[10] Suppose that some units ($D_i = 1$) are treated beginning in period 2, whereas the remaining units ($D_i = 0$) are untreated in both periods. The observed outcome for unit i in period t is $Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$. The SDIM estimator for the first-differenced outcome $Y_i := Y_{i2} - Y_{i1}$ is then equivalent to the DiD estimator,

$$\hat{\tau}_{DiD} = \frac{1}{N_1}\sum_{i:D_i=1} (Y_{i2} - Y_{i1}) - \frac{1}{N_0}\sum_{i:D_i=0} (Y_{i2} - Y_{i1}).$$

Our results for the SDIM thus have immediate implications for the DiD estimator, and we return to this special case throughout the paper. ▲

In Section 6 below, we extend our results for the SDIM to ordinary least squares estimators that adjust for observable characteristics, as well as instrumental variable estimators.

[10] Appendix B.1 extends our results in the two-period running example to non-staggered DiD settings with multiple time periods.

3.1 Expectation of the SDIM

Our first result gives the expectation of the SDIM over the randomization distribution.

Proposition 3.1.

$$\mathbb{E}_R[\hat{\tau}] = \tau_{ATE} + \frac{N}{N_0}\mathbb{C}ov_1\left[\pi_i, Y_i(0)\right] + \frac{N}{N_1}\mathbb{C}ov_1\left[\pi_i, Y_i(1)\right] \tag{3}$$

$$= \tau_{EATT} + \frac{N}{N_0}\frac{N}{N_1}\mathbb{C}ov_1\left[\pi_i, Y_i(0)\right] \tag{4}$$

where, for $\tau_i = Y_i(1) - Y_i(0)$, $\tau_{ATE} = \frac{1}{N}\sum_i \tau_i$ *and* $\tau_{EATT} = \frac{1}{N_1}\sum_i \pi_i\tau_i = \mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i\tau_i\right]$.

The first line of Proposition 3.1 shows that the expectation of the SDIM is equal to the finite-population ATE, τ_{ATE}, plus a bias term that depends on the finite-population covariances between the idiosyncratic treatment probabilities π_i and the potential outcomes. The second line of Proposition 3.1 gives an alternative expression for $\mathbb{E}_R[\hat{\tau}]$ in terms of a design-based analog to the average treatment effect on the treated, which we refer to as the *expected* ATT (EATT). In particular, τ_{EATT} is the expected value of what Imbens (2004) and Sekhon and Shem-Tov (2020) refer to as the sample average treatment effect on the treated (SATT) — i.e., the average treatment effect for the treated units in the sample — where the expectation is taken over the stochastic realization of which units are treated. This is equivalent to a convex weighted average of the treatment effects τ_i, where the weights are proportional to the idiosyncratic treatment probabilities π_i.

From Proposition 3.1, it is immediate that the SDIM will be unbiased over the randomization distribution for the EATT if the finite-population covariance between idiosyncratic treatment probabilities π_i and the untreated potential outcomes $Y_i(0)$ is equal to zero, i.e. $\sum_i(\pi_i - \frac{N_1}{N})Y_i(0) = 0$. This is satisfied under complete randomization of the treatment, in which case $\pi_i \equiv \frac{N_1}{N}$. It can also be satisfied if the idiosyncratic treatment probabilities vary across units but in a way that is not systematically related to the untreated potential outcomes on average in the finite population. Proposition 3.1 further implies the SDIM will be unbiased for the finite-population ATE if the finite-population covariance between π_i and both potential outcomes is zero.

Special case: DiD (continued). Consider the DiD example introduced above. Suppose the "no-anticipation" assumption is satisfied, i.e. $Y_{i1}(0) = Y_{i1}(1)$, so that treatment status in period 2 has no impact on the outcome in period 1. Proposition 3.1 then implies that

$$\mathbb{E}_R\left[\hat{\tau}_{DiD}\right] = \underbrace{\frac{1}{N_1}\sum_i \pi_i \tau_{i2}}_{\tau_{EATT,2}} + \underbrace{\frac{N}{N_1}\frac{N}{N_0}\mathbb{C}\text{ov}_1\left[\pi_i, Y_{i2}(0) - Y_{i1}(0)\right]}_{\delta}, \tag{5}$$

where $\tau_{i2} = Y_{i2}(1) - Y_{i2}(0)$ is unit i's treatment effect in period 2. The previous display shows that the expectation of the DiD estimator is the sum of two terms. The first is a design-based analog to the ATT in period 2, $\tau_{EATT,2}$. The second term, δ, is proportional to the finite-population covariance between idiosyncratic treatment probabilities π_i and trends in the untreated potential outcomes. Thus, the DiD estimator is unbiased for τ_{EATT} under the assumption that $\delta = 0$, which can be viewed as a design-based analog to the parallel trends assumption — i.e., if idiosyncratic treatment probabilities π_i are uncorrelated in the finite-population with changes in potential outcomes $Y_{i2}(0) - Y_{i1}(0)$. ▲

Remark 4 (Sensitivity analysis)**.** The characterization of the SDIM's bias in (3) and (4) may be useful for conducting sensitivity analyses. For example, researchers could report how large $\mathbb{C}\text{ov}_1\left[\pi_i, Y_i(0)\right]$ would need to be to produce a bias of a magnitude large enough to change a particular conclusion (e.g. the EATT is positive). Such a sensitivity analysis is related to, but different from existing design-based sensitivity analyses. For example, Rosenbaum (1987, 2002, 2005) places bounds on the relative odds ratio of treatment between two units (i.e., $\frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)}$ for $i \neq j$) and examines the extent to which the relative odds ratio must vary across units such that we no longer reject a particular sharp (Fisher) null of interest. In contrast, (3) and (4) suggest a simple approach for examining how the bias of the SDIM for the average treatment effect (on the treated) varies with the finite population covariance between idiosyncratic treatment probabilities π_i and the potential outcomes.[11] Likewise, equation (5) could be used for sensitivity analysis or partial identification of the EATT in DiD designs, as in Manski and Pepper (2018) and Rambachan and Roth (Forthcoming). ■

4 Distribution of the SDIM

We next turn our attention to the behavior of $\hat{\tau}$ over the randomization distribution. We will show that when the finite population is large, $\hat{\tau}$ is approximately normally distributed with a particular variance, and that the usual variance estimator is a conservative estimator for this variance. Our results have immediate implications for the distribution of the DiD estimator; we discuss extensions to general regression estimators with covariates and

[11] This approach is related to Aronow and Lee (2013) and Miratrix, Wager and Zubizarreta (2018), who consider sensitivity analysis for the finite-population mean under unequal-probability sampling where the sampling probabilities (analogous to p_i) are restricted to an interval $[p_{lb}, p_{ub}]$.

instrumental variables in Section 6.

4.1 Connection to unequal probability sampling

To analyze the behavior of $\hat{\tau}$ over the randomization distribution, it will be useful to connect the problem of estimating treatment effects to that of sampling from a finite population with unequal probabilities, which was previously studied by Hajek (1964) (among others). Specifically, note that $\hat{\tau}$ may be re-written as

$$\hat{\tau} = \sum_i \frac{D_i}{\pi_i}(\pi_i \tilde{Y}_i) - \frac{1}{N_0}\sum_i Y_i(0), \tag{6}$$

where $\tilde{Y}_i := \frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)$.[12] The second term, $\frac{1}{N_0}\sum_i Y_i(0)$, is non-stochastic, and therefore does not affect the variance (or higher-order moments) of the distribution of $\hat{\tau}$. The first term, $\sum_i \frac{D_i}{\pi_i}(\pi_i \tilde{Y}_i)$, is a Horvitz-Thompson estimator for the population total $\sum_i(\pi_i \tilde{Y}_i)$ under what Hajek (1964) refers to as rejective sampling. We can therefore make use of results from Hajek (1964) on the distribution of the Horvitz-Thompson estimator under rejective sampling to analyze the behavior of the SDIM over the randomization distribution.

4.2 Variance of the SDIM

As described in Hajek (1964), the exact variance of $\hat{\tau}$ depends on the second-order treatment probabilities, $\pi_{ij} = \mathbb{P}_R(D_i = 1, D_j = 1)$, which in general are complicated functions of the p_i. Fortunately, simple approximations to the variance are available which become accurate when $\sum_i \mathbb{V}_R[D_i] = \sum_i \pi_i(1-\pi_i)$ is large — that is, when the sum of the variances of the individual treatment indicators is large. The approximation we derive for the variance should therefore be accurate when the finite population is large and the treatment probabilities π_i are not too close to 0 or 1 for all units.[13] We evaluate the quality of this approximation in more detail below (see Proposition 4.4 and the simulations in Section 5).

Proposition 4.1 (Variance of the SDIM)**.**

$$\mathbb{V}_R[\hat{\tau}][1 + o(1)] = C\left[\frac{1}{N_1}\mathbb{V}ar_{\tilde{\pi}}[Y_i(1)] + \frac{1}{N_0}\mathbb{V}ar_{\tilde{\pi}}[Y_i(0)] - \frac{1}{N}\mathbb{V}ar_{\tilde{\pi}}[\tau_i]\right], \tag{7}$$

[12] The theory that follows can accommodate the case where $\pi_i = 0$ for some i, if $\frac{D_i}{\pi_i}$ is defined to be 0 whenever $\pi_i = 0$.

[13] Under a strict overlap condition of the form $\pi_i \in [\eta, 1-\eta]$ for some $\eta > 0$ for all units i, we would have that $\sum_i \mathbb{V}_R[D_i] = O(N)$. However, our results remain valid even if strict overlap fails and π_i is arbitrarily close to 0 or 1 for some units.

where $o(1) \to 0$ *as* $\sum_i \pi_i(1-\pi_i) \to \infty$, $\tilde{\pi}_i := \pi_i(1-\pi_i)$, *and* $C := \frac{\frac{1}{N}\sum_{k=1}^{N} \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}} \leqslant 1$.

Proposition 4.1 shows that the variance of $\hat{\tau}$ depends on the weighted finite-population variances of the potential outcomes and the treatment effects, where unit i is weighted proportionally to the variance of their treatment status, $\mathbb{V}_R[D_i] = \pi_i(1-\pi_i)$. The leading constant term is less than or equal to one by Jensen's inequality, with equality when π_i is constant across units.[14] Thus, in the special case of a completely randomize experiment, the right-hand side of (7) reduces to $\left(\frac{1}{N_1}\mathbb{V}\text{ar}_1[Y_i(1)] + \frac{1}{N_0}\mathbb{V}\text{ar}_1[Y_i(0)] - \frac{1}{N}\mathbb{V}\text{ar}_1[\tau_i]\right)$, which matches Neyman (1923)'s formula for completely randomized experiments up to a degrees-of-freedom correction.[15]

4.3 Estimated variance of the SDIM

Let $\hat{s}^2$ be the heteroskedasticity-robust variance estimator for $\hat{\tau}$ if the units are assumed to be sampled independently from an infinite super-population. That is, $\hat{s}^2 = \frac{1}{N_1}\hat{s}_1^2 + \frac{1}{N_0}\hat{s}_0^2$, where

$$\hat{s}_1^2 := \frac{1}{N_1}\sum_i D_i(Y_i - \bar{Y}_1)^2, \qquad \hat{s}_0^2 := \frac{1}{N_0}\sum_i (1-D_i)(Y_i - \bar{Y}_0)^2,$$

and $\bar{Y}_1 := \frac{1}{N_1}\sum_i D_i Y_i$, $\bar{Y}_0 := \frac{1}{N_0}\sum_i (1-D_i)Y_i$. Our next result gives an approximate expression for the expectation of $\hat{s}^2$ over the randomization distribution, where again the approximation is good when $\sum_i \mathbb{V}_R[D_i]$ is large.

Lemma 4.1.

$$\mathbb{E}_R\left[\hat{s}^2\right](1+o(1)) = \frac{1}{N_1}\mathbb{V}ar_\pi[Y_i(1)] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}[Y_i(0)], \tag{8}$$

where $o(1)$ *is as defined in Proposition 4.1.*

4.4 Comparison of actual and estimated variance

How does the usual estimated variance $\hat{s}^2$ compare to the variance of the SDIM over the randomization distribution? The following result shows that $\hat{s}^2$ is a (weakly) conservative estimator of the variance of the SDIM, up to the approximation errors described above.

[14] Specifically, if X is uniformly distributed on $\{\pi_1, ..., \pi_N\}$ and $g(x) = x(1-x)$, then $E[g(X)] = \frac{1}{N}\sum_i \pi_i(1-\pi_i) \leqslant g(E[X]) = \frac{N_1}{N}\frac{N_0}{N}$.

[15] Note that $\mathbb{V}\text{ar}_1[Y_i(d)] = \frac{1}{N}\sum_i (Y_i(d) - \mathbb{E}_1[Y_i(d)])^2$, which differs from the finite population variance used in Neyman (1923) by the degrees-of-freedom correction factor $\frac{N}{N-1} = 1 + o(1)$.

Proposition 4.2. *Let* $\mathbb{V}_R^{approx}[\hat{\tau}]$ *denote the expression on the right-hand side of (7), and* $\mathbb{E}_R^{approx}[\hat{s}^2]$ *the expression on the right-hand side of (8). We have that*

$$\mathbb{E}_R^{approx}\left[\hat{s}^2\right] \geqslant \mathbb{V}_R^{approx}\left[\hat{\tau}\right].$$

Moreover, the inequality holds with equality if and only if

$$Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right] = \frac{(1-\pi_i)/\pi_i}{N_0/N_1}\left(Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]\right) \text{ for all } i. \tag{9}$$

In the case of a completely randomized experiment ($\pi_i \equiv \frac{N_1}{N}$), equation (9) is satisfied if and only if treatment effects are constant, and thus Proposition 4.2 nests the well-known result from Neyman (1923) that in a completely randomized experiment, the usual variance estimator is weakly conservative, and is strictly conservative if and only if there are heterogeneous treatment effects (i.e. $\mathbb{V}\text{ar}_1\left[\tau_i\right] > 0$).

Interestingly, Proposition 4.2 implies that $\hat{s}^2$ will generally be strictly conservative when the π_i differ across units, except in knife-edge cases. For example, the following corollary shows that when treatment effects are constant and the SDIM is unbiased, the estimated variance is strictly conservative if $\pi_i \neq \frac{N_1}{N}$ for any unit i with $Y_i(0) \neq \mathbb{E}_{1-\pi}\left[Y_i(0)\right]$.[16]

Corollary 4.1. *If treatment effects are constant, i.e.* $Y_i(1) = \tau + Y_i(0)$ *for all* i*, and* $\mathbb{E}_R\left[\hat{\tau}\right] = \tau$*, then the inequality in Proposition 4.2 holds with equality if and only if* $\pi_i = \frac{N_1}{N}$ *for all* i *such that* $Y_i(0) \neq \mathbb{E}_{1-\pi}\left[Y_i(0)\right]$.

Thus, in contrast to the special case of a completely randomized experiment, the usual variance estimator may be conservative even under constant treatment effects. To develop intuition for this result, consider two data-generating processes (DGPs) where half of the units are treated. In the first DGP, we have a completely randomized experiment so that $\pi_i = 0.5$ for all i, whereas in the second DGP $\pi_i = 0.1$ for half of the units and 0.9 for the remaining half. In the first DGP, $\mathbb{V}_R\left[D_i\right] = 0.5^2$ for all i, whereas in the second DGP, $\mathbb{V}_R\left[D_i\right] = 0.1 \cdot 0.9 < 0.5^2$ for all i. Hence, the variance of the treatment indicators is strictly smaller in the second experiment. It is thus intuitive that the variance of $\hat{\tau}$ should be smaller under the second DGP, since the stochastic nature of the data only arises from the distribution of D_i, and D_i has strictly smaller variance for all i in the latter DGP. Since

[16] If treatment effects are not constant, then it is possible that the estimated variance is non-conservative with heterogeneous π_i. However, this requires the knife-edge scenario where equation (9) holds for all i, i.e. for all units, the distance between $Y_i(1)$ and its finite-population mean is exactly equal to the product of a term capturing the deviation of π_i from $\frac{N_1}{N}$ (i.e. $\frac{(1-\pi_i)/\pi_i}{N_1/N_0}$) and the deviation of $Y_i(0)$ from its finite-population mean.

$\hat{s}^2$ is non-conservative for the variance of $\hat{\tau}$ under constant treatment effects in the first experiment, it will therefore tend to over-estimate the variance in the second experiment.

The proof of Proposition 4.2 suggests that the conservativeness of $\hat{s}^2$ will tend to be larger when there is more heterogeneity in π_i. For example, under the setting in Corollary 4.1, $\mathbb{E}_R^{approx}[\hat{s}^2] - \mathbb{V}_R^{approx}[\hat{\tau}]$ is bounded below by a term proportional to $\mathbb{V}\text{ar}_1[(\pi_i - \frac{N_1}{N}) \cdot (Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)])]$. Thus, $\hat{s}^2$ will tend to be quite conservative when the heterogeneity in π_i is large, especially if $\pi_i - \frac{N_1}{N}$ is large for units with extreme values of $Y_i(0)$. The fact that conventional variance estimates tend to become more conservative when the π_i are more heterogeneous has important implications for the coverage of conventional confidence intervals, as we explore in our Monte Carlo simulations in Section 5 below.

Special case: DiD (continued) In the running DiD example, the variance estimator $\hat{s}^2$ is equivalent to the cluster-robust (at the unit level) variance estimator for $\hat{\tau}_{DiD}$ from the OLS panel regression

$$Y_{it} = \alpha_i + \lambda_t + D_i \cdot 1[t = 2]\tau_{DiD} + \epsilon_{it}. \tag{10}$$

Therefore, Proposition 4.2 implies that the cluster-robust variance estimator for $\hat{\tau}_{DiD}$ is weakly conservative for the variance of the DiD estimator over the randomization distribution. ▲

4.5 Normality and Variance Consistency

Our results so far imply that the typical variance estimator will be conservative in the sense that its expectation is weakly larger than the true variance of $\hat{\tau}$ (up to an $o(1)$ approximation error). This suggests that standard confidence intervals based on $\hat{s}$ will be conservative for $\mathbb{E}_R[\hat{\tau}]$ if (i) $\hat{\tau}$ is approximately normally distributed, and (ii) $\hat{s}^2$ is close to its expectation with high probability. Our next results show that both will be true in large finite populations satisfying certain regularity conditions.

To formalize this intuition, we follow Hajek (1964) for sampling from a finite population and Freedman (2008), Lin (2013), and Li and Ding (2017) for randomized experiments, and consider a sequence of finite populations of increasing size. More precisely, we consider sequences of finite populations indexed by m of size N_m, with N_{1m} treated units, potential outcomes $\{Y_{im}(\cdot) : i = 1, ..., N_m\}$, and assignment probabilities $\pi_{1m}, ..., \pi_{N_m}$. For brevity, we leave the subscript m implicit in our notation; all limits are implicitly taken as $m \to \infty$. We then establish a central limit theorem (CLT) and variance consistency result under mild regularity conditions on the sequence of finite populations. These results provide an approximation to the properties of $\hat{\tau}$ for finite populations with a sufficiently large number

of units. Indeed, as we show in Proposition 4.4 below, these asymptotic results translate to Berry-Esseen type bounds on the approximation quality of the CLT in any finite population of fixed size.

We impose the following assumptions on the sequence of populations.

Assumption 4.1. *The sequence of populations satisfies* $\sum_{i=1}^{N} \pi_i(1-\pi_i) \to \infty$.

Recall that $\pi_i(1-\pi_i)$ is the variance of the Bernoulli random variable D_i, so Assumption 4.1 implies that the sum of the variances of the D_i grows large. Assumption 4.1 also implies that both N_1 and N_0 go to infinity, since $\sum_{i=1}^{N} \pi_i(1-\pi_i) \leqslant \min\{\sum_i \pi_i, \sum_i (1-\pi_i)\} = \min\{N_1, N_0\}$.

Our next assumption is similar to the condition for the Lindeberg central limit theorem, and imposes that the weighted finite-population variance of $\tilde{Y}_i$ is not dominated by a small number of observations.

Assumption 4.2. *Let* $\tilde{Y}_i = \frac{1}{N_1} Y_i(1) + \frac{1}{N_0} Y_i(0)$, *and assume* $\sigma^2_{\tilde{\pi}} = \mathbb{V}ar_{\tilde{\pi}}\left[\tilde{Y}_i\right] > 0$. *Suppose that for all* $\epsilon > 0$,

$$\frac{1}{\sigma^2_{\tilde{\pi}}} \mathbb{E}_{\tilde{\pi}}\left[\left(\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right]\right)^2 1\left[\left|\tilde{Y}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right]\right| > \sqrt{\sum_i \pi_i(1-\pi_i)} \cdot \sigma_{\tilde{\pi}} \epsilon\right]\right] \to 0.$$

Our final assumption bounds the influence that any single observation has on the π- and $(1-\pi)$-weighted variances of the potential outcomes. This generalizes the assumptions in Theorem 1 in Li and Ding (2017), which establishes consistency of the Neyman variance under equal-probability sampling from a finite population.

Assumption 4.3. *Define* $m_N(1) := \max_{1 \leqslant i \leqslant N} \left(Y_i(1) - \mathbb{E}_{\pi}\left[Y_i(1)\right]\right)^2$, *and analogously* $m_N(0) := \max_{1 \leqslant i \leqslant N} \left(Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]\right)^2$. *Assume that,*

$$\frac{1}{N_1} \frac{m_N(1)}{\mathbb{V}ar_{\pi}\left[Y_i(1)\right]} \to 0 \text{ and } \frac{1}{N_0} \frac{m_N(0)}{\mathbb{V}ar_{1-\pi}\left[Y_i(0)\right]} \to 0.$$

Central limit theorem and variance consistency. Under the conditions introduced above, we can formally establish a CLT and variance consistency result.

Proposition 4.3. *Suppose Assumptions 4.1 and 4.2 hold. Then,*

$$\frac{\hat{\tau} - \mathbb{E}_R\left[\hat{\tau}\right]}{\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}} \xrightarrow{d} \mathcal{N}\left(0,\, 1\right).$$

Further, under Assumptions 4.1 and 4.3,

$$\frac{\hat{s}^2}{\frac{1}{N_1}\mathbb{V}ar_\pi\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}\left[Y_i(0)\right]} \xrightarrow{p} 1.$$

Non-asymptotic bounds. In addition to the asymptotic results shown above, we can also obtain Berry-Esseen type bounds on the quality of the normal approximation (using the approximate variance $\mathbb{V}_R^{approx}\left[\hat{\tau}\right]$) for a fixed finite population.

Proposition 4.4. *Let b_1, b_2 be positive constants, and define $t = (\hat{\tau} - \mathbb{E}_R\left[\hat{\tau}\right])/\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}$. Then there exist constants k and $\bar{N}$ such that*

$$\sup_y |P(t \leqslant y) - \Phi(y)| \leqslant \frac{k}{\sqrt{N}}$$

for any finite population of size $N \geqslant \bar{N}$ such that $\mathbb{V}_R^{approx}\left[\hat{\tau}\right] = Nb_1$ and $\mathbb{E}_1\left[\left(\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right)^4\right] < b_2$.

Proposition 4.4 is attractive in the sense that it shows that the distribution of $\hat{\tau}$ will be approximately normally distributed in finite populations that are sufficiently large (relative to the fourth moment of the potential outcomes), without appealing to arguments involving a sequence of finite populations of increasing size.

4.6 Implications for coverage of confidence intervals

The results in the previous subsection allow us to formalize the conditions under which confidence intervals of the form $\hat{\tau} \pm 1.96 \cdot \hat{s}$ will be valid for τ_{EATT} (or τ_{ATE}) when the finite population is large.

Proposition 4.5. *Suppose Assumptions 4.1-4.3 hold, and that*

(i) $\frac{b}{\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}} \to b^* \in \mathbb{R}$, *where* $b = \frac{N}{N_1}\frac{N}{N_0}\mathbb{C}ov_1\left[\pi_i, Y_i(0)\right]$ *is the bias of $\hat{\tau}$ for the EATT.*

(ii) $\sqrt{\frac{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}{\mathbb{E}_R^{approx}\left[\hat{s}^2\right]}} \to r \in (0, 1]$.

Then,

$$\frac{\hat{\tau} - \tau_{EATT}}{\hat{s}} \xrightarrow{d} \mathcal{N}\left(b^* \cdot r,\, r^2\right),$$

and the confidence interval $\hat{\tau} \pm 1.96 \cdot \hat{s}$ has asymptotic coverage for τ_{EATT} approaching

$$\Phi\left(\frac{1.96}{r} - b^*\right) - \Phi\left(\frac{-1.96}{r} - b^*\right). \tag{11}$$

The analogous result holds for τ_{ATE}, replacing b with $\frac{N}{N_1}\mathbb{C}ov_1\left[\pi_i, Y_i(1)\right] + \frac{N}{N_0}\mathbb{C}ov_1\left[\pi_i, Y_i(0)\right]$.

Part (i) of Proposition 4.5 imposes that the sequence of finite populations is such that the bias of $\hat{\tau}$ is of the same order of magnitude as its standard deviation over the randomization distribution (i.e. local to zero). Part (ii) of the proposition imposes that the conservativeness of the typical variance estimator stabilizes asymptotically (recall that $\mathbb{E}_R^{approx}\left[\hat{s}^2\right] \geqslant \mathbb{V}_R^{approx}\left[\hat{s}^2\right]$ by Proposition 4.2).

When $\hat{\tau}$ is unbiased, so that $b^* = 0$, Proposition 4.5 shows that typical confidence intervals will have correct but generally conservative coverage. Indeed, coverage will be strictly above the nominal level when the variance estimator is strictly conservative, i.e. when $r < 1$, as will typically be the case when the π_i are heterogeneous (see Section 4.4). Thus, in the running DiD example, confidence intervals based on cluster-robust standard-errors for the OLS specification (10) will have asymptotically correct but typically conservative coverage for $\tau_{EATT,2}$ under the design-based analog to the parallel trends assumption ($\delta = 0$) discussed in Section 3.

Proposition 4.5 also implies that conventional confidence intervals will have correct coverage when the bias of $\hat{\tau}$ is sufficiently small relative to the conservativeness of the variance estimator. Specifically, since the expression for coverage in (11) is continuous in b^* and is strictly above the nominal level when $r < 1$ and $b^* = 0$, it follows that coverage will still be correct when b^* is non-zero but sufficiently small. A sufficient condition to ensure at least 95% coverage is that $|b^*| \leqslant 1.96\left(\frac{1}{r} - 1\right)$. Thus, we see that conventional confidence intervals can accommodate some bias owing to the fact that heterogeneity in treatment probabilities π_i or heterogeneous effects τ_i typically induces conservativeness of the variance estimator.

5 Monte Carlo simulations

In this section, we conduct Monte Carlo simulations based on the Quarterly Workforce Indicators (QWI) from the Longitudinal Household-Employer Dynamics (LEHD) Program at the U.S. Census (United States Bureau of the Census, 2022). These simulations allow us to illustrate our main results and investigate the quality of the asymptotic approximations in Section 4 in a realistic empirical setting. The QWI provides aggregate statistics from the LEHD linked employer-employee microdata, which covers over 95% of U.S. private sector jobs. It is thus natural to view the uncertainty in policy analyses using the QWI as arising from the stochastic realization of treatment status rather than sampling from an infinite

super-population.[17]

Simulation design: Our simulation design mimics a state-level DiD analysis. We use aggregate data on the 50 U.S. states and Washington D.C. from the QWI (indexed by $i = 1, ..., N$) for the years 2012 and 2016 (indexed by $t = 1, 2$).[18] For each state and year, we set the potential outcomes $Y_{it}(1)$ and $Y_{it}(0)$ equal to the state's observed outcome in the QWI (Y_{it}). This imposes that our simulated treatments have no effect for any state, and so $\tau_{EATT,2} = \tau_{ATE,2} = 0$. We discuss extensions that incorporate heterogeneous treatment effects below. In line with the design-based model described above, the potential outcomes are held fixed throughout our simulations; the simulation draws differ in that each corresponds with a different realization of the generated placebo laws $D = (D_1, ..., D_N)'$. We simulate D from the rejective assignment mechanism (1). For each draw of the assignment vector, we calculate the DiD estimator $\hat{\tau}_{DiD}$ and a nominal 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96 \cdot \hat{s}$, where $\hat{s}$ is the heteroskedasticity-robust standard error for the first-differenced outcome (equivalently, the cluster-robust standard error for specification (10)).

To simulate D from the rejective assignment mechanism (1), we draw $D_1, \ldots, D_N$ as independent Bernoulli random variables with (unconditional) state-level treatment probabilities p_i, discarding any draws where $\sum_i D_i \neq N_1$. The state-level treatment probabilities p_i are chosen such that, for some $p^1 \in [0, 1]$, states that voted for Clinton in the 2016 presidential election have $p_i = p^1$, and states that voted for Trump have $p_i = 1 - p^1$.[19] Thus, when $p^1 = 0.5$, all states have the same probability of adopting treatment, as in a randomized experiment, whereas when $p^1 > 0.5$, Democratic states are more likely to adopt the treatment. We report results as p^1 varies over $p^1 \in \{0.50, 0.75, 0.90\}$ and fix the number of treated and untreated states at $N_1 = 25$ and $N_0 = 26$, respectively (Washington D.C. is included in the data).

We report results for two choices of the outcome Y_{it}. The first outcome is when Y_{it} corresponds with the log employment level for state i in period t. The second is when Y_{it} is the log of state-level average monthly earnings for state i in year t.[20]

Simulation results: We first report the bias of the DiD estimator. While the placebo law has no treatment effect for any state, the change in untreated potential outcomes

[17] The LEHD program even writes, "Because the estimates are not derived from a probability-based sample, no sampling error measures are applicable" (United States Bureau of the Census, 2022).

[18] Specifically, we use the QWI data for the first quarter of each of these years.

[19] We collect the state-level results from the 2016 presidential election from the MIT Election Data and Science Lab (MIT Election Data and Science Lab, 2022).

[20] Specifically, this is the log of `earnS` in the QWI, which corresponds with the average monthly earnings of individuals employed at the same firm throughout the relevant quarter.

$Y_{i2}(0) - Y_{i1}(0)$ varies across states in a way that is related to state-level voting patterns in the 2016 presidential election. As a result, the design-based parallel trends assumption, $\mathbb{C}\text{ov}_1\left[\pi_i, Y_{i2}(0) - Y_{i1}(0)\right] = 0$, is violated when $p^1 \neq 0.5$, and hence the DiD estimator is biased for the EATT over the randomization distribution in these simulations. The first row of Table 1 reports the normalized bias of the DiD estimator (i.e., $\mathbb{E}_R\left[\hat{\tau}_{DiD}\right] / \sqrt{\mathbb{V}\text{ar}_R\left[\hat{\tau}_{DiD}\right]}$) as p^1 varies for both of these two outcomes. Recall that $p^1 = 0.5$ corresponds with the case of a completely randomized experiment, and thus the bias is 0 (up to simulation error) in this case. The magnitude of the bias increases as we increase p^1, since the average value of $Y_{i2}(0) - Y_{i1}(0)$ differs between Democratic and Republican states for both of our outcomes. Note that the magnitude of the bias (relative to the standard deviation of the DiD estimator) is smaller for the log employment outcome than for the log earnings outcome.

	p_1		
	0.50	0.75	0.90
Normalized bias	0.013	0.250	0.525
Variance conservativeness	0.976	1.315	2.303
Coverage	0.939	0.967	0.991
Oracle coverage	0.949	0.943	0.917

(a) Log employment

	p_1		
	0.50	0.75	0.90
Normalized bias	0.004	0.882	1.871
Variance conservativeness	0.987	1.383	2.541
Coverage	0.944	0.917	0.888
Oracle coverage	0.952	0.854	0.516

(b) Log earnings

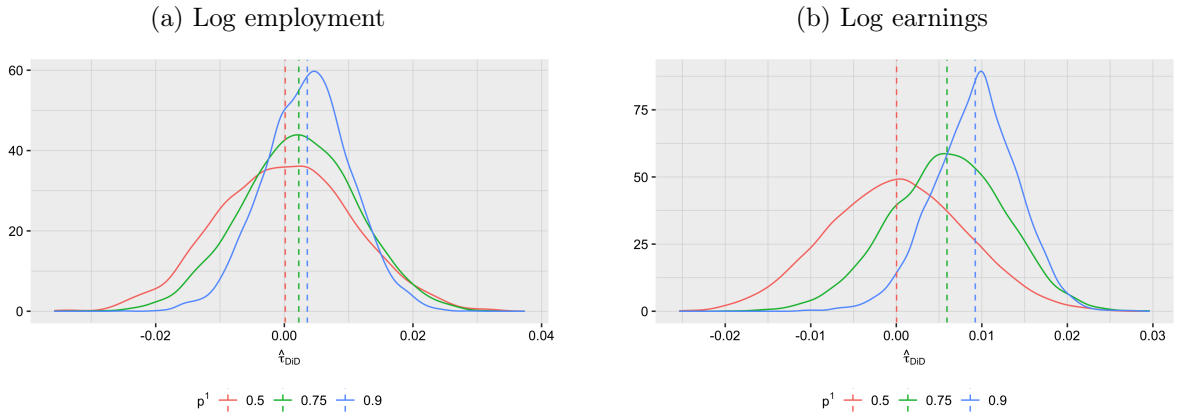
Table 1: Normalized bias, variance conservativeness, and coverage in Monte Carlo simulations.

Notes: Row 1 reports the normalized bias of the DiD estimator ($\mathbb{E}_R\left[\hat{\tau}_{DiD}\right] / \sqrt{\mathbb{V}\text{ar}_R\left[\hat{\tau}_{DiD}\right]}$) for the EATT over the randomization distribution. Row 2 reports the estimated ratio $\frac{\mathbb{E}_R\left[\hat{s}^2\right]}{\mathbb{V}\text{ar}_R\left[\hat{\tau}_{DiD}\right]}$ across simulations, which measures the conservativeness of the heteroskedasticity-robust variance estimator. Row 3 reports the coverage rate of a nominal 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96\,\hat{s}$. The columns report results as the idiosyncratic treatment probability for Democratic states, p^1, varies over $\{0.5, 0.75, 0.9\}$. Row 4 reports the coverage rate of an "oracle" 95% confidence interval that uses the true variance rather than an estimated one, $\hat{\tau}_{DiD} \pm 1.96 \sqrt{\mathbb{V}_R\left[\hat{\tau}_{DiD}\right]}$. The results are computed over 5,000 simulations with $N_1 = 25$.

Figure 1 plots the distribution of the DiD estimator over the randomization distribution. The figure illustrates that as we increase p^1, (i) the DiD estimator becomes more biased, but (ii) the distribution of the DiD estimator is less-dispersed, which leads the usual variance estimator to be conservative. The figure also shows that the distributions are approximately normally distributed, illustrating the CLT / Berry-Esseen result from Section 4.

The conservativeness of the usual heteroskedasticity-robust variance estimator is captured more directly in the second row of Table 1, which shows the ratio of the average estimated variance for $\hat{\tau}$ to the actual variance of the estimator, $\frac{\mathbb{E}_R\left[\hat{s}^2\right]}{\mathbb{V}\text{ar}_R[\hat{\tau}]}$. Recall that Proposition 4.2 and

Figure 1: Behavior of DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution.



Notes: This figure plots the behavior of the DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution. The idiosyncratic treatment probability for Democratic states, p^1, varies over $\{0.5, 0.75, 0.9\}$ (colors), holding fixed the number of treated units $N_1 = 25$. The results are computed over 5,000 simulations. The vertical dashed lines show the mean of the t-statistic for the relevant parameter values.

Corollary 4.1 established that $\hat{s}^2$ will typically be conservative for the true variance of the DiD estimator over the randomization distribution (in the sense that $\frac{\mathbb{E}_R^{approx}[\hat{s}^2]}{\mathbb{V}_R^{approx}[\hat{\tau}]} > 1$) when there is heterogeneity in the idiosyncratic treatment probabilities. Indeed, this is exactly what we observe. For simulations with $p^1 = 0.5$ (i.e., no heterogeneity in idiosyncratic treatment probabilities), $\hat{s}^2$ is, on average, approximately equal to the true variance of the DiD estimator. As p^1 increases, however, it becomes more conservative — in the most extreme case when $p^1 = 0.9$, the average estimated variance is approximately 2.5 times as large as the true variance. Recall that in our baseline specification, treatment effects are zero for all units, and thus this conservativeness is the result of heterogeneity in the π_i rather than in treatment effects.

The third row of Table 1 reports the coverage of a standard 95% confidence interval (i.e., the fraction of simulations in which the confidence interval covers the true EATT of zero). For the case with $p^1 = 0.5$, which corresponds with a completely randomized experiment, the standard confidence intervals have approximately 95% coverage for both outcomes (up to simulation error). As we increase p^1, there is a tradeoff between the fact that the estimator is biased (which leads to lower coverage) and the fact that the variance estimator is conservative (which leads to higher coverage), as formalized in Proposition 4.5. For log earnings as the outcome, the bias dominates and coverage decreases in p^1 — coverage of the EATT is only about 88.8% when $p^1 = 0.9$. By contrast, for the state-level log average employment outcome, the bias is smaller, and so the conservativeness of the variance estimator dominates. Remarkably, when $p^1 = 0.9$, the coverage rate is 99.1% owing to the conservativeness of the

variance estimator, despite the fact that the design-based analog to parallel trends does not hold exactly.

Finally, the last row of Table 1 reports the coverage of an "oracle" 95% confidence interval that uses the true variance of the DiD estimator over the randomization distribution instead of the estimated variance $\hat{s}^2$, which enables us to examine the impact of the conservative variance estimator on coverage. When $p^1 = 0.9$ for log-earnings, for example, coverage would be only 51.6% using the oracle variance, but is 88.8% using the conventional conservative variance estimator. The conservativeness of the variance estimator thus greatly mitigates the bias induced by the heterogeneity in π_i in this example.

Extensions: Appendix Section D presents several extensions to these simulations. We consider simulation designs that vary the number of treated units, with similar results. We also consider designs with treatment effect heterogeneity, which we find leads conventional confidence intervals to be even more conservative. Finally, we consider designs with varying population sizes, and find that the normal approximation works fairly well with as few as 26 states, but becomes less accurate with only 10.

6 Extensions

In this section, we develop three extensions that illustrate the usefulness of our design-based framework. First, we consider settings where treatment is assigned at the cluster level, and show that the cluster-robust variance estimator (clustered at the level at which treatment is assigned) is valid but potentially conservative from the design-based perspective when the number of clusters is large. The heteroskedasticity-robust variance estimator, in contrast, can now be invalid. Second, we consider conditions under which linear covariate adjustment can address the bias of the SDIM estimator derived in Section 3, providing two characterizations of the covariate-adjusted difference-in-means estimand: one in terms of the EATT plus a bias that depends on the finite-population covariance between the treatment probabilities and a covariate-adjusted untreated potential outcome, and another directly in terms of the relationship between the treatment probabilities and the covariates. Third, we analyze instrumental variable estimators, where the stochastic nature of the data now arises from the assignment of the instrument, and show that the IV estimand can has a causal interpretation as an instrument-propensity weighted LATE under orthogonality conditions that are weaker than complete randomly assignment of the instrument.

6.1 Clustered treatment assignment

Suppose each unit $i = 1, \ldots, N$ belongs to one of C clusters, where $c(i)$ denotes the cluster membership of unit i and N_c is the number of units in cluster c. We assume treatment is assigned at the cluster level, where the cluster level treatment assignments $D := (D_1, \ldots, D_C)'$ follow a rejective assignment mechanism. For example, units i may be individuals living in states $c(i)$, while policy is determined at the state level. Formally, letting $C_1 := \sum_c D_c$ and $C_0 := \sum_c (1 - D_c)$ denote the number of treated and untreated clusters respectively, we assume the clustered treatment assignments follows the data-generating process

$$\mathbb{P}\left(D = d \mid \sum_c D_c = N_1^C, Y(\cdot) \right) \propto \prod_c p_c^{d_c} (1 - p_c)^{d_c}. \tag{12}$$

Let $\pi_c := \mathbb{P}_R(D_c = 1)$ denote the marginal treatment probability for cluster c under (12). Let $D_i = D_{c(i)}$ denote unit i's treatment assignment. The total number of treated units $N_1 = \sum_i D_i$ is now stochastic as the number of units varies across clusters.

Suppose the researcher estimates the SDIM estimator $\hat{\tau}$ based on individual-level data on outcomes and treatments. We analyze the behavior of $\hat{\tau}$ over the randomization distribution of the clustered treatment assignments (12). Since the regularity conditions for many of our results are natural extensions of those in Section 4 to the clustered design, we defer them to Appendix C in the interest of brevity. Appendix C also provides more general results that apply to any OLS estimator, which nests the SDIM studied in the main text as a special case.

Our first result describes the distribution of $\hat{\tau}$ under finite-population asymptotics where the number of clusters grows large, analogous to those in Section 4.

Proposition 6.1.

1. *If Assumption C.1(i)-(iii) holds with* $X_i(d) = (1, d)'$, *and* $\sum_c \pi_c(1 - \pi_c) \to \infty$, *then* $\hat{\tau} - (\tau_{EATT}^{cluster} + \delta_{cluster}) \xrightarrow{p} 0$, *where*

$$\tau_{EATT}^{cluster} = \mathbb{E}_{\pi_{c(i)}}[\tau_i] \text{ and } \delta_{cluster} = \frac{N}{N - \sum_i \pi_{c(i)}} \frac{N}{\sum_i \pi_{c(i)}} \mathbb{C}ov_1\left[\pi_{c(i)}, Y_i(0)\right].$$

2. *If Assumption C.1 holds with* $X_i(d) = (1, d)'$, *and* $\sum_c \pi_c(1 - \pi_c) \to \infty$, *then*

$$\frac{\sqrt{C}(\hat{\tau} - \tau_{EATT}^{cluster} - \delta_{cluster})}{\sqrt{\Omega_{cluster}(2, 2)}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\Omega_{cluster}(2,2)$ *is the* $(2,2)$*-th element of the matrix*

$$\Omega_{cluster} := \mathbb{E}_R\left[\frac{1}{C}\sum_i X_i X_i'\right]^{-1} V_{cluster}\ \mathbb{E}_R\left[\frac{1}{C}\sum_i X_i X_i'\right]^{-1},$$

for $X_i := (1, D_i)'$ *and* $V_{cluster}$ *as defined in Proposition C.1.*

The first part of the proposition shows that $\hat{\tau}$ is consistent for $\tau_{EATT}^{cluster} + \delta_{cluster}$, where $\tau_{EATT}^{cluster}$ is an analog to the EATT (i.e. a weighted average of τ_i, with weights proportional to the probability that an individual's cluster is treated), and $\delta_{cluster}$ is a bias term related to the covariance between treatment probabilities and potential outcomes. The second part of the theorem shows that $\hat{\tau}$ is also asymptotically normally distributed as the number of clusters grows large.

If all of the clusters are the same size (i.e. N_c is constant), then analogous to the results in Proposition 3.1, we have that $\mathbb{E}_R[\hat{\tau}] = \tau_{EATT}^{cluster} + \delta_{cluster}$. However, if cluster sizes vary across clusters, then the total number of treated units (i.e. the denominator in the SDIM) is now stochastic, and thus it need not be the case that $\hat{\tau}$ is exactly unbiased for $\tau_{cluster}^{EATT} + \delta_{cluster}$, although Proposition 6.1 shows that it will be consistent as the number of clusters grows large. This is analogous to well-known results for the OLS estimator from the sampling-based perspective, where OLS is unbiased only if one conditions on the design-matrix but is consistent when the covariates are stochastic.

Cluster robust variance estimator: Let $\hat{\Omega}_{cluster}$ be the cluster-robust variance estimator (Liang and Zeger, 1986) for the coefficients from the regression of Y_i on $X_i = (1, D_i)'$ (see Appendix C for a formal definition). Our next result establishes that the cluster-robust variance estimator is weakly conservative for the true variance of the SDIM estimator over the clustered treatment assignment mechanism in finite-populations with a large number of populations.

Proposition 6.2. *If Assumption C.1(i)-(iii) and Assumption C.2 hold with* $X_i(d) = (1, d)'$*, and* $\sum_c \pi_c(1-\pi_c) \to \infty$*, then* $\hat{\Omega}_{cluster} - \Omega_{cluster}^{est} \xrightarrow{p} 0$*, for a matrix* $\Omega_{cluster}^{est}$ *such that* $\Omega_{cluster}^{est} - \Omega_{cluster} \geqslant 0$ *(i.e.,* $\Omega_{cluster}^{est} - \Omega_{cluster}$ *is positive semi-definite).*

Propositions 6.1 and 6.2 together imply that standard confidence intervals based on the cluster-robust variance estimator will have asymptotically correct but possibly conservative coverage for $\tau_{EATT}^{cluster} + \delta_{cluster}$. In Appendix C, we further characterize the probability limit of the heteroskedasticity-robust variance estimator that ignores clustering (see Proposition C.3). An immediate implication is that the sign of the asymptotic bias of the

heteroskedasticity-robust variance estimator is ambiguous, and so confidence intervals based on the conventional heteroskedasticity-robust variance estimator may not be valid even in finite populations with a large number of clusters.

Altogether, these results imply that if the need for clustering arises from the stochastic assignment of treatment, then the researcher should cluster at the level at which treatment is assigned. This recommendation is similar to that in Abadie et al. (2022), although they study a two-step data-generating process in which cluster-level treatment probabilities are initially drawn according to some fixed distribution that is unrelated to potential outcomes. Each cluster therefore has the same treatment probability marginalized over the two-step process, and hence the ATE is consistently estimable in their framework. Consequently, their variance calculations are not directly applicable to quasi-experimental settings, such as DiD, where units may have different treatment probabilities and the causal estimand may be the ATT rather than the ATE.[21]

6.2 Linear covariate adjustment

Suppose each unit $i = 1, \ldots, N$ is associated with a vector of fixed covariates $W_i \in \mathbb{R}^k$ that are unaffected by the treatment. We consider the ordinary least squares estimator that adjusts for these fixed covariates, i.e. the OLS regression of the observed outcome on a constant, the treatment indicator D_i, and the fixed covariates W_i. Our next result characterizes the causal interpretation of the estimand associated with the OLS coefficient on the treatment indicator in such a regression. This is sometimes referred to as the "covariate-adjusted" difference-in-means, and was studied in the case of completely randomized experiments by Freedman (2008) and Lin (2013), among others; our results extend the study of this estimator to settings where treatment probabilities may differ across units.

Proposition 6.3. *Assume* $\mathbb{E}_R\left[\frac{1}{N}\sum_{i=1}^N (1, D_i, W_i')'(1, D_i, W_i')\right]$ *is invertible. Let* β_D *denote the coefficient on* D_i *in the best linear projection of* Y_i *on* $(1, D_i, X_i')'$ *over the randomization distribution, as defined formally in equation* (19) *in the Appendix. Then,*

$$\beta_D = \tau_{EATT} + \frac{N}{N_1}\frac{N}{N_0}\mathbb{C}ov_1\left[\pi_i, Y_i(0) - \gamma' W_i\right],$$

where $\gamma = \theta\gamma(1) + (1-\theta)\gamma(0)$ *for* $\theta = \left(\frac{N_1}{N}\mathbb{V}ar_\pi\left[W_i\right] + \frac{N_0}{N}\mathbb{V}ar_{1-\pi}\left[W_i\right]\right)^{-1}\left(\frac{N_1}{N}\mathbb{V}ar_\pi\left[W_i\right]\right)$,

[21] Abadie et al. (2022) allow for both sampling- and design-based uncertainty simultaneously, whereas we only consider the case of design-based uncertainty. Xu (2021) and Xu and Wooldridge (2022) study clustered standard errors for non-linear estimators from a design-based perspective. Their results cover inference on a finite-population argmin that is well-defined if units have varying treatment probabilities, although existing results require the propensity score to be linear in observable covariates to give a causal interpretation to this parameter.

$\gamma(1) = \mathbb{V}ar_{\pi}[W_i]^{-1}\mathbb{C}ov_{\pi}[W_i, Y_i(1)]$, *and* $\gamma(0) = \mathbb{V}ar_{1-\pi}[W_i]^{-1}\mathbb{C}ov_{1-\pi}[W_i, Y_i(0)]$.

The proposition establishes that the estimand associated with the adjusted difference-in-means estimator can be decomposed into the EATT plus a bias term that now depends on the finite-population covariance between the treatment probabilities π_i and a covariate-adjusted untreated potential outcome $Y_i(0) - \gamma' W_i$. The coefficient γ is a weighted average of the (π-weighted) projection of $Y_i(1)$ onto W_i and the corresponding ($(1-\pi)$-weighted) projection of $Y_i(0)$ on W_i.

Alternatively, we can characterize the OLS estimand in terms of the relationship between the treatment probabilities π_i and the covariates W_i.

Proposition 6.4. *Let* $\bar{W}_i = (1, W_i')'$. *Under the conditions of Proposition 6.3,*

$$\beta_D = \tau_{OLS} + \mathbb{E}_1[\pi_i(1-\hat{\pi}_i)]^{-1}\mathbb{C}ov_1[\pi_i - \hat{\pi}_i, Y_i(0)]$$

where $\hat{\pi}_i = \omega' \bar{W}_i$ *for* $\omega = \mathbb{E}_1\left[\bar{W}_i \bar{W}_i'\right]^{-1}\mathbb{E}_1\left[\bar{W}_i \pi_i\right]$, *and* $\tau_{OLS} = \mathbb{E}_1[\pi_i(1-\hat{\pi}_i)]^{-1}\mathbb{E}_1[\pi_i(1-\hat{\pi}_i)\tau_i]$.

Proposition 6.4 shows that β_D is the sum of two terms. The first, τ_{OLS}, is a weighted average of treatment effects with weights proportional to $\pi_i(1-\hat{\pi}_i)$, where $\hat{\pi}_i$ is the prediction of the best linear predictor of π_i given the covariates $\bar{W}_i$. The second term is a bias term that depends on the covariance between $Y_i(0)$ and $\pi_i - \hat{\pi}_i$, i.e. the difference between the actual treatment probability π_i and the best linear prediction given the covariates $\hat{\pi}_i$. In the special case where the π_i are linear in observed covariates, we have that $\hat{\pi}_i = \pi_i$, in which case Proposition 6.4 implies that $\beta_D = \mathbb{E}_{\tilde{\pi}}[\tau_i]$. This is a weighted average of treatment effects with weights proportional to the variance of the treatment indicator, $\tilde{\pi}_i = \pi_i(1-\pi_i) = \mathbb{V}_R[D_i]$. Proposition 6.4 thus nests as a special case the finding that when the propensity score is linear, OLS gives a variance-weighted average of treatment effects; see Angrist (1998) and Abadie et al. (2020) for similar results in a super-population and design-based setting, respectively. Proposition 6.4 generalizes this finding to the case where the propensity score may not be linear in covariates.

In Appendix C, we provide regularity conditions under which $\sqrt{N}(\hat{\beta}_D - \beta_D)$ is asymptotically normally distributed, and show that the typical heteroskedasticity-robust standard errors are consistent for an upper bound on the asymptotic variance.[22]

[22] Although OLS is consistent in an experiment (i.e., when $\pi_i = \frac{N_1}{N}$ for all i), Freedman (2008) and Lin (2013) showed that the OLS estimator is biased for the ATE over the randomization distribution. This bias, however, is $O(N^{-1})$, and thus is second-order under conventional asymptotics.

6.3 Instrumental variables

We now analyze the properties of two-stage least squares instrumental variables (IV) estimators from a design-based perspective. Let $Z_i \in \{0,1\}$ be an instrument, $D_i(z) \in \{0,1\}$ be the potential treatment status for $z \in \{0,1\}$, and $Y_i(d)$ be the potential outcome for $d \in \{0,1\}$. The notation $Y_i(d)$ encodes the exclusion restriction that the instrument Z_i only causally affects the outcome through the treatment D_i. We also maintain the typical monotonocity assumption: $D_i(1) \geqslant D_i(0)$ for all units $i = 1, \ldots, N$. The observed data is (Y_i, D_i, Z_i), where $Y_i = Y_i(D_i(Z_i))$ and $D_i = D_i(Z_i)$ for each unit.

In a slight change from our previous analysis, we now view the instrument as stochastic, holding fixed (i.e. conditioning on) the potential treatments $D(\cdot) = \{D_i(\cdot) \colon i = 1, \ldots, N\}$ and potential outcomes $Y(\cdot) = \{Y_i(\cdot) \colon i = 1, \ldots, N\}$. This is sensible, since in IV settings researchers often discuss how the instrument is determined by idiosyncratic factors. We define N_1^Z to be the number of units with $Z_i = 1$ and N_0^Z be the number of units with $Z_i = 0$, and conduct our analysis conditional on N_1^Z.

In canonical IV frameworks, the instrument is typically assumed to be independent of the potential treatments and outcomes (see, e.g., Angrist and Imbens (1994) for a sampling-based setting, and Kang et al. (2018) for a design-based setting).[23] We instead allow the probability that $Z_i = 1$ to vary across units and to be arbitrarily related to the potential treatments and outcomes. The assignment of the instrument Z_i mimics the assignment mechanism in (1) and satisfies

$$\mathbb{P}\left(Z = z \middle| \sum_i Z_i = N_1^Z, D(\cdot), Y(\cdot)\right) \propto \prod_i p_i^{z_i}(1-p_i)^{1-z_i} \tag{13}$$

for all $Z \in \{0,1\}^N$ such that $\sum_i z_i = N_1^Z$, and zero otherwise. To avoid additional notation, let $\mathbb{P}_R(\cdot)$, $\mathbb{E}_R[\cdot]$, $\mathbb{V}_R[\cdot]$ now denote probabilities, expectations, and variances respectively under the randomization distribution (13). Denote the marginal assignment probability as $\pi_i^Z := \mathbb{P}_R(Z_i = 1)$.

Consider the popular two-stage least-squares (2SLS) estimator, defined as $\hat{\beta}_{2SLS} := \hat{\tau}_{RF}/\hat{\tau}_{FS}$ with

$$\hat{\tau}_{RF} = \frac{1}{N_1^Z}\sum_i Z_i Y_i - \frac{1}{N_0^Z}\sum_i (1-Z_i)Y_i \text{ and } \hat{\tau}_{FS} := \frac{1}{N_1^Z}\sum_i Z_i D_i - \frac{1}{N_0^Z}\sum_i (1-Z_i)D_i.$$

In order to analyze the behavior of $\hat{\beta}_{2SLS}$ over the randomization distribution, observe that

[23] Hong, Leung and Li (2020) consider a design-based IV setting where the instrument is randomly assigned within strata defined by observable characteristics.

$\hat{\tau}_{RF}$ is a SDIM estimator for the "reduced-form" effect of Z_i on Y_i, whereas $\hat{\tau}_{FS}$ is a SDIM estimator for the "first-stage" effect of Z_i on D_i. Proposition 3.1 and the monotonicity assumption therefore together imply that

$$\mathbb{E}_R\left[\hat{\tau}_{RF}\right] = \frac{1}{N_1^Z}\sum_{i\in\mathcal{C}} \pi_i^Z (Y_i(1) - Y_i(0)) + \frac{N}{N_1^Z}\frac{N}{N_0^Z}\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right]$$
$$\mathbb{E}_R\left[\hat{\tau}_{FS}\right] = \frac{1}{N_1^Z}\sum_{i\in\mathcal{C}} \pi_i^Z + \frac{N}{N_1^Z}\frac{N}{N_0^Z}\mathbb{C}\text{ov}_1\left[\pi_i^Z, D_i(0)\right],$$

where $\mathcal{C} := \{i\colon D_i(1) > D_i(0)\}$ is the set of complier units. Define the 2SLS estimand as $\beta_{2SLS} := \frac{\mathbb{E}_R[\hat{\tau}_{RF}]}{\mathbb{E}_R[\hat{\tau}_{FS}]}$.

The generalization of our results to vector-valued outcomes in Appendix B implies that $\hat{\beta}_{2SLS}$ is normally distributed around β_{2SLS} in large finite-populations. Concretely, if the sequence of finite-populations satisfies the assumptions in Proposition B.1(4), then

$$\sqrt{N}\begin{pmatrix} \hat{\tau}_{RF} - \mathbb{E}_R\left[\hat{\tau}_{RF}\right] \\ \hat{\tau}_{FS} - \mathbb{E}_R\left[\hat{\tau}_{FS}\right] \end{pmatrix} \to_d \mathcal{N}\left(0,\, \Sigma_\tau\right),$$

where $\Sigma_\tau = \lim_{N\to\infty} N\mathbb{V}_{R_Z}\left[\begin{pmatrix} \hat{\tau}_{RF} \\ \hat{\tau}_{FS} \end{pmatrix}\right]$. Assuming further that the sequence of finite-populations satisfies $\left(\mathbb{E}_R\left[\hat{\tau}_{RF}\right], \mathbb{E}_R\left[\hat{\tau}_{FS}\right]\right) \to (\tau_{RF}^*, \tau_{FS}^*)$ with $\tau_{FS}^* > 0$, then the uniform delta method (e.g., Theorem 3.8 in van der Vaart (2000)) implies that[24]

$$\sqrt{N}(\hat{\beta}_{2SLS} - \beta_{2SLS}) \to_d N(0, g'\Sigma_\tau g),$$

where g is the gradient of $h(x,y) = x/y$ evaluated at $(\tau_{RF}^*, \tau_{FS}^*)$. Typical standard errors for IV will therefore be correct for β_{2SLS} but potentially conservative from the design-based view in large finite-populations with a strong first-stage. If the instrument were instead assigned at a cluster-level, then the results in Section 6.1 would likewise imply that standard errors should be clustered at the level at which the instrument is determined.

What is the causal interpretation of the estimand β_{2SLS}? Notice that if $\pi_i^Z \equiv \frac{N_1^Z}{N}$, so that all units receive $Z = 1$ with equal probability, then $\beta_{2SLS} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}(Y_i(1) - Y_i(0))$, which is a design-based analog to the canonical local average treatment effect (LATE) for compliers

[24] It is well-known in sampling-based instrumental variables settings that the delta method fails under "weak-instrument asymptotics" in which $\mathbb{E}_R\left[\hat{\tau}_{FS}\right]$ drifts towards zero (Staiger and Stock, 1997). Similar issues apply here. However, the test static used to form Anderson-Rubin confidence intervals, which are robust to weak identification, can be written as a quadratic form in a SDIM statistic (see, e.g., Li and Ding (2017)). Our results could thus also be applied to analyze the properties of Anderson-Rubin based CIs under weak identification asymptotics.

(Angrist et al., 1996; Kang et al., 2018). Our results also imply that β_{2SLS} maintains a causal interpretation under the weaker restriction $\mathbb{C}\text{ov}_1\left[\pi_i^Z, Y_i(D_i(0))\right] = \mathbb{C}\text{ov}_1\left[\pi_i^Z, D_i(0)\right] = 0$, which only requires the probability that $Z_i = 1$ to be orthogonal to the potential outcomes and potential treatments associated with $Z_i = 0$. Under this assumption,

$$\beta_{2SLS} = \frac{1}{\sum_{i\in\mathcal{C}} \pi_i^Z} \sum_{i\in\mathcal{C}} \pi_i^Z \left(Y_i(1) - Y_i(0)\right),$$

and thus the parameter β_{2SLS} is a weighted average treatment effect among the compliers. The weights given to each complier are proportional to π_i^Z, the probability that $Z_i = 1$ under the assignment mechanism (13).

7 Conclusion

This paper develops a design-based framework of uncertainty suitable for quasi-experimental settings. We derive formulas for the bias of common estimators such as DiD as a function of the idiosyncratic treatment probabilities. We show further that common estimators of the variance tend to be conservative when there is heterogeneity in the treatment probabilities π_i (even under constant treatment effects). This conservativeness helps to mitigate undercoverage of conventional confidence intervals when the estimator is biased. Thus, for example, confidence intervals for DiD may have correct coverage of the EATT even if the design-based analog to parallel trends does not hold exactly. Our framework also has useful implications for the choice of the appropriate level of clustering and the interpretation of IV estimators when the instrument is not completely randomly assigned.

The analysis in this paper could be extended in a variety of directions. First, the analysis might be extended to settings where the stochastic nature of the data arises both from the assignment of treatment and from sampling a subset of units from a finite population, as in Abadie et al. (2020). Second, our results suggest that a variety of mis-specification robust tools and sensitivity analyses which have been developed under the assumption of asymptotic normality from a sampling-based perspective could also potentially be applied in finite population contexts as well (e.g., Andrews, Gentzkow and Shapiro, 2017; Armstrong and Kolesár, 2018, 2020; Andrews, Gentzkow and Shapiro, 2020; Bonhomme and Weidner, 2022). However, the finite population setting studied here differs from the usual sampling-based approach in that the variance matrix is only conservatively estimated, and the degree to which it is conservative may depend on the extent to which the baseline model is violated. It would be useful to study which of the theoretical guarantees in the aforementioned papers (e.g., size control, optimality) are robust to this modification.

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," Technical Report 2022. arXiv:1710.02926.

_ , _ , Guido W. Imbens, and Jeffrey M. Wooldridge, "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "On the Informativeness of Descriptive Statistics for Structural Estimates," *Econometrica*, 2020, *88* (6), 2231–2258.

_ , _ , and Jesse Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 2017, *132* (4), 1553–1592.

Angrist, Joshua and Guido Imbens, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Angrist, Joshua D., "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 1998, *66* (2), 249–288. Publisher: [Wiley, Econometric Society].

_ and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

_ , Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

Armstrong, Timothy and Michal Kolesár, "Optimal Inference in a Class of Regression Models," *Econometrica*, 2018, *86*, 655–683.

Armstrong, Timothy B. and Michal Kolesár, "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 2020, *11* (1), 1–39.

Aronow, Peter M. and Donald K. K. Lee, "Interval estimation of population means under unknown but bounded probabilities of sample selection," *Biometrika*, 2013, *100* (1), 235–240.

_ and Joel A. Middleton, "A class of unbiased estimators of the average treatment effect in randomized experiments," *Journal of Causal Inference*, 2015, *1* (1), 135–154.

Athey, Susan and Guido W. Imbens, "Design-based analysis in Difference-In-Differences settings with staggered adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79. Annals Issue in Honor of Gary Chamberlain.

Basu, Debabrata, "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association*, 1977, *72* (358), 355–366.

Berger, Yves G., "Rate of convergence to normal distribution for the Horvitz-Thompson estimator," *Journal of Statistical Planning and Inference*, April 1998, *67* (2), 209–226.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, February 2004, *119* (1), 249–275.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard, "Panel Experiments and Dynamic Causal Effects: A Finite Population Perspective," *Quantitative Economics*, 2021, *12* (4), 1171–1196.

Bonhomme, Stéphane and Martin Weidner, "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 2022, *13* (3), 907–954.

Borusyak, Kirill and Peter Hull, "Non-Random Exposure to Exogenous Shocks: Theory and Applications," September 2020.

_ **and Xavier Jaravel**, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2016.

Buehler, Robert J., "Some Validity Criteria for Statistical Inferences," *The Annals of Mathematical Statistics*, 1959, *30* (4), 845–863. Publisher: Institute of Mathematical Statistics.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (9), 2964–96.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif, "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 2019, *109* (12), 4178–4219.

Fisher, R. A., *The Design of Experiments*, Oxford, England: Oliver & Boyd, 1935.

Fisher, Sir Ronald Aylmer, *Statistical Methods and Scientific Inference*, Oliver and Boyd, 1959. Google-Books-ID: oyXPAAAAMAAJ.

Freedman, David A., "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 2008, *40* (2), 180–193.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Hajek, Jaroslav, "Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population," *Annals of Mathematical Statistics*, December 1964, *35* (4), 1491–1523. Publisher: Institute of Mathematical Statistics.

Heckman, James, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," NBER Chapters, National Bureau of Economic Research, Inc 1976.

Hong, Han, Michael P Leung, and Jessie Li, "Inference on finite-population treatment effects under limited overlap," *The Econometrics Journal*, January 2020, *23* (1), 32–47.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, February 2004, *86* (1), 4–29. Publisher: MIT Press.

_ **and Donald B. Rubin**, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press, 2015.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico, "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms," *The Quarterly Journal of Economics*, February 2016, *131* (1), 157–218.

Kang, Hyunseung, Laura Peck, and Luke Keele, "Inference for instrumental variables: a randomization inference approach," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2018, *181* (4), 1231–1254.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach, "School Finance Reform and the Distribution of Student Achievement," *American Economic Journal: Applied Economics*, April 2018, *10* (2), 1–26.

Li, Xinran and Peng Ding, "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference," *Journal of the American Statistical Association*, October 2017, *112* (520), 1759–1769.

Liang, Kung-Yee and Scott L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, 1986, *73* (1), 13–22.

Lin, Winston, "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, 2013, *7* (1), 295–318.

Manski, Charles F. and John V. Pepper, "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *Review of Economics and Statistics*, 2018, *100* (2), 232–244.

McKelvey, Richard D. and Thomas R. Palfrey, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, July 1995, *10* (1), 6–38.

Miratrix, Luke W., Stefan Wager, and Jose R. Zubizarreta, "Shape-constrained partial identification of a population mean under unknown probabilities of sample selection," *Biometrika*, 2018, *105* (1), 103–114.

MIT Election Data and Science Lab, "U.S. President 1976–2020," 2022.

Müller, Ulrich K. and Andriy Norets, "Credibility of Confidence Sets in Nonstandard Econometric Problems," *Econometrica*, 2016, *84* (6), 2183–2213.

Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923, *5* (4), 465–472. Publisher: Institute of Mathematical Statistics.

Pashley, Nicole E., Guillaume W. Basse, and Luke W. Miratrix, "Conditional as-if analyses in randomized experiments," *Journal of Causal Inference*, January 2021, *9* (1), 264–284. Publisher: De Gruyter.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, Forthcoming.

Rosenbaum, Paul R., "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 1987, *74* (1), 13–26. Publisher: [Oxford University Press, Biometrika Trust].

_ , *Observational Studies*, Springer Science & Business Media, January 2002. Google-Books-ID: K0OglGXtpGMC.

_ , "Sensitivity Analysis in Observational Studies," in B. S. Everitt and D. C. Howell, eds., *Encyclopedia of Statistics in Behavioral Science*, 2005.

Roth, Jonathan, "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights*, September 2022, *4* (3), 305–322.

_ **and Pedro H. C. Sant'Anna**, "Efficient Estimation for Staggered Rollout Designs," *arXiv:2102.01291 [econ, math, stat]*, June 2021. arXiv: 2102.01291.

Savje, Frederik and Angele Delevoye, "Consistency of the Horvitz-Thompson estimator under general sampling and experimental designs," *Journal of Statistical Planning and Inference*, 2020, *207*, 190–197.

Sekhon, Jasjeet S. and Yotam Shem-Tov, "Inference on a New Class of Sample Average Treatment Effects," *Journal of the American Statistical Association*, February 2020, pp. 1–18. Publisher: Taylor & Francis.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586. Publisher: [Wiley, Econometric Society].

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

United States Bureau of the Census, "Quarterly Workforce Indicators," Technical Report 2022.

van der Vaart, A. W., *Asymptotic Statistics*, Cambridge University Press, June 2000.

Xu, Ruonan, "Potential outcomes and finite-population inference for M-estimators," *The Econometrics Journal*, January 2021, *24* (1), 162–176.

_ **and Jeffrey M Wooldridge**, "A Design-Based Approach to Spatial Correlation," 2022, p. 70.

A Proofs for results in main text

Proof of Proposition 3.1

Proof. Recall $\mathbb{E}_R[D_i] = \pi_i$ and $\tau_i = Y_i(1) - Y_i(0)$. Hence, we have that

$$\begin{aligned}\mathbb{E}_R[\hat{\tau}] &= \mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i Y_i(1) + \frac{1}{N_0}\sum_i D_i Y_i(0)\right] \\ &= \frac{1}{N_1}\sum_i \pi_i \underbrace{(Y_i(0)+\tau_i)}_{=Y_i(1)} - \frac{1}{N_0}\sum_i (1-\pi_i) Y_i(0) \\ &= \underbrace{\frac{1}{N_1}\sum_i \pi_i \tau_i}_{=:\tau_{EATT}} + \frac{N}{N_0}\frac{N}{N_1}\underbrace{\left(\frac{1}{N}\sum_i \left(\pi_i - \frac{N_1}{N}\right) Y_i(0)\right)}_{=\mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(0)]},\end{aligned} \tag{14}$$

which yields the second expression in the Proposition. To derive the first expression, note that

$$\tau_{EATT} = \frac{1}{N_1}\sum_i (\pi_i - \frac{N_1}{N})\tau_i + \frac{1}{N}\sum_i \tau_i = \frac{N}{N_1}\mathbb{C}\mathrm{ov}_1[\pi_i, \tau_i] + \tau_{ATE}.$$

Further, since $\tau_i = Y_i(1) - Y_i(0)$, we have that $\mathbb{C}\mathrm{ov}_1[\pi_i, \tau_i] = \mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(1)] - \mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(0)]$, and hence

$$\tau_{EATT} = \tau_{ATE} + \frac{N}{N_1}\mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(1)] - \frac{N}{N_1}\mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(0)].$$

Substituting this expression into (14) and simplifying then yields

$$\mathbb{E}_R[\hat{\tau}] = \tau_{ATE} + \frac{N}{N_1}\mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(1)] + \frac{N}{N_0}\mathbb{C}\mathrm{ov}_1[\pi_i, Y_i(0)],$$

as needed. □

Proof of Proposition 4.1

Proof. Since $\hat{\tau}$ can be represented as a Horvitz-Thompson estimator under rejective sampling, Theorem 6.1 in Hajek (1964) implies

$$\mathbb{V}_R[\hat{\tau}][1+o(1)] = \left[\sum_{k=1}^N \pi_k(1-\pi_k)\right]\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\tilde{Y}_i\right] = \left[\sum_{k=1}^N \pi_k(1-\pi_k)\right]\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right]. \tag{15}$$

Standard decomposition arguments for completely randomized experiments (e.g. Imbens and Rubin (2015)), modified to replace unweighted variances with weighted variances, yield

$$\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right]=\frac{N}{N_1N_0}\left(\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(1)\right]+\frac{1}{N_0}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(0)\right]-\frac{1}{N}\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[\tau_i\right]\right),$$

which together with the previous display yields the desired result. □

Proof of Lemma 4.1

Proof. We will show that $\mathbb{E}_R\left[\hat{s}_1^2\right](1+o(1))=\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]$. The equality $\mathbb{E}_R\left[\hat{s}_0^2\right](1+o(1))=\mathbb{V}\mathrm{ar}_{1-\pi}\left[Y_i(0)\right]$ can be obtained analogously, from which the result is immediate. Observe that

$$\begin{aligned}\mathbb{E}_R\left[\hat{s}_1^2\right]&=\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_iY_i^2-\bar{Y}_1^2\right]=\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_iY_i^2-(\bar{Y}_1-\mathbb{E}_{\pi}\left[Y_i(1)\right]+\mathbb{E}_{\pi}\left[Y_i(1)\right])^2\right]\\&=\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_iY_i^2\right]-\mathbb{E}_{\pi}\left[Y_i(1)\right]^2-2\mathbb{E}_{\pi}\left[Y_i(1)\right]\mathbb{E}_R\left[\bar{Y}_1-\mathbb{E}_{\pi}\left[Y_i(1)\right]\right]-\mathbb{E}_R\left[(\bar{Y}_1-\mathbb{E}_{\pi}\left[Y_i(1)\right])^2\right]\\&=\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]-\mathbb{V}_R\left[\bar{Y}_1\right],\end{aligned}$$

where the last equality is obtained using the fact that $\mathbb{E}_R\left[D_i\right]=\pi_i$, and hence $\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_iY_i^2\right]=\mathbb{E}_{\pi}\left[Y_i(1)^2\right]$ and $\mathbb{E}_R\left[\bar{Y}_1-\mathbb{E}_{\pi}\left[Y_i(1)\right]\right]=0$. Applying Theorem 6.1 in Hajek (1964) as in the proof to Proposition 4.1, we see that

$$\mathbb{V}_R\left[\bar{Y}_1\right](1+o(1))=\left[\sum_k \pi_k(1-\pi_k)\right]\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(1)/N_1\right].$$

Next, observe that

$$\begin{aligned}\left[\sum_k \pi_k(1-\pi_k)\right]\mathbb{V}\mathrm{ar}_{\tilde{\pi}}\left[Y_i(1)/N_1\right]&=\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1)-\mathbb{E}_{\tilde{\pi}}\left[Y_i(1)\right])^2\\&\leqslant\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1)-\mathbb{E}_{\pi}\left[Y_i(1)\right])^2\\&\leqslant\frac{1}{N_1^2}\sum_i \pi_i(Y_i(1)-\mathbb{E}_{\pi}\left[Y_i(1)\right])^2=\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]\\&\leqslant\left[\sum_k \pi_k(1-\pi_k)\right]^{-1}\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]=o(1)\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]\end{aligned}$$

where the first inequality uses the fact that $\mathbb{E}_{\tilde{\pi}}\left[Y_i(1)\right]=\arg\min_u \sum_i \pi_i(1-\pi_i)(Y_i(1)-u)^2$, the second inequality uses the fact that $\pi_i(1-\pi_i)\leqslant\pi_i$, and the third inequality uses the fact that $N_1=\sum_i \pi_i\geqslant\sum_i \pi_i(1-\pi_i)$. Combining the previous three displays, we see that $\mathbb{E}_R\left[\hat{s}_1^2\right]=(1+o(1))\mathbb{V}\mathrm{ar}_{\pi}\left[Y_i(1)\right]$, as we wished to show. □

Proof of Proposition 4.2

Proof. From (15), we see that the right-hand side of (7) is equivalent to

$$\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)-\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right]\right)^2.$$

Since for any X, $\mathbb{E}_{\tilde{\pi}}[X_i] = \arg\min_{\mu} \sum_{i=1}^{N} \pi_i(1-\pi_i)(X_i-\mu)^2$, it follows that this is bounded above by

$$\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)-\left(\mathbb{E}_{\pi}\left[\frac{1}{N_1}Y_i(1)\right]+\mathbb{E}_{1-\pi}\left[\frac{1}{N_0}Y_i(0)\right]\right)\right)^2, \tag{16}$$

and the bound holds with equality if and only if

$$\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right] = \frac{1}{N_1}\mathbb{E}_{\pi}[Y_i(1)]+\frac{1}{N_0}\mathbb{E}_{1-\pi}[Y_i(0)]. \tag{17}$$

Let $\dot{Y}_i(1) = Y_i(1) - \mathbb{E}_{\pi}[Y_i(1)]$ and $\dot{Y}_i(0) = Y_i(0) - \mathbb{E}_{1-\pi}[Y_i(0)]$. Then the expression in (16) can be written as

$$\begin{aligned}
&\sum_{i=1}^{N} \pi_i(1-\pi_i)\left(\frac{1}{N_1}\dot{Y}_i(1)+\frac{1}{N_0}\dot{Y}_i(0)\right)^2 \\
&= \left[\frac{1}{N_1^2}\sum_{i=1}^{N}\pi_i\dot{Y}_i(1)^2+\frac{1}{N_0^2}\sum_{i=1}^{N}(1-\pi_i)\dot{Y}_i(0)^2-\right. \\
&\left.\frac{1}{N_1^2}\sum_{i=1}^{N}\pi_i^2\dot{Y}_i(1)^2-\frac{1}{N_0^2}\sum_{i=1}^{N}(1-\pi_i)^2\dot{Y}_i(0)^2+\frac{2}{N_1N_0}\sum_{i=1}^{N}\pi_i(1-\pi_i)\dot{Y}_i(1)\dot{Y}_i(0)\right] \\
&= \left[\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\pi}[Y_i(1)]+\frac{1}{N_0}\mathbb{V}\mathrm{ar}_{1-\pi}[Y_i(0)]-\frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{\pi_i}{N_1/N}\dot{Y}_i(1)-\frac{1-\pi_i}{N_0/N}\dot{Y}_i(0)\right)^2\right],
\end{aligned}$$

from which the first claim is immediate. Furthermore, we immediately observe that $\frac{\mathbb{V}_R^{approx}[\hat{\tau}]}{\mathbb{E}_R^{approx}[\hat{s}^2]} = 1$ if and only if both (17) holds and

$$\frac{\pi_i}{N_1/N}Y_i(1)-\frac{1-\pi_i}{N_0/N}Y_i(0) = \frac{\pi_i}{N_1/N}\mathbb{E}_{\pi}[Y_i(1)]-\frac{1-\pi_i}{N_0/N}\mathbb{E}_{1-\pi}[Y_i(0)] \text{ for all } i. \tag{18}$$

Note that equation (9) is just a re-arrangement of the terms in (18). To complete the proof, it thus suffices to show that (18) actually implies (17). To do this, we multiply both sides of (18) by $(1-\pi_i)/N$ and sum across i to obtain that

$$s\cdot\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1)+\frac{1}{N_0}Y_i(0)\right]-\mathbb{E}_{1-\pi}[Y_i(0)] = \frac{s}{N_1}\mathbb{E}_{\pi}[Y_i(1)]-\frac{1}{N_0}\sum_i(1-\pi_i)^2\mathbb{E}_{1-\pi}[Y_i(0)],$$

where $s = \sum_i \pi_i(1-\pi_i)$. Re-arranging terms, we obtain that

$$\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right] = \frac{1}{N_1}\mathbb{E}_{\pi}\left[Y_i(1)\right] + \frac{1}{N_0}\frac{1}{s}\left(N_0 - \sum_i (1-\pi_i)^2\right)\mathbb{E}_{1-\pi}\left[Y_i(0)\right].$$

Note, however, that

$$N_0 - \sum_i (1-\pi_i)^2 = N_0 - \sum_i (1-\pi_i) + \sum_i \pi_i(1-\pi_i) = s,$$

and thus,

$$\mathbb{E}_{\tilde{\pi}}\left[\frac{1}{N_1}Y_i(1) + \frac{1}{N_0}Y_i(0)\right] = \frac{1}{N_1}\mathbb{E}_{\pi}\left[Y_i(1)\right] + \frac{1}{N_0}\mathbb{E}_{1-\pi}\left[Y_i(0)\right],$$

as needed. □

Proof of Corollary 4.1

Proof. Note that we can re-write (18) as

$$\frac{\pi_i}{N_1}(Y_i(1) - \mathbb{E}_{\pi}\left[Y_i(1)\right]) - \frac{1-\pi_i}{N_0}(Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]) = 0 \text{ for all } i.$$

Under constant effects, $\tau_{EATT} = \tau$. Further, from display (14), we see that $\mathbb{E}_R\left[\hat{\tau}\right] = \tau_{EATT} + \mathbb{E}_{\pi}\left[Y_i(0)\right] - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]$, and thus if $\mathbb{E}_R\left[\hat{\tau}\right] = \tau_{EATT}$, then $\mathbb{E}_{\pi}\left[Y_i(0)\right] = \mathbb{E}_{1-\pi}\left[Y_i(0)\right]$. Additionally, under the constant effects assumption, $Y_i(1) - \mathbb{E}_{\pi}\left[Y_i(1)\right] = Y_i(0) - \mathbb{E}_{\pi}\left[Y_i(0)\right]$, and hence $Y_i(1) - \mathbb{E}_{\pi}\left[Y_i(1)\right] = Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]$. Substituting into the previous display and re-arranging terms, we obtain that

$$\left(\frac{\pi_i}{N_1} - \frac{1-\pi_i}{N_0}\right)(Y_i(0) - \mathbb{E}_{1-\pi}\left[Y_i(0)\right]) = 0 \text{ for all } i,$$

from which the result follows.

□

Proof of Proposition 4.3

Proof. First, viewing $\hat{\tau}$ as a Horwitz-Thompson estimator under rejective sampling as in (6), the central limit theorem follows immediately from Theorem 1 in Berger (1998).[25]

Second, to show convergence of $\hat{s}^2/\mathbb{E}_R^{approx}\left[\hat{s}^2\right]$, it suffices to show that $\dfrac{\hat{s}_1^2}{\mathbb{V}\text{ar}_{\pi}\left[Y_i(1)\right]} \to_p 1$ and $\dfrac{\hat{s}_0^2}{\mathbb{V}\text{ar}_{1-\pi}\left[Y_i(0)\right]} \to_p 1$. We provide a proof for the former; the latter proof is analogous.

[25] Hajek (1964) states a similar result where the Horvitz-Thompson estimator uses an approximation to the marginal probabilities $\pi_i = \mathbb{E}_R\left[D_i\right]$ in terms of the underlying idiosyncratic probabilities p_i.

For notational convenience, let $v_1 = \mathbb{V}\text{ar}_\pi\left[Y_i(1)\right]$. From the definition of $\hat{s}_1^2$, we can write

$$\frac{\hat{s}_1^2}{v_1} = \frac{1}{v_1}\left(\left(\frac{1}{N_1}\sum_i D_i(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2\right) - (\bar{Y}_1 - \mathbb{E}_\pi\left[Y_i(1)\right])^2\right).$$

Now, $\frac{1}{N_1}\sum_i D_i(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2$ can be viewed as a Horvitz-Thompson estimator of $\frac{1}{N_1}\sum_i \pi_i(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2 = v_1$, and thus by Theorem 6.1 in Hajek (1964), its variance is equal to

$$(1 + o(1))\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}\left[(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2\right].$$

Note further that

$$\begin{aligned}\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}\left[(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2\right] &\leqslant \frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^4 \\ &\leqslant \frac{1}{N_1^2} m_N(1) \sum_i \pi_i(Y_i(1) - \mathbb{E}_\pi\left[(Y_i(1)\right])^2 \\ &= \frac{1}{N_1} m_N(1) \mathbb{V}\text{ar}_\pi\left[Y_i(1)\right].\end{aligned}$$

Applying Chebychev's inequality, we have

$$\frac{1}{N_1}\sum_i (D_i(Y_i(1) - \mathbb{E}_\pi\left[Y_i(1)\right])^2 - v_1 = O_p\left(\sqrt{\frac{1}{N_1} m_N(1) \mathbb{V}\text{ar}_\pi\left[Y_i(1)\right]}\right).$$

Next, viewing $\bar{Y}_1$ as a Horvitz-Thomson estimator, we see that its variance is $(1 + o(1))\left(\frac{1}{N_1^2}\sum_i \pi_i(1-\pi_i)\right) \cdot \mathbb{V}\text{ar}_{\tilde{\pi}}\left[Y_i(1)\right]$, which by similar logic to that above is bounded above by $(1 + o(1))\frac{1}{N_1}\mathbb{V}\text{ar}_\pi\left[Y_i(1)\right]$. Thus, by Chebychev's inequality,

$$\bar{Y}_1 - \mathbb{E}_\pi\left[Y_i(1)\right] = O_p\left(\sqrt{\frac{1}{N_1}\mathbb{V}\text{ar}_\pi\left[Y_i(1)\right]}\right).$$

Combining the results above, it follows that

$$\frac{\hat{s}_1^2}{v_1} = \frac{1}{v_1}\left(v_1 + O_p\left(\sqrt{\frac{m_N(1) v_1}{N_1}}\right) + O_p\left(\frac{1}{N_1} v_1\right)\right) = 1 + O_p\left(\sqrt{\frac{m_N(1)}{v_1 N_1}}\right) + O_p\left(\frac{1}{N_1}\right).$$

However, the first O_p term converges to 0 by assumption, and since Assumption 4.1 implies that $N_1 \to \infty$, the second O_p term converges to 0 as well. □

Proof of Proposition 4.4

Proof. Viewing $\hat{\tau}$ as a Horvitz-Thompson estimator under rejective sampling once again, the result follows immediately from Theorem 3 in Berger (1998). □

Proof of Proposition 4.5

Proof. From Proposition 4.3, we have that $\frac{\hat{\tau}-\mathbb{E}_R[\hat{\tau}]}{\sqrt{\mathbb{V}_R^{approx}[\hat{\tau}]}} \xrightarrow{d} \mathcal{N}(0,\,1)$. Observe that we can write

$$\frac{\hat{\tau}-\tau_{EATT}}{\hat{s}} = \frac{\sqrt{\mathbb{E}_R^{approx}\left[\hat{s}^2\right]}}{\hat{s}}\sqrt{\frac{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}{\mathbb{E}_R^{approx}\left[\hat{s}^2\right]}}\left(\frac{\hat{\tau}-\mathbb{E}_R\left[\hat{\tau}\right]}{\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}}+\frac{b}{\sqrt{\mathbb{V}_R^{approx}\left[\hat{\tau}\right]}}\right),$$

where $\mathbb{E}_R\left[\hat{\tau}\right] = \tau_{EATT}+b$ by Proposition 3.1. However, by Proposition 4.3 and the continuous mapping theorem,

$$\frac{\sqrt{\mathbb{E}_R^{approx}\left[\hat{s}^2\right]}}{\hat{s}} \xrightarrow{p} 1.$$

It then follows from Slutky's lemma and the assumptions of the proposition that

$$\frac{\hat{\tau}-\tau_{EATT}}{\hat{s}} \xrightarrow{d} r\cdot\left(\mathcal{N}(0,\,1)+b^*\right) = \mathcal{N}\left(b^*\cdot r,\, r^2\right).$$

□

Proof of Proposition 6.1

Proof. To prove these results, we will show that the second-element of $\beta_{cluster}$ defined in Proposition C.1 equals $\tau_{cluster}^{EATT}+\delta_{cluster}$ when $X_i(d) = (1, d)'$. The stated claims then immediately follow by applying Proposition C.1. Defining $N_1^C = \sum_c \pi_c N_c = \sum_i \pi_{c(i)}$, $N_0^C = N - N_1^C = \sum_i (1-\pi_{c(i)})$, observe that

$$\left(\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XX}_c'(1)\right]+\frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XX}_c'(0)\right]\right)^{-1} = \frac{C}{N_0^C N_1^C}\begin{pmatrix} N_1^C & -N_1^C \\ -N_1^C & N\end{pmatrix},$$
$$\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XY}_c(1)\right]+\frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XY}_c(0)\right] = C^{-1}\sum_i\begin{pmatrix} Y_i(0)+\pi_{c(i)}\tau_i \\ \pi_{c(i)}(Y_i(0)+\tau_i)\end{pmatrix}.$$

Multiplying out, we therefore arrive at

$$\beta_{cluster} = \left(\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XX}_c'(1)\right]+\frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XX}_c'(0)\right]\right)^{-1}\left(\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XY}_c(1)\right]+\frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XY}_c(0)\right]\right) =$$

$$\frac{1}{N_0^C N_1^C}\begin{pmatrix} N_1^C & -N_1^C \\ -N_1^C & N\end{pmatrix}\sum_i\begin{pmatrix} Y_i(0)+\pi_{c(i)}\tau_i \\ \pi_{c(i)}(Y_i(0)+\tau_i)\end{pmatrix} = \begin{pmatrix} \frac{1}{N_0^C}\sum_i(1-\pi_{c(i)})Y_i(0) \\ \frac{1}{N_1^C}\sum_i \pi_{c(i)}\tau_i + \sum_i\left(\frac{\pi_{c(i)}}{N_1^C}-\frac{1-\pi_{c(i)}}{N_0^C}\right)Y_i(0)\end{pmatrix}.$$

Re-arranging the second element then yields

$$\beta_{cluster,2} = \mathbb{E}_{\pi_{c(i)}}\left[\tau_i\right]+\frac{N}{\sum_i \pi_{c(i)}}\frac{N}{N-\sum_i \pi_{c(i)}}\mathbb{C}\text{ov}_1\left[\pi_{c(i)}, Y_i(0)\right]$$

as desired. □

Proof of Proposition 6.2

Proof. This is a special case of Proposition C.2 below with $X_i(d) = (1, d)'$. □

Proof of Proposition 6.3

Proof. Let $E_R^*[\cdot \mid \cdot]$ denote the best linear projection under the randomization distribution with covariates. That is, for unit-level variables $A_i \in \mathbb{R}$, $B_i \in \mathbb{R}^p$, $E_R^*[A_i \mid B_i] = \beta_B' B_i$ for

$$\beta_B := \arg\min_{\beta} \mathbb{E}_R \left[\frac{1}{N} \sum_{i=1}^{N} (A_i - \beta' B_i)^2 \right].$$

Define $\beta = (\beta_0, \beta_D, \beta_W')'$ as the coefficients in the best linear projection of Y_i on $(1, D_i, W_i')'$

$$\beta := \arg\min_{\beta \in \mathbb{R}^{k+2}} \mathbb{E}_R \left[\frac{1}{N} \sum_{i=1}^{N} (Y_i - (1, D_i, W_i')\beta')^2 \right]. \tag{19}$$

To prove the first claim, observe that

$$E_R^*[W_i | 1, D_i] = D_i \mathbb{E}_\pi [W_i] + (1 - D_i) \mathbb{E}_{1-\pi} [W_i].$$

By the Frisch-Waugh-Lovell Theorem,

$$\beta_W = \mathbb{E}_R \left[\frac{1}{N} \sum_i (W_i - E^*[W_i|1, D_i])(W_i - E^*[W_i|1, D_i])' \right]^{-1} \mathbb{E}_R \left[\frac{1}{N} \sum_i (W_i - E^*[W_i|1, D_i]) Y_i \right] =$$

$$\mathbb{E}_R \left[\frac{1}{N} \sum_i D_i (W_i - \mathbb{E}_\pi [W_i])(W_i - \mathbb{E}_\pi [W_i])' + \frac{1}{N} \sum_i (1 - D_i)(W_i - \mathbb{E}_{1-\pi} [W_i])(W_i - \mathbb{E}_{1-\pi} [W_i])' \right]^{-1} \times$$

$$\mathbb{E}_R \left[\frac{1}{N} \sum_i D_i (W_i - \mathbb{E}_\pi [W_i]) Y_i(1) + \frac{1}{N} \sum_i (1 - D_i)(W_i - \mathbb{E}_{1-\pi} [W_i]) Y_i(0) \right] =$$

$$\left(\frac{N_1}{N} \mathbb{V}\mathrm{ar}_\pi [W_i] + \frac{N_0}{N} \mathbb{V}\mathrm{ar}_{1-\pi} [W_i] \right)^{-1} \left(\frac{N_1}{N} \mathbb{E}_\pi [(W_i - \mathbb{E}_\pi [W_i]) Y_i(1)] + \frac{N_0}{N} \mathbb{E}_{1-\pi} [(W_i - \mathbb{E}_{1-\pi} [W_i]) Y_i(0)] \right) =$$

$$\theta \gamma(1) + (1 - \theta) \gamma(0) = \gamma.$$

Note, however, that $E_R^*[Y_i \mid 1, D, W] = E_R^*[Y_i - \beta_W' W_i \mid 1, D]$. It follows that

$$\begin{aligned} \beta_D &= \mathbb{E}_R \left[\frac{1}{N_1} \sum_i D_i (Y_i - \gamma' W_i) - \frac{1}{N_0} \sum_i (1 - D_i)(Y_i - \gamma' W_i) \right] \\ &= \tau_{EATT} + \frac{N_1}{N} \frac{N_0}{N} \mathbb{C}\mathrm{ov}_1 [\pi_i, Y_i(0) - \gamma' W_i], \end{aligned}$$

where the last equality is obtained from applying Proposition 3.1 to the transformed outcome $Y_i - \gamma' W_i$. □

Proof of Proposition 6.4

Proof. By the Frisch-Waugh-Lovell Theorem,

$$E_R^*[Y_i|D_i - \hat{\pi}_i] = \beta_D(D_i - \hat{\pi}_i),$$

and so

$$\beta_D = \mathbb{E}_R\left[\frac{1}{N}\sum_i (D_i - \hat{\pi}_i)^2\right]^{-1} \mathbb{E}_R\left[\frac{1}{N}\sum_i (D_i - \hat{\pi}_i)Y_i\right].$$

Writing $(D_i - \hat{\pi}_i)^2 = D_i - 2D_i\hat{\pi}_i + \hat{\pi}_i^2$ and $Y_i = Y_i(0) + D_i\tau_i$ and evaluating the expectation over the randomization distribution yields

$$\begin{aligned}
\beta_D = \mathbb{E}_1\left[\pi_i - 2\pi_i\hat{\pi}_i + \hat{\pi}_i^2\right]^{-1} \mathbb{E}_R\left[\frac{1}{N}\sum_i (D_i - \hat{\pi}_i)Y_i(0)\right] + \\
\mathbb{E}_1\left[\pi_i - 2\pi_i\hat{\pi}_i + \hat{\pi}_i^2\right]^{-1} \mathbb{E}_R\left[\frac{1}{N}\sum_i D_i(1 - \hat{\pi}_i)\tau_i\right] \\
= \mathbb{E}_1\left[\pi_i - 2\pi_i\hat{\pi}_i + \hat{\pi}_i^2\right]^{-1} \mathbb{E}_1\left[(\pi_i - \hat{\pi}_i)Y_i(0)\right] + \\
\mathbb{E}_1\left[\pi_i - 2\pi_i\hat{\pi}_i + \hat{\pi}_i^2\right]^{-1} \mathbb{E}_1\left[\pi_i(1 - \hat{\pi}_i)\tau_i\right]. \qquad (20)
\end{aligned}$$

Note, however, that $\mathbb{E}_1[\pi_i - \hat{\pi}_i] = 0$, since a constant is included in W_i and thus the regression residuals average to 0, and hence $\mathbb{E}_1[(\pi_i - \hat{\pi}_i)Y_i(0)] = \mathbb{C}\text{ov}_1[\pi_i - \hat{\pi}_i, Y_i(0)]$. Additionally,

$$\mathbb{E}_1\left[\pi_i - 2\pi_i\hat{\pi}_i + \hat{\pi}_i^2\right] = \mathbb{E}_1[\pi_i(1 - \hat{\pi}_i)] + \mathbb{E}_1[\hat{\pi}_i(\hat{\pi}_i - \pi_i)] = \mathbb{E}_1[\pi_i(1 - \hat{\pi}_i)],$$

where $\mathbb{E}_1[\hat{\pi}_i(\hat{\pi}_i - \pi_i)] = 0$ since by construction regression residuals are orthogonal to the regressors. Substituting these expressions into (20) yields the desired result. □

Design-Based Uncertainty for Quasi-Experiments

Online Appendix

Ashesh Rambachan Jonathan Roth

November 21, 2022

This appendix contains additional results and additional Monte Carlo simulations for the paper "Design-Based Uncertainty for Quasi-Experiments" by Ashesh Rambachan and Jonathan Roth. Section B generalizes our analysis of the SDIM estimator under the rejective assignment mechanism to vector-valued outcomes. Section C provides more general analysis of OLS estimators under clustered treatment assignments. Section D contains additional results from our Monte Carlo simulations.

B Extension to vector-valued outcomes

In this appendix, we generalize our results for the SDIM estimator in Sections 3-4 to the vector-valued outcomes case. We apply these results to analyze IV estimators from a design-based perspective in Section 6.3 of the main text, and non-staggered DiD estimators with multiple time periods below.

We extend our notation from the main text, so that $\mathbf{Y}_i \in \mathbb{R}^K$ is the vector-valued outcome. For a fixed vector-valued characteristic $\mathbf{X}_i$, $\mathbb{E}_w[\mathbf{X}_i] := \frac{1}{\sum_i w_i}\sum_i w_i \mathbf{X}_i$ and $\mathbb{V}\text{ar}_w[\mathbf{X}_i] = \frac{1}{\sum_i w_i}\sum_i (\mathbf{X}_i - \mathbb{E}_w[\mathbf{X}_i])(\mathbf{X}_i - \mathbb{E}_w[\mathbf{X}_i])'$. Further, as shorthand, define $S_{1,w} := \mathbb{V}\text{ar}_w[\mathbf{Y}_i(1)]$, $S_{0,w} := \mathbb{V}\text{ar}_w[\mathbf{Y}_i(0)]$, $S_{10,w} := \mathbb{E}_w[(\mathbf{Y}_i(1) - \mathbb{E}_w[\mathbf{Y}_i(1)])(\mathbf{Y}_i(0) - \mathbb{E}_w[\mathbf{Y}_i(0)])']$ to be the weighted finite-population variances and covariance of $\mathbf{Y}_i(1)$ and $\mathbf{Y}_i(0)$. Finally, the vector-valued ATE is $\boldsymbol{\tau}_{ATE} := \frac{1}{N}\sum_i(\mathbf{Y}_i(1) - \mathbf{Y}_i(0))$, and the vector-valued EATT is $\boldsymbol{\tau}_{EATT} := \frac{1}{N_1}\sum_i \pi_i(\mathbf{Y}_i(1) - \mathbf{Y}_i(0))$.

We analyze the behavior over the randomization distribution (1) of the vector-valued SDIM estimator $\hat{\boldsymbol{\tau}} = \frac{1}{N_1}\sum_i D_i\mathbf{Y}_i - \frac{1}{N_0}\sum_i(1-D_i)\mathbf{Y}_i$ and associated variance estimators

$$\hat{\mathbf{s}} := \frac{1}{N_1}\hat{\mathbf{s}}_1 + \frac{1}{N_0}\hat{\mathbf{s}}_0,$$
$$\hat{\mathbf{s}}_1 := \frac{1}{N_1}\sum_i D_i(\mathbf{Y}_i - \bar{\mathbf{Y}}_1)(\mathbf{Y}_i - \bar{\mathbf{Y}}_1)', \quad \hat{\mathbf{s}}_0 := \frac{1}{N_0}\sum_i (1-D_i)(\mathbf{Y}_i - \bar{\mathbf{Y}}_0)(\mathbf{Y}_i - \bar{\mathbf{Y}}_0)',$$

where $\bar{\mathbf{Y}}_1 := \frac{1}{N_1}\sum_i D_i\mathbf{Y}_i$ and $\bar{\mathbf{Y}}_0 := \frac{1}{N_0}\sum_i(1-D_i)\mathbf{Y}_i$.

We introduce the following regularity conditions on the sequence of finite populations.

Assumption B.1. *Suppose $N_1/N \to p_1 \in (0,1)$, and $S_{1,w}, S_{0,w}, S_{10,w}$ have finite limits for $w \in \{\pi, 1-\pi, \tilde{\pi}\}$.*

Assumption B.2. $\max_{1\leqslant i\leqslant N}||\mathbf{Y}_i(1) - \mathbb{E}_\pi[\mathbf{Y}_i(1)]||^2/N \to 0$ *and* $\max_{1\leqslant i\leqslant N}||\mathbf{Y}_i(0) - \mathbb{E}_{1-\pi}[\mathbf{Y}_i(0)]||^2/N \to 0$*, where $||\cdot||$ is the Euclidean norm.*

Assumption B.3. *Let* $\tilde{\mathbf{Y}}_i = \frac{1}{N_1}\mathbf{Y}_i(1) + \frac{1}{N_0}\mathbf{Y}_i(0)$*, and let* λ_{min} *be the minimal eigenvalue of* $\Sigma_{\tilde{\pi}} = \mathbb{V}ar_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]$*. Assume* $\lambda_{min} > 0$ *and for all* $\epsilon > 0$*,*

$$\frac{1}{\lambda_{min}}\mathbb{E}_{\tilde{\pi}}\left[\left\|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right\|^2 \cdot 1\left[\left\|\tilde{\mathbf{Y}}_i - \mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right\| > \sqrt{\sum_i \pi_i(1-\pi_i)\cdot\lambda_{min}}\cdot\epsilon\right]\right] \to 0.$$

Assumption B.1 requires that the fraction of treated units and the (weighted) variance and covariances of the potential outcomes have finite limits along the sequence of finite populations. Assumption B.2 is a multivariate analog of Assumption 4.3 in that it requires that no single observation dominate the π or $(1-\pi)$-weighted variance of the potential outcomes. Assumption B.3 is a multivariate generalization of the Lindeberg-type condition in Assumption 4.2.

Proposition B.1 (Results for vector-valued outcomes)**.**

(1)

$$\begin{aligned}\mathbb{E}_R\left[\hat{\boldsymbol{\tau}}\right] &= \boldsymbol{\tau}_{ATE} + \frac{N}{N_0}\left(\frac{1}{N}\sum_i\left(\pi_i - \frac{N_1}{N}\right)\mathbf{Y}_i(0)\right) + \frac{N}{N_1}\left(\frac{1}{N}\sum_i\left(\pi_i - \frac{N_1}{N}\right)\mathbf{Y}_i(1)\right),\\ &= \boldsymbol{\tau}_{EATT} + \frac{N}{N_0}\frac{N}{N_1}\left(\frac{1}{N}\sum_i\left(\pi_i - \frac{N_1}{N}\right)\mathbf{Y}_i(0)\right).\end{aligned}$$

(2) Under Assumptions 4.1 and B.1,

$$\begin{aligned}\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right] + o(N^{-1}) &= \frac{\frac{1}{N}\sum_{k=1}^N \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}}\left[\frac{1}{N_1}\mathbb{V}ar_{\tilde{\pi}}\left[\mathbf{Y}_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{\tilde{\pi}}\left[\mathbf{Y}_i(0)\right] - \frac{1}{N}\mathbb{V}ar_{\tilde{\pi}}\left[\boldsymbol{\tau}_i\right]\right]\\ &\leqslant \frac{1}{N_1}\mathbb{V}ar_{\pi}\left[\mathbf{Y}_i(1)\right] + \frac{1}{N_0}\mathbb{V}ar_{1-\pi}\left[\mathbf{Y}_i(0)\right],\end{aligned}$$

where $A \leqslant B$ *if* $B - A$ *is positive semi-definite.*

(3) Under Assumptions 4.1, B.1, and B.2,

$$\hat{\mathbf{s}}_1 - \mathbb{V}ar_{\pi}\left[\mathbf{Y}_i(1)\right] \xrightarrow{p} 0, \qquad \hat{\mathbf{s}}_0 - \mathbb{V}ar_{1-\pi}\left[\mathbf{Y}_i(0)\right] \xrightarrow{p} 0.$$

(4) Under Assumptions 4.1, B.1, and B.3,

$$\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right]^{-\frac{1}{2}}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right) \xrightarrow{d} \mathcal{N}\left(0,\, I\right).$$

Assumption B.1 implies $\Sigma_\tau = \lim_{N\to\infty} N\mathbb{V}_R\left[\hat{\boldsymbol{\tau}}\right]$ *exists, so the previous display can alternatively be written as*

$$\sqrt{N}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}\left(0,\, \Sigma_{\boldsymbol{\tau}}\right).$$

Proof. The proof of claim (1) is analogous to the proof of Proposition 3.1 in the scalar case.

We next prove claim (2). For simplicity, let $A_n = \mathbb{V}_R[\hat{\tau}]$, let B_n be the right-hand-side of the first equality in claim (2), and let C_n be the right-hand side of the inequality in claim (2). We first prove the inequality. Note that by the definition of a semi-definite matrix, it suffices to show that $l'B_n l \leqslant l'C_n l$ for all $l \in \mathbb{R}^K$. However, letting $Y_i(d) = l'\mathbf{Y}_i(d)$, the desired inequality follows from Proposition 4.2. Next, observe that $A_n - B_n = o(N^{-1})$ if and only if $D_n := NA_n - NB_n = o(1)$, which holds if and only if $l'D_n l = o(1)$ for all $l \in L := \{e_j \,|\, 1 \leqslant j \leqslant K\} \cup \{e_j - e_{j'} \,|\, 1 \leqslant j, j' \leqslant K\}$, where e_j is the jth basis vector in $\mathbb{R}^K$. To obtain the last equivalence, note that $e_j' D_n e_j = [D_n]_{jj}$ (the (j,j) element of D_n), whereas exploiting the fact that D_n is symmetric, $(e_j - e_{j'})' D_n (e_j - e_{j'}) = [D_n]_{jj} + [D_n]_{j'j'} - 2[D_n]_{jj'}$, and so convergence of $l'D_n l$ to zero for all $l \in L$ is equivalent to convergence of each of the elements of D_n. Next, note that if $Y_i(d) = l'\mathbf{Y}_i(d)$, then $\hat{\tau}$ as defined in (2) is equal to $l'\hat{\boldsymbol{\tau}}$ and $\mathbb{V}\text{ar}_{\tilde{\pi}}[Y_i(d)] = l'\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(d)]\, l$. It follows from Proposition 4.1 that

$$N \cdot l'\mathbb{V}_R[\hat{\boldsymbol{\tau}}]\, l[1+o(1)] = \frac{\frac{1}{N}\sum_{k=1}^N \pi_k(1-\pi_k)}{\frac{N_0}{N}\frac{N_1}{N}} l' \left[\frac{N}{N_1}\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(1)] + \frac{N}{N_0}\mathbb{V}\text{ar}_{\tilde{\pi}}[\mathbf{Y}_i(0)] - \mathbb{V}\text{ar}_{\tilde{\pi}}[\tau_i] \right] l, \tag{21}$$

which implies that $l'D_n l = l'(NA_n)l \cdot o(1)$. However, Assumption B.1, together with the inequality in claim (2), implies that the right-hand side of the previous display is $O(1)$, and thus $l'(NA_n)l = O(1)$, from which the desired result follows.

The proof of claim (3) is similar to the proof of Lemma A3 in Li and Ding (2017), which gives a similar result in the case of completely randomized experiments. We provide a proof for the convergence of $\hat{\boldsymbol{s}}_1$; the convergence of $\hat{\boldsymbol{s}}_0$ is similar. As in the proof to claim (2), it suffices to show that $l'\hat{\boldsymbol{s}}_1 l - l'\mathbb{V}\text{ar}_{\pi}[\mathbf{Y}_i(1)]\, l \to_p 0$ for all $l \in L$. Let $Y_i(d) = l'\mathbf{Y}_i(1)$. Then

$$\begin{aligned} l'\hat{\boldsymbol{s}}_1 l &= \frac{1}{N_1}\sum_i D_i (l'\mathbf{Y}_i(1) - \frac{1}{N_1}\sum_j D_j l'\mathbf{Y}_j(1))^2 \\ &= \left(\frac{1}{N_1}\sum_i D_i (l'\mathbf{Y}_i(1) - l'\mathbb{E}_\pi[\mathbf{Y}_i(1)])^2 \right) + \left(\frac{1}{N_1}\sum_i D_i l'\mathbf{Y}_i(1) - \mathbb{E}_\pi[l'\mathbf{Y}_i(1)] \right)^2, \end{aligned} \tag{22}$$

where the second line uses the bias variance decomposition. The first term can be viewed as a Horvitz-Thompson estimator of $\frac{1}{N_1}\sum_i \pi_i (l'\mathbf{Y}_i(1) - \mathbb{E}_\pi[l'\mathbf{Y}_i(1)])^2 = \mathbb{V}\text{ar}_\pi[l'\mathbf{Y}_i(1)]$ under rejective sampling, and thus has variance equal to

$$(1+o(1))\frac{1}{N_1^2}\left(\sum_i \pi_i(1-\pi_i)\right) \mathbb{V}\text{ar}_{\tilde{\pi}}\left[(l'\mathbf{Y}_i(1) - \mathbb{E}_\pi[l'\mathbf{Y}_i(1)])^2\right].$$

Further, observe that

$$\frac{1}{N_1^2}\left(\sum_i \pi_i(1-\pi_i)\right)\mathbb{V}\text{ar}_{\tilde{\pi}}\left[(l'\mathbf{Y}_i(1)-\mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2\right] \leqslant$$
$$\frac{1}{N_1}\mathbb{E}_\pi\left[(l'\mathbf{Y}_i(1)-\mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^4\right] \leqslant$$
$$\frac{1}{N_1}\max_i\left\{(l'\mathbf{Y}_i(1)-\mathbb{E}_\pi\left[l'\mathbf{Y}_i(1)\right])^2\right\}\cdot\mathbb{V}\text{ar}_\pi\left[l'\mathbf{Y}_i(1)\right] \leqslant$$
$$\left[||l||^2\frac{N}{N_1}\right]\left[\max_i||\mathbf{Y}_i(1)-\mathbb{E}_\pi\left[\mathbf{Y}_i(1)\right]||^2/N\right]\cdot\left[l'\mathbb{V}\text{ar}_\pi\left[\mathbf{Y}_i(1)\right]l\right]=o(1)$$

where the first inequality is obtained using the fact that $\mathbb{V}\text{ar}_{\tilde{\pi}}[X] \leqslant \mathbb{E}_{\tilde{\pi}}[X^2]$, expanding the definition of $\mathbb{E}_{\tilde{\pi}}[\cdot]$, and using the inequality $\pi_i(1-\pi_i) \leqslant \pi_i$, analogous to the argument in the proof to Proposition 4.3 in the scalar case; the final inequality uses the Cauchy-Schwarz inequality and factors out l; and we obtain that the final term is $o(1)$ by noting that the first and final bracketed terms are $O(1)$ by Assumption B.1 and the middle term is $o(1)$ by Assumption B.2. Applying Chebychev's inequality, it follows that the first term in (22) is equal to $\mathbb{V}\text{ar}_\pi\left[l'\mathbf{Y}_i(1)\right]+o(1)$.

To complete the proof of the claim, we show that the second term in (22) is $o(1)$. Note that we can view $\frac{1}{N_1}\sum_i D_i l'\mathbf{Y}_i(1)$ as a Horvitz-Thompson estimator of $\mathbb{E}_\pi\left[l'\mathbf{Y}_i\right]$. Following similar arguments to that in the proceeding paragraph, we have that its variance is bounded above by $\frac{1}{N_1}l'\mathbb{V}\text{ar}_\pi\left[\mathbf{Y}_i(1)\right]l$, which is $o(1)$ by Assumption B.1 combined with the fact that Assumption 4.1 implies $N_1 \to \infty$. Applying Chebychev's inequality again, we obtain that the second term in (22) is $o(1)$, as needed.

To prove claim (4), appealing to the Cramer-Wold device, it suffices to show that for any $l \in \mathbb{R}^K\backslash\{0\}$, $Y_i = l'\mathbf{Y}_i$, and $\hat{\tau}$ as defined in (2), $\mathbb{V}_R\left[\hat{\tau}\right]^{-\frac{1}{2}}(\hat{\tau}-\tau) \to_d \mathcal{N}(0,1)$. This follows from Proposition 4.3, provided that we can show that Assumption B.3 implies that Assumption 4.2 holds when $Y_i = l'\mathbf{Y}_i$ for any conformable vector l. Indeed, recall that $\sigma_{\tilde{\pi}}^2 = l'\Sigma_{\tilde{\pi}}l \geqslant \lambda_{min}||l||^2$, and hence $\frac{1}{\lambda_{min}} \geqslant \frac{1}{||l||^2}\frac{1}{\sigma_{\tilde{\pi}}^2}$. From the Cauchy-Schwarz inequality

$$\left|\left|\tilde{\mathbf{Y}}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right|^2\cdot||l||^2 \geqslant (\tilde{Y}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])^2.$$

Together with the previous inequality, this implies that

$$\frac{1}{\lambda_{min}}\mathbb{E}_{\tilde{\pi}}\left[\left|\left|\tilde{\mathbf{Y}}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right|^2\cdot 1\left[\left|\left|\tilde{\mathbf{Y}}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{\mathbf{Y}}_i\right]\right|\right| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)\cdot\lambda_{min}}\cdot\epsilon\right]\right] \geqslant$$
$$\frac{1}{\sigma_{\tilde{\pi}}^2}\mathbb{E}_{\tilde{\pi}}\left[(\tilde{Y}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])^2\cdot 1\left[\left|(\tilde{Y}_i-\mathbb{E}_{\tilde{\pi}}\left[\tilde{Y}_i\right])\right| \geqslant \sqrt{\sum_i \pi_i(1-\pi_i)}\cdot\sigma_{\tilde{\pi}}\epsilon\right]\right],$$

from which the result follows.

□

B.1 Non-staggered difference-in-differences

We apply the multiple outcomes results to provide a design-based analysis of non-staggered difference-in-differences (DiD) estimators with more than two periods (e.g., Chapter 5 of Angrist and Pischke (2009)), extending those for the two-period DiD model in the main text.

Set-up: Suppose we observe panel data for a finite-population of N units for periods $t = -\underline{T}, \ldots, \bar{T}$. Units with $D_i = 1$ receive a treatment of interest beginning at period $t = 1$.[26] The observed outcome for unit i at period t is $Y_{it} = Y_{it}(D_i)$. We assume the treatment has no effect prior to implementation, so that $Y_{it}(1) = Y_{it}(0)$ for all $t < 1$ (i.e., "no-anticipation"). It is common to estimate the ATT in period t by the difference-in-differences estimator

$$\hat{\beta}_t = \hat{\tau}_t - \hat{\tau}_0 \quad \text{where} \quad \hat{\tau}_t = \frac{1}{N_1}\sum_i D_i Y_{it} - \frac{1}{N_0}\sum_i (1 - D_i) Y_{it}. \tag{23}$$

The DiD estimators $\{\hat{\beta}_t \colon t = 1, \ldots, \bar{T}\}$ correspond with the coefficients from the dynamic two-way fixed effects (TWFE) or "event-study" regression specification

$$Y_{it} = \alpha_i + \phi_t + \sum_{s \neq 0} D_i \times 1[s = t] \times \beta_s + \epsilon_{it}. \tag{24}$$

From equation (23), we see that $\hat{\beta}_t$ is the difference in the SDIM estimators for the outcome in period t and period 0. Letting $\mathbf{Y}_i = (Y_{i,-\underline{T}}, ..., Y_{i,\bar{T}})'$, claim (1) of Proposition B.1 implies

$$\mathbb{E}_R\left[\hat{\beta}_t\right] = \tau_{EATT,t} + \frac{N}{N_0}\frac{N}{N_1}\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0) - Y_{i0}(0)\right],$$

where $\tau_{EATT,t} = \frac{1}{N_1}\sum_i \pi_i (Y_{it}(1) - Y_{it}(0))$ is the EATT in period t, and we use the fact that $\tau_0 = 0$ by the no-anticipation assumption. Thus, the bias in $\hat{\beta}_t$ is proportional to the finite population covariance between π_i and trends in the untreated potential outcomes, $Y_{it}(0) - Y_{i0}(0)$. It follows that $\hat{\beta}_t$ is unbiased for τ_t over the randomization distribution if $\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0) - Y_{i0}(0)\right] = 0$, or equivalently, if

$$\mathbb{E}_R\left[\frac{1}{N_1}\sum_i D_i (Y_{it}(0) - Y_{i0}(0))\right] = \mathbb{E}_R\left[\frac{1}{N_0}\sum_i (1 - D_i)(Y_{it}(0) - Y_{i0}(0))\right],$$

which mimics the familiar "parallel trends" assumption from the sampling-based model.

Additionally, if the sequence of populations satisfies the assumptions in claim (4) of

[26] We focus on the case with non-staggered treatment timing since it may be difficult to causally interpret the estimand of standard two-way fixed effects models under treatment effect heterogeneity and staggered treatment timing (Borusyak and Jaravel, 2016; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Athey and Imbens, 2022). Nonetheless, the results discussed in this section could potentially be extended to other estimators with a more sensible causal interpretation under staggered timing e.g. Callaway and Sant'Anna (2021); Sun and Abraham (2021).

Proposition B.1, then

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - (\boldsymbol{\tau}_{EATT} + \boldsymbol{\delta})) \to_d \mathcal{N}(0, \Sigma), \tag{25}$$

where $\hat{\boldsymbol{\beta}}$ is the vector that stacks the period-specific estimators $\hat{\beta}_t$, $\Sigma = \lim_{N\to\infty} N\mathbb{V}_R\left[\hat{\beta}_t\right]$, and $\boldsymbol{\tau}_{EATT}$, $\boldsymbol{\delta}$ are the vectors that stack $\tau_{EATT,t}$ and $\delta_t = \frac{N}{N_0}\frac{N}{N_1}\mathbb{C}\text{ov}_1\left[\pi_i, Y_{it}(0) - Y_{i0}(0)\right]$. Claim (3) implies that the variance estimator $\hat{\mathbf{s}}$ is asymptotically conservative for $\hat{\boldsymbol{\beta}}$. It is easily verified that $\hat{\mathbf{s}}$ corresponds with the cluster-robust variance estimator for (24) that clusters at level i (up to degrees of freedom corrections). The resulting normal limiting model in (25) has been studied by Roth (2022) and Rambachan and Roth (Forthcoming) from a sampling-based perspective in which parallel trends may fail.[27] These results show that it also has a sensible interpretation from a design-based perspective.

C Extension to general OLS estimators with clustered assignment

This section extends our analysis of the SDIM estimator under the rejective assignment mechanism in two ways. First, we consider general regression estimators beyond the simple difference-in-means. Second, we allow for clustered treatment assignment. This nests our results in the main text on the SDIM under individual-level treatment assignment as a special case where (i) the regression estimator is the SDIM, and (ii) each cluster corresponds with exactly 1 unit.

As in Section 6.1, suppose each unit $i = 1, \ldots, N$ belongs to one of $c = 1, \ldots, C$ clusters, where $c(i)$ denotes the cluster membership of unit i. The treatment is assigned at the cluster level, where the cluster level treatment assignments $D := (D_1, \ldots, D_C)'$ follow a rejective assignment mechanism (12). Suppose that the researcher estimates the ordinary least squares (OLS) coefficients $\hat{\beta}$ from the regression $Y_i = X_i'\beta + \epsilon_i$, where $X_i = D_i X_i(1) + (1 - D_i)X_i(0)$ is a vector of covariates potentially depending on D_i. Note that if $X_i(d) = (1, d)'$, then the second element of $\hat{\beta}$ corresponds with the SDIM.

We analyze the properties of the OLS estimator along a sequence of finite-populations along which the number of clusters C grows large, similar to the asymptotics in Section 4. Before stating our results, we introduce some notation. Let $\widetilde{XX}_c'(d) = \sum_{i:c(i)=c} X_i(d)X_i(d)'$ and $\widetilde{XY}_c(d) = \sum_{i:c(i)=c} X_i(d)Y_i(d)$. Analogous to the notation in the main text, for a cluster-level function of the potential outcome $A_c(d)$, we will write, $\mathbb{E}_{w_c}\left[A_c(d)\right]$ to denote the sum

[27] One difference from the design-based view is that Σ is only conservatively estimable.

$\frac{1}{\sum_c w_c}\sum_c A_c(d)$. Using this notation, $\hat{\beta}$ can be written as

$$\begin{aligned}\hat{\beta} &= \left(\sum_i X_i X_i'\right)^{-1}\left(\sum_i X_i Y_i\right) \\ &= \left(\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{XX}_c'(1) + \frac{C_0}{C}\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XX}_c'(0)\right)^{-1} \times \\ &\quad \left(\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{XY}_c(1) + \frac{C_0}{C}\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XY}_c(0)\right)\end{aligned}$$

We provide the proofs of all results in Section C.1.

Our first result shows $\hat{\beta}$ is consistent for

$$\beta_{cluster} := \left(\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XX}_c'(1)\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XX}_c'(0)\right]\right)^{-1}\left(\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{XY}_c(1)\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{XY}_c(0)\right]\right),$$

and asymptotically normally distributed under the clustered randomization distribution.

Assumption C.1.

(i) (Moments have limits) $\mathbb{E}_{\pi_c}\left[\widetilde{XY}_c(1)\right]$, $\mathbb{E}_{1-\pi_c}\left[\widetilde{XY}_c(0)\right]$, $\mathbb{E}_{\pi_c}\left[\widetilde{XX}_c'(1)\right]$, $\mathbb{E}_{1-\pi_c}\left[\widetilde{XX}_c'(0)\right]$, *and* $\frac{C_1}{C}$ *have finite limits, with* $\lim\frac{C_1}{C} \in (0,1)$.

(ii) (Full-rank regressors) $\frac{C_1}{C}\mathbb{E}_{\pi}\left[\widetilde{XX}_c'(1)\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi}\left[\widetilde{XX}_c'(0)\right]$ *has a full-rank limit.*

(iii) (Bounded variances) There exists $M < \infty$ *such that* $\mathbb{V}ar_{\tilde{\pi}_c}\left[(\widetilde{XX}_c'(d))_{jk}\right] < M$ *and* $\mathbb{V}ar_{\tilde{\pi}_c}\left[(\widetilde{XY}_c(d))_j\right] < M$ *for* $d = 0, 1$ *and* $j, k = 1, ..., dim(X_i)$.

(iv) (Lindeberg condition) Assumption B.3 is satisfied for $\mathbf{Y}_i = \widetilde{X}\epsilon_c(1) - \widetilde{X}\epsilon_c(0) - \mathbb{E}_{\pi_c}\left[\widetilde{X}\epsilon_c(1) - \widetilde{X}\epsilon_c(0)\right]$, *where* $\epsilon_i(d) = Y_i(d) - X_i(d)'\beta_{cluster}$ *and* $\widetilde{X}\epsilon_c(d) = \sum_{i:c(i)=c} X_i(d)\epsilon_i(d)$.

Proposition C.1 (Consistency and asymptotic normality)**.**

(1) If $\sum_c \pi_c(1-\pi_c) \to \infty$ *and Assumption C.1 parts (i)-(iii) hold,* $\hat{\beta} - \beta_{cluster} \xrightarrow{p} 0$.

(2) Define $V_{cluster} := C^{-1}\left(\sum_c \tilde{\pi}_c\right)\mathbb{V}ar_{\tilde{\pi}_c}\left[\sum_{i:\, c(i)=c} X_i(1)\epsilon_i(1) - X_i(0)\epsilon_i(0)\right]$. *If* $\sum_c \pi_c(1-\pi_c) \to \infty$ *and Assumption C.1 holds,*

$$\Omega_{cluster}^{-1/2}\sqrt{C}\left(\hat{\beta} - \beta_{cluster}\right) \xrightarrow{d} \mathcal{N}(0,\, I),$$

where $\Omega_{cluster} := \mathbb{E}_R\left[\frac{1}{C}\sum_i X_i X_i'\right]^{-1} V_{cluster}\mathbb{E}_R\left[\frac{1}{C}\sum_i X_i X_i'\right]^{-1}$.

We next analyze the cluster-robust variance estimator (Liang and Zeger, 1986),

$$\hat{\Omega}_{cluster} := \left(\frac{1}{C}\sum_i X_i X_i'\right)^{-1} \hat{V}_{cluster} \left(\frac{1}{C}\sum_i X_i X_i'\right)^{-1}, \tag{26}$$

where

$$\hat{V}_{cluster} := \frac{1}{C}\sum_c \widetilde{X\hat{\epsilon}}_c \widetilde{X\hat{\epsilon}}_c' \tag{27}$$

for $\hat{\epsilon}_i = Y_i - X_i'\hat{\beta}$ and $\widetilde{X\hat{\epsilon}}_c = \sum_{i:\, c(i)=c} X_i\hat{\epsilon}_i$. In the case with an individual-level treatment assignment (i.e., $C = N$), the cluster-robust variance estimator is equivalent to the Eicker-Huber-White heteroskedasticity-robust variance estimator. Our next result establishes that $\hat{V}_{cluster}$ is consistent for an upper bound of $V_{cluster}$ defined in Proposition C.1 in finite populations with a large number of clusters.

Assumption C.2.

(i) $\mathbb{E}_{\pi_c}\left[\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]$ *and* $\mathbb{E}_{1-\pi_c}\left[\widetilde{X\epsilon}_c(0)\widetilde{X\epsilon}_c(0)'\right]$ *have limits.*

(ii) There exists $\tilde{M}_1 > 0$ *such that* $\|\mathbb{V}ar_{\tilde{\pi}_c}\left[\widetilde{X\epsilon}_c(d)\widetilde{X\epsilon}_c(d)'\right]\| < \tilde{M}_1$ *for* $d = 0, 1$*, where* $\|A\|$ *denotes the Frobenius norm of a matrix* A*.*

(iii) There exists $\tilde{M}_2 > 0$ *such that* $\mathbb{E}_1\left[\|\widetilde{X\epsilon}_c(d)\|^2\right] < \tilde{M}_2$ *and* $\mathbb{E}_1\left[\|\widetilde{XX}_c'(d)\|^2\right] < \tilde{M}_2$ *for* $d = 0, 1$*.*

Proposition C.2 (Variance consistency)**.** *If Assumption C.1(i)-(iii) and Assumption C.2 hold, and* $\sum_c \pi_c(1-\pi_c) \to \infty$*, then* $\hat{V}_{cluster} - V^{est}_{cluster} \xrightarrow{p} 0$ *for*

$$V^{est}_{cluster} := \frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{X}\epsilon_c(1)\widetilde{X}\epsilon_c(1)'\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{X}\epsilon_c(0)\widetilde{X}\epsilon_c(0)'\right]$$

Furthermore, $V^{est}_{cluster} \geqslant V_{cluster}$ *(i.e.,* $V^{est}_{cluster} - V_{cluster}$ *is positive semi-definite).*

Corollary C.1. *Define* $\Omega^{est}_{cluster} := \mathbb{E}_R\left[\sum_i X_i X_i\right]^{-1} V^{est}_{cluster} \mathbb{E}_R\left[\sum_i X_i X_i\right]^{-1}$*. Under the same conditions as Proposition C.2,* $\hat{\Omega}_{cluster} - \Omega^{est}_{cluster} \xrightarrow{p} 0$*, and* $\Omega^{est}_{cluster} \geqslant \Omega_{cluster}$*.*

Finally, we show that the Eicker-Huber-White (EHW) covariance estimator need not be valid under the clustered treatment assignment mechanism considered here. Specifically, consider the Eicker-Huber-White variance estimator $\hat{V}_{EHW} = \frac{1}{N}\sum_i X_i X_i'\hat{\epsilon}_i^2$. Under the clustered treatment assignment mechanism, it can be equivalently rewritten as

$$\hat{V}_{EHW} = \frac{C_1}{N}\frac{1}{C_1}\sum_c D_c\left(\widetilde{XX'\hat{\epsilon}^2}_c(1)\right) + \frac{C_0}{N}\frac{1}{C_0}\sum_c (1-D_c)\left(\widetilde{XX'\hat{\epsilon}^2}_c(0)\right),$$

where $\widetilde{XX'\hat{\epsilon}^2}_c(d) = \sum_{i:\, c(i)=c} X_i(d)X_i(d)'\hat{\epsilon}_i^2$. Define $\widetilde{XX'\epsilon^2}_c(d) = \sum_{i:\, c(i)=c} X_i(d)X_i(d)'\epsilon_i(d)^2$ analogously. Our next result characterizes the probability limit of $\hat{V}_{EHW}$.

Assumption C.3.

(i) $\mathbb{E}_{\pi_c}\left[\widetilde{XX'\epsilon^2}_c(1)\right]$, $\mathbb{E}_{1-\pi_c}\left[\widetilde{XX'\epsilon^2}_c(0)\right]$, N/C, C_1/C *have finite limits with* $\lim C_1/C \in (0,1)$ *and* $\lim N/C < \infty$.

(ii) There exists $\tilde{M}_3$ *such that* $\|\mathbb{V}ar_{\tilde{\pi}_c}\left[\widetilde{XX'\epsilon^2}_c(d)\right]\| \leqslant \tilde{M}_3$ *for* $d = 0, 1$.

(iii) There exists $\tilde{M}_4$ *such that* $\mathbb{E}_1\left[\widetilde{W(d)}_c\right] < \tilde{M}_4$ *and* $\mathbb{E}_1\left[\widetilde{V(d)}_c\right] < \tilde{M}_4$ *for* $d = 0, 1$, *where* $\widetilde{W(d)}_c = \sum_{i:\, c(i)=c} \|X_i(1)\epsilon_i(d)\|^2$ *and* $\widetilde{V(d)}_c = \sum_{i:\, c(i)=c} \|X_i(d)X_i(d)'\|^2$.

Proposition C.3. *If Assumptions C.1 and C.3(i)-(iii) hold, and* $\sum_c \pi_c(1-\pi_c) \to \infty$, *then* $\hat{V}_{EHW} - V^{EHW}_{cluster} \xrightarrow{p} 0$ *for*

$$V^{EHW}_{cluster} := \frac{C_1}{N}\mathbb{E}_{\pi_c}\left[\widetilde{XX'\epsilon^2}_c(1)\right] + \frac{C_0}{N}\mathbb{E}_{1-\pi_c}\left[\widetilde{XX'\epsilon^2}_c(0)\right].$$

Furthermore, $V_{cluster} - \frac{N}{C}V^{EHW}_{cluster}$ *equals*

$$\frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\sum_{i\neq j:\, c(i),c(j)=c} \eta_i(1)\eta_j(1)'\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\sum_{i\neq j:\, c(i),c(j)=c} \eta_i(0)\eta_j(0)'\right] -$$

$$\mathbb{E}_1\left[(\pi_c\eta_c(1) + (1-\pi_c)\eta_c(0))(\pi_c\eta_c(1) + (1-\pi_c)\eta_c(0))'\right] - \mathbb{E}_1\left[\tilde{\pi}_c\right]\mathbb{E}_{\tilde{\pi}_c}\left[\eta_c(1) - \eta_c(0)\right]\mathbb{E}_{\tilde{\pi}_c}\left[\eta_c(1) - \eta_c(0)\right]'$$

where $\eta_i(d) = X_i(d)\epsilon_i(d)$ *and* $\eta_c(d) = \sum_{i:\, c(i)=c} \eta_i(d)$.

Proposition C.3 implies that the usual heteroskedasticity-robust variance estimator can be invalid in large populations if there is clustered treatment assignment (i.e. if $N \neq C$). To see this, consider the SDIM, which corresponds with $X_i = (1, D_i)'$. Suppose there is no within-cluster heterogeneity in potential outcomes (i.e., $Y_i(d) = Y_{c(i)}(d)$ for all i and $d \in \{0,1\}$) and all clusters are the same size (i.e., $N_c = N/C$). In this case, $V^{est}_{cluster} = \frac{N}{C}V^{EHW}_{cluster}$. If further there is no across-cluster treatment effect heterogeneity nor heterogeneity in cluster-specific treatment probabilities, $V_{cluster} = V^{est}_{cluster}$ by the same logic as Corollary 4.1 in the main text for the non-clustered case, and the heteroskedasticity-robust variance estimator is thus too small whenever $N/C > 1$. If there is either treatment effect heterogeneity or heterogeneity in cluster-specific treatment probabilities, then $V_{cluster} \leqslant V^{est}_{cluster}$ (generally with strict inequality), in which case the heteroskedasticity-robust variance estimator is valid whenever $C/N \geqslant V_{cluster}/V^{est}_{cluster}$. Abadie et al. (2022) establish a similar result for a setting in which units have the same probability of receiving treatment marginalized over a two-stage assignment process; thus treatment probabilities in Abadie et al. (2022) are not related to potential outcomes, and so their calculations are not directly applicable to quasi-experimental settings.

C.1 Proofs of results for general OLS estimators under clustering

Proof of Proposition C.1

Proof. To establish claim (1), let p_c^* be the limit of $\frac{C_1}{C}$, let $\mu_{\pi_c}\left[\widetilde{XX'_c}(1)\right]$ be the limit of $\mathbb{E}_{\pi_c}\left[\widetilde{XX'_c}(1)\right]$, and define $\mu_{\pi_c}[\cdot]$ and $\mu_{1-\pi_c}[\cdot]$ of other variables analogously. Let

$$\beta^*_{cluster} = \left(p_c^*\mu_{\pi_c}\left[\widetilde{XX'_c}(1)'\right] + (1-p_c^*)\mu_{1-\pi_c}\left[\widetilde{XX'_c}(0)'\right]\right)^{-1}\left(p_c^*\mu_{\pi_c}\left[\widetilde{XY_c}(1)\right] + (1-p_c^*)\mu_{1-\pi_c}\left[\widetilde{XY_c}(0)\right]\right).$$

It is immediate from Assumption C.1(i)-(ii) that $\beta_{cluster} \to \beta^*_{cluster}$, so it suffices to show that $\hat{\beta} \xrightarrow{p} \beta^*_{cluster}$. Note that we can write $\hat{\beta}$ as

$$\left(\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{XX'}(1) + \frac{C_0}{C}\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XX'_c}(0)\right)^{-1}\left(\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{XY_c}(1) + \frac{C_0}{C}\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XY_c}(0)\right).$$

Using Theorem 6.1 in Hajek (1964) as in the proof to Proposition 4.1, we have that

$$\begin{aligned}\mathbb{V}\mathrm{ar}_R\left[\frac{1}{C_1}\sum_c D_c(\widetilde{XX'_c}(1))_{jk}\right] &= (1+o(1))C_1^{-2}\left(\sum_c \tilde{\pi}_c\right)\mathbb{V}\mathrm{ar}_{\tilde{\pi}_c}\left[(\widetilde{XX'_c}(1))_{jk}\right] \\ &\leqslant (1+o(1))C_1^{-1}M \to 0,\end{aligned}$$

where we obtain the inequality from Assumption C.1(iii) combined with the fact that $\tilde{\pi}_c \leqslant \pi_c$ for all c and thus $\sum_c \tilde{\pi}_c \leqslant \sum_c \pi_c = C_1$. Combining the previous display with Chebychev's inequality, we obtain that $\frac{1}{C_1}\sum_c D_c\widetilde{XX'_c}(1) - \mathbb{E}_R\left[\frac{1}{C_1}\sum_c D_c\widetilde{XX'_c}(1)\right] \xrightarrow{p} 0$. But $\mathbb{E}_R\left[\frac{1}{C_1}\sum_c D_c\widetilde{XX'_c}(1)\right] = \mathbb{E}_{\pi_c}\left[\widetilde{XX'_c}(1)\right] \to \mu_{\pi_c}\left[\widetilde{XX'_c}(1)\right]$, and hence $\frac{1}{C_1}\sum_c D_c\widetilde{XX'_c}(1) \xrightarrow{p} \mu_{\pi_c}\left[\widetilde{XX'_c}(1)\right]$. An analogous argument yields that $\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XX'_c}(0) \xrightarrow{p} \mu_{1-\pi_c}\left[\widetilde{XX'_c}(0)\right]$, $\frac{1}{C_1}\sum_c D_c\widetilde{XY_c}(1) \xrightarrow{p} \mu_{\pi_c}\left[\widetilde{XY_c}(1)\right]$, and $\frac{1}{C_0}\sum_c (1-D_c)\widetilde{XY_c}(0) \xrightarrow{p} \mu_{1-\pi_c}\left[\widetilde{XY_c}(0)\right]$. These convergences together with the continuous mapping theorem yield that $\hat{\beta} \xrightarrow{p} \beta^*_{cluster}$, as we wished to show.

To show the second claim, define $\epsilon_i = D_i\epsilon_i(1) + (1-D_i)\epsilon_i(0)$ (and recall that $\epsilon_i(d) = Y_i(d) - X_i(d)'\beta_{cluster}$), so that

$$\hat{\beta} = \beta_{cluster} + \left(\frac{1}{C}\sum_i X_iX_i'\right)^{-1}\left(\frac{1}{C}\sum_i X_i\epsilon_i\right).$$

and

$$\sqrt{C}(\hat{\beta} - \beta_{cluster}) = \left(\frac{1}{C}\sum_i X_iX_i'\right)^{-1}\left(\frac{1}{\sqrt{C}}\sum_i X_i\epsilon_i\right).$$

In the proof of claim (1), we established that $\left(\frac{1}{C}\sum_i X_iX_i'\right)^{-1}$ is consistent for $\mathbb{E}_R\left[\frac{1}{C}\sum_i X_iX_i'\right]^{-1}$. We therefore focus on establishing the asymptotic normality of $\frac{1}{\sqrt{C}}\sum_i X_i\epsilon_i$. Towards this, notice that standard arguments for linear projections imply that

$$\mathbb{E}_R\left[\frac{1}{C}\sum_i X_i\epsilon_i\right] = \frac{C_1}{C}\mathbb{E}_{\pi_c}\left[\widetilde{X\epsilon_c}(1)\right] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}\left[\widetilde{X\epsilon_c}(0)\right] = 0, \tag{28}$$

where $\widetilde{X\epsilon_c}(d) = \sum_{i:\, c(i)=c} X_i(d)\epsilon_i(d)$ as before. By adding/subtracting $C_1\mathbb{E}_{\pi_c}\left[\widetilde{X\epsilon_c}(0)\right]$ from the previous display and applying the identity $C_1\mathbb{E}_{\pi_c}[v_c] + C_0\mathbb{E}_{1-\pi_c}[v_c] = C\mathbb{E}_1[v_c]$ for any cluster-level attribute v_c, we obtain that

$$C_1\mathbb{E}_{\pi_c}\left[\widetilde{X\epsilon_c}(1) - \widetilde{X\epsilon_c}(0)\right] + \sum_c \widetilde{X\epsilon_c}(0) = 0.$$

It therefore follows that

$$\begin{aligned}\sum_i X_i\epsilon_i &= \sum_c D_c\widetilde{X\epsilon_c}(1) + \sum_c (1-D_c)\widetilde{X\epsilon_c}(0) \\ &= \sum_c D_c\left(\left(\widetilde{X\epsilon_c}(1) - \widetilde{X\epsilon_c}(0)\right) - \mathbb{E}_{\pi_c}\left[\widetilde{X\epsilon_c}(1) - \widetilde{X\epsilon_c}(0)\right]\right)\end{aligned}$$

Therefore, $\sum_i X_i\epsilon_i$ can be represented as Horvitz-Thompson estimator under clustered rejective sampling. Applying the multivariate generalization of Theorem 1 in Berger (1998) as in the proof to Proposition 4, we therefore conclude that

$$V_{cluster}^{-1/2}\frac{1}{\sqrt{C}}\sum_i X_i\epsilon_i \xrightarrow{d} \mathcal{N}(0,\, I),$$

where $V_{cluster}$ is defined in the statement of claim (2). Claim (2) follows by applying Slutsky's lemma. □

Proof of Proposition C.2

Proof. To show the first claim, observe that

$$\hat{V}_{cluster} = \frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{X\hat{\epsilon}_c}(1)\widetilde{X\hat{\epsilon}_c}(1)' + \frac{C_0}{C}\frac{1}{C_0}\sum_c (1-D_c)\widetilde{X\hat{\epsilon}_c}(0)\widetilde{X\hat{\epsilon}_c}(0)'.$$

Furthermore, $\widetilde{X\hat{\epsilon}_c}(d) = \widetilde{X\epsilon_c}(d) - \widetilde{XX}'_c(d)(\hat{\beta} - \beta_{cluster})$. It follows that

$$\begin{aligned}\frac{1}{C_1}\sum_c D_c\widetilde{X\hat{\epsilon}_c}(1)\widetilde{X\hat{\epsilon}_c}(1)' = \underbrace{\frac{1}{C_1}\sum_c D_c\widetilde{X\epsilon_c}(1)\widetilde{X\epsilon_c}(1)'}_{=(A)} - \\ \underbrace{\frac{1}{C_1}\sum_c D_c\widetilde{X}\epsilon_c(1)(\hat{\beta} - \beta_{cluster})'\widetilde{XX}'_c(1)'}_{=(B)} - \underbrace{\frac{1}{C_1}\sum_c D_c\left(\widetilde{X}\epsilon_c(1)(\hat{\beta} - \beta_{cluster})'\widetilde{XX}'_c(1)'\right)'}_{=(B')} + \\ \underbrace{\frac{1}{C_1}\sum_c D_c\widetilde{XX}'_c(1)(\hat{\beta} - \beta_{cluster})(\hat{\beta} - \beta_{cluster})'\widetilde{XX}'_c(1)'}_{=(C)}\end{aligned} \tag{29}$$

Consider the term labeled (A) in (29) and observe that

$$\left\|\mathbb{V}_R\left[\frac{1}{C_1}\sum_c D_c\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]\right\| = (1+o(1))C_1^{-2}(\sum_c \tilde{\pi}_c)\left\|\mathbb{V}\text{ar}_{\tilde{\pi}_c}\left[\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]\right\|$$
$$\leqslant (1+o(1))C_1^{-1}\tilde{M}_1 \to 0,$$

where we use Assumption C.2(ii) to bound $||\mathbb{V}\text{ar}_{\tilde{\pi}_c}\left[\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]||$. Hence, by Chebychev's inequality, $\frac{1}{C_1}\sum_c D_c\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)' \xrightarrow{p} \mu_{\pi_c}\left[\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]$, where we define $\mu_{\pi_c}[\cdot]$ as in the proof to Proposition C.1. Next, consider the term labeled (C) in (29). Recall that the Frobenius norm is sub-multiplicative, so that $\|QR\| \leqslant \|Q\|\|R\|$ for any matrices Q, R. Hence, we have that

$$\begin{aligned}
\|(C)\| &\leqslant \frac{1}{C_1}\sum_c D_c||\widetilde{XX}'_c(1)(\hat{\beta}-\beta_{cluster})(\hat{\beta}-\beta_{cluster})'\widetilde{XX}'_c(1)'|| \\
&\leqslant ||(\hat{\beta}-\beta_{cluster})(\hat{\beta}-\beta_{cluster})'||\frac{1}{C_1}\sum_c D_c||\widetilde{XX}'_c(1)||^2 \\
&\leqslant ||(\hat{\beta}-\beta_{cluster})(\hat{\beta}-\beta_{cluster})'||\frac{C}{C_1}\frac{1}{C}\sum_c ||\widetilde{XX}'_c(1)||^2 \\
&\leqslant ||(\hat{\beta}-\beta_{cluster})(\hat{\beta}-\beta_{cluster})'||\frac{C}{C_1}\tilde{M}_2 \xrightarrow{p} 0
\end{aligned}$$

where the last inequality uses Assumption C.2(iii), and we use the fact that C/C_1 has a finite limit by Assumption C.1(i) and $\hat{\beta}-\beta_{cluster} \xrightarrow{p} 0$ by Proposition C.1. Finally,

$$\begin{aligned}
\|(B)\| &\leqslant \frac{1}{C_1}\sum_c D_c||\widetilde{X\epsilon}_c(1)(\hat{\beta}-\beta_{cluster})'\widetilde{XX'}_c(1)'|| \\
&\leqslant \frac{1}{C_1}\sum_c D_c||\widetilde{X\epsilon}_c(1)||\cdot||\widetilde{XX}'_c(1)||\cdot||(\hat{\beta}-\beta_{cluster})|| \\
&\leqslant \frac{C}{C_1}\frac{1}{C}\sum_c ||\widetilde{X\epsilon}_c(1)||\cdot||\widetilde{XX}'_c(1)||\cdot||(\hat{\beta}-\beta_{cluster})|| \\
&\leqslant \frac{C_1}{C}\sqrt{\frac{1}{C}\sum_c ||\widetilde{X\epsilon}_c(1)||^2}\cdot\sqrt{\frac{1}{C}\sum_c ||\widetilde{XX}'_c(1)||^2}\cdot||(\hat{\beta}-\beta_{cluster})|| \\
&\leqslant \frac{C_1}{C}\tilde{M}_2||\hat{\beta}-\beta_{cluster}|| \xrightarrow{p} 0,
\end{aligned}$$

where the fourth inequality uses Cauchy-Schwarz, the fifth inequality uses Assumption C.2(iii) and we use the fact that $\hat{\beta}-\beta_{cluster} \xrightarrow{p} 0$ as shown above. We have thus shown that $\frac{1}{C_1}\sum_c D_c\widetilde{X\hat{\epsilon}}_c(1)\widetilde{X\hat{\epsilon}}_c(1)' \xrightarrow{p} \mu_{\pi_c}\left[\widetilde{X\epsilon}_c(1)\widetilde{X\epsilon}_c(1)'\right]$. By analogous argument, we can show that $\frac{1}{C_0}\sum_c(1-D_c)\widetilde{X\hat{\epsilon}}_c(0)\widetilde{X\hat{\epsilon}}_c(0)' \xrightarrow{p} \mu_{1-\pi_c}\left[\widetilde{X\epsilon}_c(0)\widetilde{X\epsilon}_c(0)'\right]$. The first part of the result then follows from the continuous mapping theorem.

To show the second claim, let $\eta_c(d) = \sum_{i:c(i)=c} X_i(d)\epsilon_i(d)$, $\dot{\eta}_c(1) = \dot{\eta}_c(1) - \mathbb{E}_{\pi_c}[\eta_c(1)]$, and $\dot{\eta}_c(0) = \dot{\eta}_c(0) - \mathbb{E}_{1-\pi_c}[\eta_c(0)]$. Then,

$$
\begin{aligned}
V_{cluster} =& \frac{1}{C}\sum_c \pi_c(1-\pi_c)\left(\eta_c(1)-\eta_c(0)-\mathbb{E}_{\tilde{\pi}_c}[\eta_c(1)-\eta_c(0)]\right)\left(\eta_c(1)-\eta_c(0)-\mathbb{E}_{\tilde{\pi}_c}[\eta_c(1)-\eta_c(0)]\right)' \\
\leqslant& \frac{1}{C}\sum_c \pi_c(1-\pi_c)\left(\dot{\eta}_c(1)-\dot{\eta}_c(0)\right)\left(\dot{\eta}_c(1)-\dot{\eta}_c(0)\right)' \\
=& \frac{1}{C}\left(\sum_c \pi_c\dot{\eta}_c(1)\dot{\eta}_c(1)' + \sum_c (1-\pi_c)\dot{\eta}_c(0)\dot{\eta}_c(0)' - \right. \\
& \qquad \left. \left(\sum_c \pi_c^2\dot{\eta}_c(1)\dot{\eta}_c(1)' + \sum_c (1-\pi_c)^2\dot{\eta}_c(0)\dot{\eta}_c(0)' + \sum_c \pi_c(1-\pi_c)(\dot{\eta}_c(1)\dot{\eta}_c(0)' + \dot{\eta}_c(0)\dot{\eta}_c(1)')\right)\right) \\
=& \frac{C_1}{C}\mathbb{V}\mathrm{ar}_{\pi_c}[\eta_c(1)] + \frac{C_0}{C}\mathbb{V}\mathrm{ar}_{1-\pi_c}[\eta_c(0)] - \frac{1}{C}\sum_c (\pi_c\dot{\eta}_c(1)+(1-\pi_c)\dot{\eta}_c(0))(\pi_c\dot{\eta}_c(1)+(1-\pi_c)\dot{\eta}_c(0))' \\
\leqslant& \frac{C_1}{C}\mathbb{E}_{\pi_c}[\eta_c(1)\eta_c(1)'] + \frac{C_0}{C}\mathbb{E}_{1-\pi_c}[\eta_c(0)\eta_c(0)'] = V_{cluster}^{est}.
\end{aligned}
$$

□

Proof of Corollary C.1

Proof. The proof is immediate from Proposition C.2 combined with the fact that $\frac{1}{C}\sum_i X_i X_i' - \mathbb{E}_R\left[\frac{1}{C}\sum_i X_i X_i'\right] \xrightarrow{p} 0$ as shown in the proof to Proposition C.1. □

Proof of Proposition C.3

Proof. To show the first claim, it is immediate from Assumption C.3(i) that $V_{cluster}^{EHW}$ converges to

$$(1/n_c^*)p_c^*\mu_{\pi_c}[\widetilde{XX'\epsilon^2}_c(1)] + (1/n_c^*)(1-p_c^*)\mu_{1-\pi_c}[\widetilde{XX'\epsilon^2}_c(0)],$$

where $n_c^* = \lim N/C$, $p_c^* = \lim C_1/C$, and $\mu_{\pi_c}[\cdot]$ is defined as in the proof to Proposition C.1. It therefore suffices to show that $\hat{V}_{EHW}$ converges in probability to the same limit. To show this, recall that $\hat{\epsilon}_i = D_i\hat{\epsilon}_i(1) + (1-D_i)\hat{\epsilon}_i(0)$ for $\hat{\epsilon}_i(d) = \epsilon_i(d) - X_i(d)'(\hat{\beta} - \beta_{cluster})$ and $X_i(d)\hat{\epsilon}_i(d) = X_i(d)\epsilon_i(d) - X_i(d)X_i(d)'(\hat{\beta} - \beta_{cluster})$. Therefore, we can write $\frac{C_1}{N}\frac{1}{C_1}\sum_c D_c\left(\widetilde{XX'\hat{\epsilon}^2}_c(1)\right)$ as

$$
\underbrace{\frac{C}{N}\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\widetilde{XX'\epsilon^2}_c(1)}_{(A)} + \underbrace{\frac{C}{N}\frac{1}{C}\sum_c D_c \sum_{i:\, c(i)=c} X_i(1)\epsilon_i(1)(\hat{\beta}-\beta_{cluster})'X_i(1)X_i(1)' +}_{(B)}
$$

$$
\underbrace{\frac{C}{N}\frac{1}{C}\sum_c D_c \sum_{i:\, c(i)=c} X_i(1)X_i(1)'(\hat{\beta}-\beta_{cluster})X_i'(1)\epsilon_i(1) +}_{(B\prime)}
$$

$$\underbrace{\frac{C}{N}\frac{1}{C}\sum_c D_c\left(\sum_{i:\,c(i)=c} X_i(1)X_i'(1)(\hat{\beta}-\beta_{cluster})(\hat{\beta}-\beta_{cluster})'X_i(1)X_i'(1)\right)}_{(C)}.$$

First, consider the term (A), and observe that

$$\begin{aligned}\left\|\mathbb{V}_R\left[\frac{1}{C_1}\sum_c D_c\widetilde{XX'\epsilon^2}_c(1)\right]\right\| &= (1+o(1))C_1^{-2}\left(\sum_c \tilde{\pi}_c\right)\left\|\mathbb{V}\mathrm{ar}_{\tilde{\pi}_c}\left[\widetilde{XX'\epsilon^2}_c(1)\right]\right\| \\ &\leqslant (1+o(1))C_1^{-1}\tilde{M}_3 \to 0,\end{aligned}$$

where we use Assumption C.3(ii) to bound $\left\|\mathbb{V}\mathrm{ar}_{\tilde{\pi}_c}\left[\widetilde{XX'\epsilon^2}_c(1)\right]\right\|$. Hence, $\frac{1}{C_1}\sum_c D_c\widetilde{XX'\epsilon^2}_c(1) \xrightarrow{p} \mu_{\pi_c}\left[\widetilde{XX'\epsilon^2}_c\right]$ by Chebyshev's Inequality. Next, consider term (B) and observe that

$$\begin{aligned}\|(B)\| &\leqslant \frac{1}{C}\sum_c D_c \sum_{i:\,c(i)=c}\|X_i(1)\epsilon_i(1)(\hat{\beta}-\beta_{cluster})'X_i(1)X_i(1)'\| \\ &\leqslant \|\hat{\beta}-\beta_{cluster}\|\left(\frac{1}{C}\sum_c D_c\sum_{i:\,c(i)=c}\|X_i(1)\epsilon_i(1)\|\|X_i(1)X_i(1)'\|\right) \\ &\leqslant \|\hat{\beta}-\beta_{cluster}\|\left(C^{-1}\sum_c \widetilde{W(1)}_c\widetilde{V(1)}_c\right) \\ &\leqslant \|\hat{\beta}-\beta_{cluster}\|\sqrt{C^{-1}\sum_c \widetilde{W(1)}_c}\sqrt{C^{-1}\sum_c \widetilde{V(1)}_c} \\ &\leqslant \|\hat{\beta}-\beta_{cluster}\|\tilde{M}_4\end{aligned}$$

where the first inequality applies the triangle inequality, the second inequality applies the submultiplicative property of the Frobenius norm, the third inequality uses the positivity of the norm, and the fourth inequality uses the Cauchy-Schwarz inequality. Since $\hat{\beta}-\beta_{cluster} \xrightarrow{0}$, it follows that $\|(B)\| \xrightarrow{p} 0$ by Assumption C.3(iii). The analogous argument gives that (B')

converges in probability to zero. Finally, consider term (C) and observe that

$$\begin{aligned}
\|(C)\| &\leqslant \frac{1}{C_1}\sum_c D_c \sum_{i\colon c(i)=c} \|X_i(1)X_i'(1)(\hat\beta-\beta_{cluster})(\hat\beta-\beta_{cluster})'X_i(1)X_i'(1)\| \\
&\leqslant \|(\hat\beta-\beta_{cluster})(\hat\beta-\beta_{cluster})'\| \left(\frac{1}{C_1}\sum_c D_c \sum_{i\colon c(i)=c} \|X_i(1)X_i'(1)\|^2\right) \\
&= \|(\hat\beta-\beta_{cluster})(\hat\beta-\beta_{cluster})'\| \left(\frac{1}{C_1}\sum_c D_c \widetilde{V(d)}_c\right) \\
&\leqslant \|(\hat\beta-\beta_{cluster})(\hat\beta-\beta_{cluster})'\| \frac{C}{C_1}\left(\frac{1}{C}\sum_c \widetilde{V(d)}_c\right) \\
&\leqslant \|(\hat\beta-\beta_{cluster})(\hat\beta-\beta_{cluster})'\| \frac{C}{C_1}\tilde M_4,
\end{aligned}$$

which converges in probability to zero since $\hat\beta-\beta_{cluster} \xrightarrow{p} 0$ and $\frac{C_1}{C}$ has a finite limit. Putting this together, it follows that $\frac{C}{N}\frac{C_1}{C}\frac{1}{C_1}\sum_c D_c\left(\widetilde{XX'\hat\epsilon^2}_c(1)\right) \xrightarrow{p} (1/n_c^*)p_c^*\mu_{\pi_c}[\widetilde{XX'\epsilon^2}_c(1)]$ by the continuous mapping theorem. By the same argument, we can show $\frac{C}{N}\frac{C_0}{C}\frac{1}{C_0}\sum_c(1-D_c)\left(\widetilde{XX'\hat\epsilon^2}_c(0)\right) \xrightarrow{p} (1/n_c^*)(1-p_c^*)\mu_{1-\pi_c}[\widetilde{XX'\epsilon^2}_c(0)]$. The first claim then follows by another application of the continuous mapping theorem.

To show the second claim, we first observe that $V_{cluster}$ can be expanded into

$$C^{-1}\sum_c \pi_c(1-\pi_c)\left(\eta_c(1)-\eta_c(0)-\mathbb{E}_{\tilde\pi_c}\left[\eta_c(1)-\eta_c(0)\right]\right)\left(\eta_c(1)-\eta_c(0)-\mathbb{E}_{\tilde\pi_c}\left[\eta_c(1)-\eta_c(0)\right]\right)' =$$

$$\underbrace{C^{-1}\sum_c \pi_c(1-\pi_c)(\eta_c(1)-\eta_c(0))(\eta_c(1)-\eta_c(0))'}_{(a)} - \left(C^{-1}\sum_c \tilde\pi_c\right)\mathbb{E}_{\tilde\pi_c}\left[\eta_c(1)-\eta_c(0)\right]\mathbb{E}_{\tilde\pi_c}\left[\eta_c(1)-\eta_c(0)\right]'.$$

Further expanding out, notice that (a) equals

$$C^{-1}\sum_c \pi_c(1-\pi_c)\left(\eta_c(1)\eta_c(1)'+\eta_c(0)\eta_c(0)'-\eta_c(1)\eta_c(0)'-\eta_c(0)\eta_c(1)'\right) =$$

$$C^{-1}\sum_c \pi_c\eta_c(1)\eta_c(1)' + C^{-1}\sum_c(1-\pi_c)\eta_c(0)\eta_c(0)' -$$

$$C^{-1}\sum_c\left(\pi_c^2\eta_c(1)\eta_c(1)'+(1-\pi_c)^2\eta_c(0)\eta_c(0)'+\pi_c(1-\pi_c)(\eta_c(1)\eta_c(0)'+\eta_c(0)\eta_c(1)')\right) =$$

$$\underbrace{C^{-1}\sum_c \pi_c\eta_c(1)\eta_c(1)' + C^{-1}\sum_c(1-\pi_c)\eta_c(0)\eta_c(0)'}_{(b)} - C^{-1}\sum_c(\pi_c\eta_c(1)+(1-\pi_c)\eta_c(0))(\pi_c\eta_c(1)+(1-\pi_c)\eta_c(0))'.$$

Then, using the identity $\eta_c(d)\eta_c(d)' = \sum_{i\colon c(i)=c}\sum_{j\colon c(j)=c}\eta_i(d)\eta_j(d)' = \sum_{i\colon c(i)=c}\eta_i(d)\eta_i(d)' +$

$\sum_{i \neq j : c(i), c(j) = c} \eta_i(d) \eta_j(d)'$, we further expand out (b) as

$$C^{-1} \sum_c \pi_c \eta_c(1) \eta_c(1)' + C^{-1} \sum_c (1 - \pi_c) \eta_c(0) \eta_c(0)' =$$

$$C^{-1} \sum_c \pi_c \sum_{i \colon c(i) = c} \eta_i(1) \eta_i(1)' + C^{-1} \sum_c (1 - \pi_c) \sum_{i \colon c(i) = c} \eta_i(0) \eta_i(0)' +$$

$$C^{-1} \sum_c \pi_c \sum_{i \neq j \colon c(i), c(j) = c} \eta_i(1) \eta_j(1)' + C^{-1} \sum_c (1 - \pi_c) \sum_{i \neq j \colon c(i), c(j) = c} \eta_i(0) \eta_j(0)' =$$

$$\frac{N}{C} V_{cluster}^{EHW} + \frac{C_1}{C} \mathbb{E}_{\pi_c} \left[\sum_{i \neq j \colon c(i), c(j) = c} \eta_i(1) \eta_j(1)' \right] + \frac{C_0}{C} \mathbb{E}_{1 - \pi_c} \left[\sum_{i \neq j \colon c(i), c(j) = c} \eta_i(0) \eta_j(0)' \right].$$

Putting this altogether, we therefore have shown that $V_{cluster}$ equals

$$\frac{N}{C} V_{cluster}^{EHW} + \frac{C_1}{C} \mathbb{E}_{\pi_c} \left[\sum_{i \neq j \colon c(i), c(j) = c} \eta_i(1) \eta_j(1)' \right] + \frac{C_0}{C} \mathbb{E}_{1 - \pi_c} \left[\sum_{i \neq j \colon c(i), c(j) = c} \eta_i(0) \eta_j(0)' \right] -$$

$$\mathbb{E}_1 \left[(\pi_c \eta_c(1) + (1 - \pi_c) \eta_c(0)) (\pi_c \eta_c(1) + (1 - \pi_c) \eta_c(0))' \right] - \mathbb{E}_1 \left[\tilde{\pi}_c \right] \mathbb{E}_{\tilde{\pi}_c} \left[\eta_c(1) - \eta_c(0) \right] \mathbb{E}_{\tilde{\pi}_c} \left[\eta_c(1) - \eta_c(0) \right]'.$$

□

D Additional Monte Carlo simulations

This appendix considers extensions to the simulations in Section 5, where (i) the number of treated units varies, (ii) there is treatment effect heterogeneity, and (iii) the size of the finite population varies.

D.1 Varying the number of treated units

In Section 5 of the main text, we report Monte Carlo simulations that documented the behavior of DiD estimates for the effect of a placebo law on state-level log average employment and state-level log average monthly earnings from the QWI when the number of treated and untreated units was approximately equal ($\frac{N_1}{N} = \frac{25}{51}$). We now report the same results for the fraction of treated units varying over $N_1 \in \{\lfloor 0.4\, N \rfloor, \lfloor 0.6\, N \rfloor\}$ in Table 2, where $\lfloor \cdot \rfloor$ is the floor function. The results are qualitatively similar as the case with $N_1 = \lfloor 0.5\, N \rfloor$ in the main text.

D.2 Treatment effect heterogeneity

In Section 5 of the main text, we report Monte Carlo simulations that documented the behavior of DiD estimators for the effect of a placebo law on state-level average employment and state-level log average monthly earnings from the QWI. These simulations were conducted

	p_1		
	0.50	0.75	0.90
Normalized bias	−0.008	0.249	0.629
Variance conservativeness	1.035	1.316	2.910
Coverage	0.943	0.968	0.995
Oracle coverage	0.946	0.944	0.909

(a) Log employment with $N_1 = \lfloor 0.4\,N \rfloor$

	p_1		
	0.50	0.75	0.90
Normalized bias	−0.001	0.850	2.016
Variance conservativeness	0.981	1.311	2.713
Coverage	0.945	0.914	0.897
Oracle coverage	0.952	0.863	0.438

(b) Log earnings with $N_1 = \lfloor 0.4\,N \rfloor$

	p_1		
	0.50	0.75	0.90
Normalized bias	0.008	0.250	0.394
Variance conservativeness	0.989	1.257	1.648
Coverage	0.942	0.963	0.979
Oracle coverage	0.948	0.947	0.932

(c) Log employment with $N_1 = \lfloor 0.6\,N \rfloor$

	p_1		
	0.50	0.75	0.90
Normalized bias	−0.015	0.819	1.405
Variance conservativeness	1.005	1.265	1.886
Coverage	0.944	0.903	0.891
Oracle coverage	0.949	0.866	0.701

(d) Log earnings with $N_1 = \lfloor 0.6\,N \rfloor$

Table 2: Normalized bias, variance conservativeness, and coverage in Monte Carlo simulations with $N_1 \in \{\lfloor 0.4\,N \rfloor, \lfloor 0.6\,N \rfloor\}$.

Notes: Row 1 reports the normalized bias of the DiD estimator ($\mathbb{E}_R[\hat{\tau}_{DiD}] / \sqrt{\mathbb{V}\mathrm{ar}_R[\hat{\tau}_{DiD}]}$) for the EATT over the randomization distribution. Row 2 reports the estimated ratio $\frac{\mathbb{E}_R[\hat{s}^2]}{\mathbb{V}\mathrm{ar}_R[\hat{\tau}_{DiD}]}$ across simulations, which measures the conservativeness of the heteroskedasticity-robust variance estimator. Row 3 reports the estimated coverage rate of a 95% confidence interval for the EATT based on the limiting normal approximation of the randomization distribution of the DiD estimator and the heteroskedasticity-robust variance estimator $\hat{s}^2$. Row 4 reports the coverage rate of an "oracle" 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96 \sqrt{\mathbb{V}_R[\hat{\tau}_{DiD}]}$. The columns report results as the idiosyncratic treatment probability p^1 varies over $\{0.5, 0.75, 0.9\}$. The results are computed over 5,000 simulations with $N = 51$.

without treatment effect heterogeneity, setting $Y_{it}(1) = Y_{it}(0)$ both to equal the observed state-level outcomes Y_{it}.

We now report results from Monte Carlo simulations that incorporate treatment effect heterogeneity. As in the main text, we use aggregate data on the 50 U.S. states and Washington D.C. from the QWI (indexed by $i = 1, \ldots, N$) for the years 2012 and 2016 (indexed by $t = 1, 2$). For each state and year, we set the untreated potential outcome $Y_{it}(0)$ equal to the state's observed outcome in the QWI. We impose "no-anticipation" by setting $Y_{i1}(1) = Y_{i1}(0)$. We draw the treated potential outcome at $t = 2$ as $Y_{i2}(1) = Y_{i1}(0) + \lambda\sqrt{\mathbb{V}\text{ar}_1\left[Y_{i2}(0) - Y_{i1}(0)\right]}Z_i$, where Z_i is drawn from a standard normal distribution and $\lambda \in \{0.5, 1\}$. We draw the Z_i once and hold them fixed throughout the simulations. To ease interpretation, we recenter the draws of the unit-specific treatment effects $\lambda\sqrt{\mathbb{V}\text{ar}_1\left[Y_{i2}(0) - Y_{i1}(0)\right]}Z_i$ so that the EATT $\tau_{EATT,2}$ equals zero.

We simulate D from the rejective assignment mechanism using the state-level results in the 2016 presidential election as in the main text, and we fix the number of treated states at $N_1 = \lfloor 0.5\, N \rceil$. We again report results for two choices of the outcome Y_{it}: the log employment level for state i in period t, and the log of state-level average quarterly earnings for state i in year t.

Simulation results: Table 3 summarizes the normalized bias, variance conservativeness, and coverage in the Monte Carlo simulations. The first row reproduces the results in Table 1 without treatment effect heterogeneity (i.e., $\lambda = 0$). For a particular choice of the idiosyncratic treatment probabilities p^1, the bias of the DiD estimator for the EATT is fixed as the standard deviation of unit-specific treatment effects varies in these simulations. But, as the standard deviation of unit-specific treatment effects increases, the standard errors become noticeably more conservative. For example, for the log earnings outcome and $p^1 = 0.75$, the variance estimator is approximately 1.4 times too large when $\lambda = 0$, approximately 1.5 times too large when $\lambda = 0.5$, and approximately 2 times too large when $\lambda = 1$. As a result of this conservativeness, coverage rates increase for both outcomes as λ increases: e.g., for log-earnings with $p^1 = 0.75$, coverage is 91.7% with $\lambda = 0$, 93.5% with $\lambda = 0.5$, and 97.4% with $\lambda = 1$.

In Figure 2, we plot how the randomization distribution of the DiD estimator varies as we vary both the idiosyncratic treatment probabilities and the standard deviation of unit-specific treatment effects.

D.3 Varying Population Sizes

In Section 5, we reported results where the finite population was the 50 U.S. states and Washington D.C. We now report simulations where the size of the finite population varies. Specifically, we consider simulations designs with $N \in \{10, 26, 51\}$, where the smaller populations are obtained by choosing a subset of the 51 units in ascending order of their associated FIPS codes.

In Figure 3, we fix the standard deviation of unit-specific treatment effects to be $\lambda = 0$, and plot how the randomization distribution of the DiD estimator varies as we vary both the idiosyncratic treatment probabilities p^1 and the total number of states N. For $N = 10$, the distributions appear to be symmetric, but have oscillations that are not characteristic

	0.50	0.75	0.90
Normalized bias	0.013	0.250	0.525
Variance conservativeness	0.976	1.315	2.303
Coverage	0.939	0.967	0.991
Oracle coverage	0.949	0.943	0.917

(a) Log employment with $\lambda = 0$

	0.50	0.75	0.90
Normalized bias	0.004	0.882	1.871
Variance conservativeness	0.987	1.383	2.541
Coverage	0.944	0.917	0.888
Oracle coverage	0.952	0.854	0.516

(b) Log earnings with $\lambda = 0$

	p_1		
	0.50	0.75	0.90
Normalized bias	0.008	0.263	0.486
Variance conservativeness	1.071	1.495	2.761
Coverage	0.953	0.977	0.996
Oracle coverage	0.953	0.943	0.924

(c) Log employment with $\lambda = 0.5$

	p_1		
	0.50	0.75	0.90
Normalized bias	0.015	0.882	1.856
Variance conservativeness	1.068	1.517	2.925
Coverage	0.956	0.935	0.930
Oracle coverage	0.956	0.861	0.531

(d) Log earnings with $\lambda = 0.5$

	p_1		
	0.50	0.75	0.90
Normalized bias	0.000	0.225	0.453
Variance conservativeness	1.238	1.594	2.794
Coverage	0.967	0.980	0.999
Oracle coverage	0.952	0.944	0.924

(e) Log employment with $\lambda = 1$

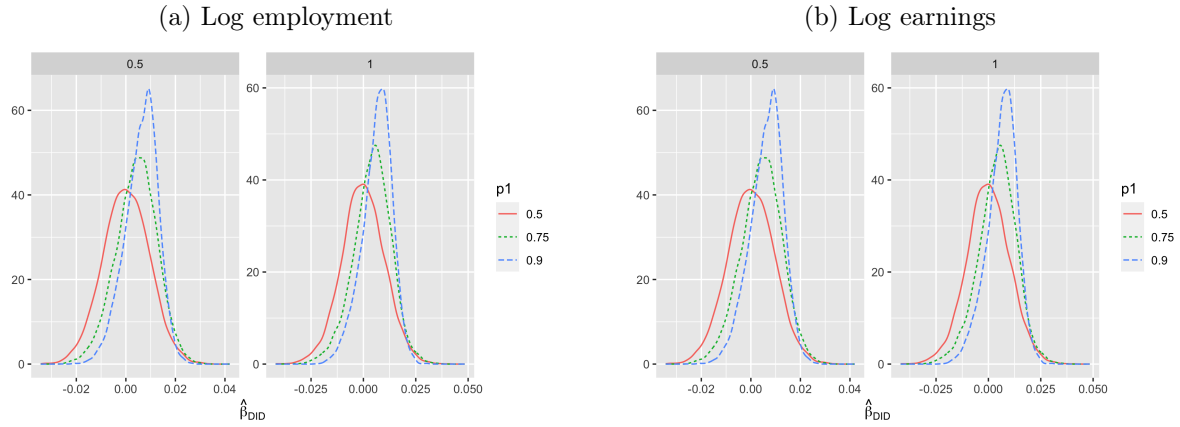
	p_1		
	0.50	0.75	0.90
Normalized bias	−0.033	0.857	1.910
Variance conservativeness	1.269	1.959	4.052
Coverage	0.965	0.974	0.981
Oracle coverage	0.951	0.861	0.513

(f) Log earnings with $\lambda = 1$

Table 3: Normalized bias, variance conservativeness, and coverage in Monte Carlo simulations.

Notes: Within a particular table, Row 1 reports the normalized bias of the DiD estimator ($\mathbb{E}_R[\hat{\tau}_{DiD}]/\sqrt{\mathbb{V}\text{ar}_R[\hat{\tau}_{DiD}]}$) for the EATT over the randomization distribution; Row 2 reports the estimated ratio $\frac{\mathbb{E}_R[\hat{s}^2]}{\mathbb{V}\text{ar}_R[\hat{\tau}_{DiD}]}$ across simulations, which measures the conservativeness of the heteroskedasticity-robust variance estimator; Row 3 reports the coverage rate of a nominal 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96\,\hat{s}$; and Row 4 reports coverage of an oracle confidence interval that uses the true variance rather than an estimated one. The columns report results as the idiosyncratic treatment probability p^1 varies over $\{0.5, 0.75, 0.9\}$. The results are computed over 5,000 simulations with $N_1 = \lfloor 0.5\,N \rfloor$ and $N = 51$. Panels (a)-(f) vary the outcome and the degree of treatment heterogeneity (λ).

Figure 2: Behavior of DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution with treatment effect heterogeneity.

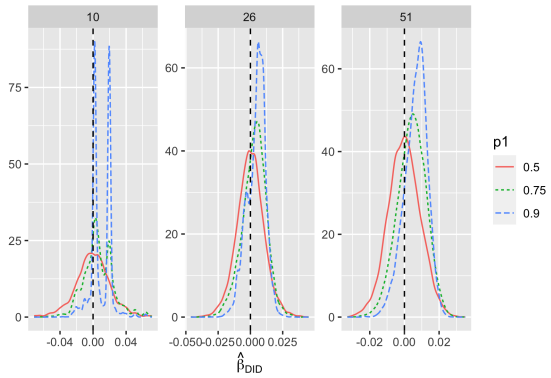


Notes: This figure plots the behavior of the DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution. The idiosyncratic treatment probabilities p^1 varies over $\{0.5, 0.75, 0.9\}$ (colors), and the standard deviation of unit-specific treatment effects λ varies over $\{0.5, 1\}$ (columns). The results are computed over $5,000$ simulations with $N_1 = \lfloor 0.5\, N \rfloor$ and $N = 51$.

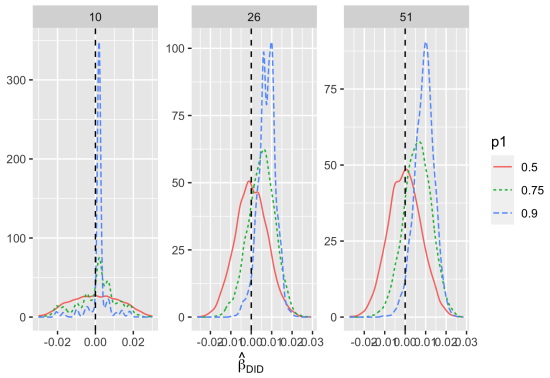
of a normal distribution (particularly for $p^1 = 0.9$). But, as N is increased to 26 (or 51), the distributions appear to be approximately normally distributed, illustrating the finite-population central limit theorem in Proposition 4.3. Table 4 summarizes how the coverage rate of a nominal 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96\, \hat{s}$ varies. Interestingly, for $N_c = 10$, despite the non-normal distribution we find that the coverage rate never drops below 91.9% for the log employment outcome and 92.3% for the log earnings outcome, although of course this finding may not generalize beyond the specific data-generating process studied here.

Figure 3: Behavior of DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution varying the size of the finite population.

(a) Log employment



(b) Log earnings



Notes: This figure plots the behavior of the DiD estimator $\hat{\tau}_{DiD}$ over the randomization distribution. The idiosyncratic treatment probabilities p^1 varies over $\{0.5, 0.75, 0.9\}$ (colors), and the total number of units N varies over $\{10, 26, 51\}$ (columns). The results are computed over $5,000$ simulations with $N_1 = \lfloor 0.5\, N \rfloor$ and $\lambda = 0$.

	p_1		
	0.5	0.75	0.90
N = 10	0.919	0.932	0.982
N = 26	0.935	0.966	0.995
N = 51	0.937	0.965	0.990

(a) Log employment with $\lambda = 0$

	p_1		
	0.5	0.75	0.90
N = 10	0.923	0.976	0.999
N = 26	0.938	0.929	0.946
N = 51	0.945	0.911	0.889

(b) Log earnings with $\lambda = 0$

Table 4: Coverage in Monte Carlo simulations varying the size of the finite population.

Notes: This table reports the coverage rate of a nominal 95% confidence interval of the form $\hat{\tau}_{DiD} \pm 1.96\, \hat{s}$ as the size of the finite population N varies over $\{10, 26, 51\}$ (rows) and the idiosyncratic treatment probability p^1 varies over $\{0.5, 0.75, 0.9\}$ (columns) The results are computed over 5,000 simulations with with $N_1 = \lfloor 0.5\, N \rfloor$ and $\lambda = 0$.