Efficient Estimation for Staggered Rollout Designs[*]

Jonathan Roth[†]   Pedro H.C. Sant'Anna[‡]

June 29, 2021

Abstract

This paper studies efficient estimation of causal effects when treatment is (quasi-) randomly rolled out to units at different points in time. We solve for the most efficient estimator in a class of estimators that nests two-way fixed effects models and other popular generalized difference-in-differences methods. A feasible plug-in version of the efficient estimator is asymptotically unbiased with efficiency (weakly) dominating that of existing approaches. We provide both t-based and permutation-test based methods for inference. We illustrate the performance of the plug-in efficient estimator in simulations and in an application to Wood et al. (2020a)'s study of the staggered rollout of a procedural justice training program for police officers. We find that confidence intervals based on the plug-in efficient estimator have good coverage and can be as much as five times shorter than confidence intervals based on existing state-of-the-art methods. As an empirical contribution of independent interest, our application provides the most precise estimates to date on the effectiveness of procedural justice training programs for police officers.

1 Introduction

Researchers are often interested in the causal effect of a treatment that is first implemented for different units at different times. In some cases, such as our application to the rollout of a police training program, the timing of the treatment can be explicitly randomized. In other settings, the researcher does not explicitly control the timing of treatment, but may argue that the timing of the treatment was determined by idiosyncratic, quasi-random factors.[1]

Staggered rollouts are commonly analyzed using methods that extend the simple two-period difference-in-differences (DiD) estimator to the staggered setting. This includes two-way fixed effects (TWFE) regression models, as well as recent alternatives that yield a more intuitive causal parameter when there are heterogeneous treatment effects (Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020). These methods all exploit a generalized parallel trends assumption for estimation. Although (quasi-) random treatment timing implies the parallel trends assumption, however, it has additional implications beyond parallel trends.

In this paper we exploit these additional implications to obtain estimators that are more efficient than DiD-based approaches under (quasi-) random treatment timing. We derive the most efficient estimator in a large class of estimators that nests many existing DiD-based approaches, and show how to conduct both t-based and permutation-based inference. Our results are complementary to the active literature on staggered DiD, since we provide more efficient estimates when treatment is (quasi-) randomly assigned, but our results may not be applicable in observational settings where the researcher is confident in parallel trends but not in (quasi-) random treatment timing.

We begin by introducing a design-based framework that formalizes the notion that treatment timing is (quasi-) randomly assigned. There are T periods, and unit i is first treated in period $G_i \in \mathcal{G} \subseteq \{1, ..., T, \infty\}$, with $G_i = \infty$ denoting that i is never treated (or treated after period T). We make two key assumptions in this model. First, we assume that the treatment timing G_i is (quasi-) randomly assigned. Second, we rule out anticipatory effects of treatment — for example, a unit's outcome in period two does not depend on whether it was first treated in period three or in period four.

Within this framework, we show that pre-treatment outcomes play a similar role to fixed covariates in a randomized experiment, and generalized DiD estimators can be viewed as applying a crude form of covariate adjustment. To see this, it is instructive to first consider

---

[1] For example, Lafortune et al. (2018) "exploits the apparent randomness of reform timing"; Rossin-Slater (2017) "leverages quasi-experimental variation in the timing of IHVPE program initiation across states and years"; and Denning (2017) writes that "the assumption is that timing of a vote authorizing annexation is exogenous."

the special case where we observe data for two periods ($T = 2$), some units are first treated in period 2 ($G_i = 2$), and the remaining units are treated in a later period or never treated ($G_i = \infty$). This special case is analogous to conducting a randomized experiment in period 2, with the outcome in period 1 serving as a pre-treatment covariate. The commonly used difference-in-differences estimator is $\hat{\theta}^{DiD} = (\bar{Y}_{22} - \bar{Y}_{2\infty}) - (\bar{Y}_{12} - \bar{Y}_{1\infty})$, where $\bar{Y}_{tg}$ is the mean outcome for treatment group g at period t. It is clear that $\hat{\theta}^{DiD}$ is a special case of the class of estimators of the form

$$
\hat{\theta}_\beta = \underbrace{(\bar{Y}_{22} - \bar{Y}_{2\infty})}_{\text{Post-treatment diff}} - \beta \underbrace{(\bar{Y}_{12} - \bar{Y}_{1\infty})}_{\text{Pre-treatment diff}} \tag{1}
$$

which adjust the post-treatment difference in means by β times the pre-treatment difference in means. Under the assumption of (quasi-) random treatment timing, the estimator $\hat{\theta}_\beta$ is unbiased for the average treatment effect (ATE) for any β, since the post-treatment difference in means is unbiased for the ATE and the pre-treatment difference in means is mean-zero. The value of β that minimizes the variance of the estimator depends on the covariances of the potential outcomes between periods, however. Intuitively, we want to put more weight on lagged outcomes when they are more informative about post-treatment outcomes. DiD, which imposes the fixed weight $\beta = 1$, will thus generally be inefficient, and one can obtain an (asymptotically) more efficient estimator by estimating the optimal weights from the data.

Our main theoretical results extend this logic to the case of staggered treatment timing, and provide formal methods for efficient estimation and inference. We begin by introducing a flexible class of causal parameters that can highlight treatment effect heterogeneity across both calendar time and time since treatment. Following Athey and Imbens (2021), we define $\tau_{t,gg'}$ to be the average effect on the outcome in period t of changing the initial treatment date from g' to g. For example, in the simple two-period case described above, $\tau_{2,2\infty}$ corresponds with the average treatment effect (ATE) on the second-period outcome of being treated in period two relative to never being treated. We then consider the class of estimands that are linear combinations of these building blocks, $\theta = \sum_{t,g,g'} a_{t,g,g'} \tau_{t,gg'}$. Our framework thus allows for arbitrary treatment effect dynamics, and accommodates a variety of ways of summarizing these dynamic effects, including several aggregation schemes proposed in the recent literature.

We then consider the large class of estimators that start with a sample analog to the target parameter and adjust by a linear combination of differences in pre-treatment outcomes. More precisely, we consider estimators of the form $\hat{\theta}_\beta = \sum_{t,g} a_{t,g,g'} \hat{\tau}_{t,gg'} - \hat{X}'\beta$, where the first term is a sample analog to θ, and the second term adjusts linearly using a vector $\hat{X}$ that compares outcomes for cohorts treated at different dates at points in time before either was treated.

For example, in the simple two-period case described above, $\hat{X} = \bar{Y}_{12} - \bar{Y}_{1\infty}$ is the difference-in-means in period 1. We show that several estimation procedures for the staggered setting are part of this class for an appropriately defined estimand and $\hat{X}$, including the classical TWFE estimator as well as recent procedures proposed by Callaway and Sant'Anna (2020), de Chaisemartin and D'Haultfœuille (2020), and Sun and Abraham (2020). All estimators of this form are unbiased for θ under the assumptions of random treatment timing and no anticipation.

We then derive the most efficient estimator in this class. The optimal coefficient β^* depends on covariances between the potential outcomes over time, and thus the estimators previously proposed in the literature will only be efficient for special covariance structures. Although the covariances of the potential outcomes are generally not known ex ante, one can estimate a "plug-in" version of the efficient estimator that replaces the "oracle" coefficient β^* with a sample analog $\hat{\beta}^*$. We show that the plug-in efficient estimator is asymptotically unbiased and as efficient as the oracle estimator under large population asymptotics similar to those in Lin (2013) and Li and Ding (2017) for covariate adjustment in cross-sectional experiments.

Our results suggest two complementary approaches to inference. First, we show that the plug-in efficient estimator is asymptotically normally distributed in large populations, which allows for asymptotically valid confidence intervals of the familiar form $\hat{\theta}_{\hat{\beta}*} \pm 1.96\hat{se}$.[2] Second, an appealing feature of our (quasi-) random treatment timing framework is that it permits us to construct Fisher randomization tests (FRTs), also known as permutation tests. Following Wu and Ding (2020) and Zhao and Ding (2020) for cross-sectional randomized experiments, we consider FRTs based on a studentized version of our efficient estimator. These FRTs have the dual advantages that they are finite-sample exact under the sharp null of no treatment effects, and asymptotically valid for the weak null of no average effects. In a Monte Carlo study calibrated to our application, we find that both the t-based and FRT-based approaches yield reliable inference, and CIs based on the plug-in efficient estimator are substantially shorter than those for the procedures of Callaway and Sant'Anna (2020), Sun and Abraham (2020), and de Chaisemartin and D'Haultfœuille (2020).[3]

As an illustration of our method and standalone empirical contribution, we revisit the randomized rollout of a procedural justice training program for police officers in Chicago. The original study by Wood et al. (2020a) found large and statistically significant reductions in complaints and officer use of force, and these findings were influential in policy debates

---

[2] As is common in finite-population settings, the covariance estimate may be conservative if there are heterogeneous treatment effects.

[3] The R package `staggered` allows for easy implementation of the plug-in efficient estimator, available at https://github.com/jonathandroth/staggered.

about policing (Doleac, 2020). Unfortunately, an earlier version of our analysis revealed a statistical error in the analysis of Wood et al. (2020a) which led their estimates to be inflated. In Wood et al. (2020b), we collaborated with the original authors to correct this error, finding no significant effects on complaints against police officers and borderline significant effects on officer use of force, but with wide confidence intervals that included both near-zero and meaningfully large treatment effects estimates. We find that the use of the methodology proposed in this paper allows us to obtain substantially more precise estimates of the effect of the training program. Although we again find no statistically significant effects on complaints and borderline significant effects on force, the standard errors from using our methodology are between 1.3 and 5.6 times smaller than from the Callaway and Sant'Anna (2020) estimator used in Wood et al. (2020b). For complaints, for example, we are able to rule out reductions larger than 11% of the pre-treatment mean, compared with an upper bound of 26% in the previous analysis. Our confidence intervals for the effects on complaints are also three times smaller than those for a smaller pilot study in Seattle (Owens et al., 2018).

Related Literature. This paper contributes to an active literature on DiD and related methods in settings with staggered treatment timing. Several recent papers have demonstrated the failures of TWFE models to recover a sensible causal estimand under treatment effect heterogeneity and have proposed alternative estimators with better properties (Borusyak and Jaravel, 2017; Goodman-Bacon, 2018; de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2020; Sun and Abraham, 2020). Most of this literature has focused on obtaining consistent estimates under a generalized parallel trends assumption, whereas we focus on *efficient* estimation under the stronger assumption of (quasi-) random treatment timing.

Two related papers that have studied (quasi-) random treatment timing are Athey and Imbens (2021) and Shaikh and Toulis (2019). The former studies a model of random treatment timing similar to ours, but focuses on the interpretation of the TWFE estimand. The latter paper adopts a different framework of randomization in which treatment timing is random only conditional on observables, and no two units can be treated at the same time. Neither paper considers the efficient choice of estimator as we do.

Our technical results extend results in statistics on efficient covariate adjustment in cross-sectional experiments (Freedman, 2008a,b; Lin, 2013; Li and Ding, 2017) to the setting of staggered treatment timing, where pre-treatment outcomes play a similar role to fixed covariates in a cross-sectional experiment. In the special two-period case, our results are related to McKenzie (2012), who showed that DiD is inefficient under random treatment assignment in a model with homogeneous treatment effects; see Remark 3 for additional

details.

Our paper also relates to the literature on clinical trials using a stepped wedge design, which is a staggered rollout in which all units are ultimately treated (Brown and Lilford, 2006; Davey et al., 2015; Turner et al., 2017; Thompson et al., 2017). Until recently, this literature has focused on estimation using mixed effects regression models. Lindner and Mcconnell (2021) points out, however, that such models may be difficult to interpret under heterogeneity, and recommends using DiD-based approaches like Sun and Abraham (2020) instead. Our approach has the potential to offer large gains in precision relative to such DiD-based approaches. Our paper is also complementary to Ji et al. (2017), who propose using randomization-based inference procedures to test Fisher's sharp null hypothesis in stepped wedge designs. By contrast, we consider Neymanian inference on average treatment effects, and also show how an FRT with a studentized statistic is both finite-sample exact for sharp nulls and asymptotically valid for inference on average effects.

Finally, our work is related to Xiong et al. (2019) and Basse et al. (2020), who consider how to optimally design a staggered rollout experiment to maximize the efficiency of a fixed estimator. By contrast, we solve for the most efficient estimator given a fixed experimental design.

2 Model and Theoretical Results

2.1 Model

There is a finite population of N units. We observe data for T periods, $t = 1, .., T$. A unit's treatment status is denoted by $G_i \in \mathcal{G} \subseteq \{1, ..., T, \infty\}$, where G_i is the first period in which unit i is treated, and $G_i = \infty$ denotes that a unit is never treated (or treated after period T). We assume that treatment is an absorbing state.[4] We denote by $Y_{it}(g)$ the potential outcome for unit i in period t when treatment starts at time g, and define the vector $Y_i(g) = (Y_{i1}(g), ..., Y_{iT}(g))' \in \mathbb{R}^T$. We let $D_{ig} = 1[G_i = g]$. The observed vector of outcomes for unit i is then $Y_i = \sum_i D_{ig} Y_i(g)$.

Following Neyman (1923) for randomized experiments and Athey and Imbens (2021) for settings with staggered treatment timing, our model is design-based: We treat as fixed (or condition on) the potential outcomes and the number of units first treated at each period (N_g). The only source of uncertainty in our model comes from the vector of times at which units are first-treated, $G = (G_1, ..., G_N)'$, which is stochastic.

---

[4]If treatment turns on and off, the parameters we estimate can be viewed as the intent-to-treat effect of first being treated at a particular date; see Sun and Abraham (2020) and de Chaisemartin and D'Haultfoeuille (2021) for related discussion for DiD models.

Remark 1 (Design-based uncertainty). Design-based models are particularly attractive in settings where it is difficult to define the super-population, such as when all 50 states are observed (Manski and Pepper, 2018), or in our application where the near-universe of police officers in Chicago is observed. Even when there is a super-population, the design-based view allows for valid inference on the sample average treatment effect (SATE); see Abadie et al. (2020), Sekhon and Shem-Tov (2020) for additional discussion.

Our first main assumption is that the treatment timing is (quasi-) randomly assigned, meaning that any permutation of the treatment timing vector is equally likely.

Assumption 1 (Random treatment timing). *Let D be the random $N \times |\mathcal{G}|$ matrix with (i, g)th element D_{ig}. Then $\mathbb{P}(D = d) = (\prod_{g \in \mathcal{G}} N_g!)/N!$ if $\sum_i d_{ig} = N_g$ for all g, and zero otherwise.*

We discuss extensions to clustered and conditional random assignment in Section 2.8.

Remark 2 (Comparison to parallel trends). The random timing assumption in Assumption 1 is stronger than the usual parallel trends assumption, which only requires that treatment probabilities are orthogonal to trends in the potential outcomes (see Rambachan and Roth (2020) for a discussion of parallel trends in design-based models). However, Assumption 1 can be ensured by design in settings where treatment timing can be explicitly randomized, such as our application in Section 4. Moreover, we show in Roth and Sant'Anna (2021) that without random treatment timing, the parallel trends assumption will be sensitive to functional form unless one imposes strong assumptions about the distribution of the potential outcomes. Thus, parallel trends will typically be most plausible in contexts where Assumption 1 is plausible, and indeed researchers often justify the parallel trends assumption in practice by arguing that treatment is timing is "quasi-random" or "quasi-experimental."

We also assume that the treatment has no causal impact on the outcome in periods before it is implemented. This assumption is plausible in many contexts, but may be violated if individuals learn of treatment status beforehand and adjust their behavior in anticipation (Abbring and van den Berg, 2003; Lechner, 2010; Malani and Reif, 2015).[5]

Assumption 2 (No anticipation). *For all i, $Y_{it}(g) = Y_{it}(g')$ for all $g, g' > t$.*

Note that this assumption does not restrict the possible dynamic effects of treatment – that is, we allow for $Y_{it}(g) \neq Y_{it}(g')$ whenever $t \geqslant min(g, g')$, so that treatment effects can arbitrarily depend on calendar time and the time that has elapsed since treatment. Rather,

---

[5]If anticipatory behavior is only possible within m periods of treatment (e.g., because treatment is announced m periods in advance), the initial treatment can be re-defined as $G_i - m$.

we only require that, say, a unit's outcome in period one does not depend on whether it was ultimately treated in period two or period three.

Example 1 (Special case: two periods). Consider the special case of our model in which there are two periods ($T = 2$) and units are either treated in period two or never treated ($\mathcal{G} = \{2, \infty\}$). Under random treatment timing and no anticipation, this special case is isomorphic to a cross-sectional experiment where the outcome $Y_i = Y_{i2}$ is the second period outcome, the binary treatment $D_i = 1[G_i = 2]$ is whether a unit is treated in period two, and the covariate $X_i = Y_{i1} \equiv Y_{i1}(\infty)$ is the pre-treatment outcome (which by the No Anticipation assumption does not depend on treatment status). Covariate adjustment in cross-sectional randomized experiments has been studied previously by Freedman (2008a,b), Lin (2013), and Li and Ding (2017), and our results will nest many of the existing results in the literature as a special case. The two-period special case also allows us to study the canonical difference-in-differences estimator, while avoiding complications discussed in the recent literature related to extending this estimator to the staggered case. We will therefore come back to this example throughout the paper to provide intuition and connect our results to the previous literature.

Notation. All expectations ($\mathbb{E}[\cdot]$) and probability statements ($\mathbb{P}(\cdot)$) are taken over the distribution of G conditional on the potential outcomes and the number of units treated at each period, $(N_g)_{g \in \mathcal{G}}$, although we suppress this conditioning for ease of notation. For a non-stochastic attribute W_i (e.g. a function of the potential outcomes), we denote by $\mathbb{E}_f[W_i] = N^{-1} \sum_i W_i$ and $\mathbb{V}\mathrm{ar}_f[W_i] = (N-1)^{-1} \sum_i (W_i - \mathbb{E}_f[W_i])(W_i - \mathbb{E}_f[W_i])'$ the finite-population expectation and variance of W_i.

2.2 Target Parameters

In our staggered treatment setting, the effect of being treated may depend on both the calendar time (t) as well as the time one was first treated (g). We therefore consider a large class of target parameters that allow researchers to highlight various dimensions of heterogeneous treatment effects across both calendar time and time since treatment.

Following Athey and Imbens (2021), we define $\tau_{it,gg'} = Y_{it}(g) - Y_{it}(g')$ to be the causal effect of switching the treatment date from date g' to g on unit i's outcome in period t. We define $\tau_{t,gg'} = N^{-1} \sum_i \tau_{it,gg'}$ to be the average treatment effect (ATE) of switching treatment from g' to g on outcomes at period t. We will consider scalar estimands of the form

$$\theta = \sum_{t,g,g'} a_{t,gg'} \tau_{t,gg'}, \tag{2}$$

8

i.e. weighted sums of the average treatment effects of switching from treatment g' to g, with $a_{t,gg'} \in \mathbb{R}$ being arbitrary weights. Researchers will often be interested in weighted averages of the $\tau_{t,gg'}$, in which case the $a_{t,gg'}$ will sum to 1, although our results allow for arbitrary $a_{t,gg'}$.[6] The results extend easily to vector-valued θ's where each component is of the form in the previous display; we focus on the scalar case for ease of notation. The no anticipation assumption (Assumption 2) implies that $\tau_{t,gg'} = 0$ if $t < min(g, g')$, and so without loss of generality we make the normalization that $a_{t,gg'} = 0$ if $t < min(g, g')$.

Example 1 (continued). In our simple two-period example, a natural target parameter is the average treatment effect (ATE) in period two. This corresponds with setting $\theta = \tau_{2,2\infty} = N^{-1} \sum_i Y_{i2}(2) - Y_{i2}(\infty)$.

We now describe a variety of intuitive parameters that can be captured by this framework in the general staggered setting. Researchers are often interested in the effect of receiving treatment at a particular time relative to not receiving treatment at all. We will define $ATE(t, g) := \tau_{t,g\infty}$ to be the average treatment effect on the outcome in period t of being first-treated at period g relative to not being treated at all. The $ATE(t, g)$ is a close analog to the cohort average treatment effects on the treated (ATTs) considered in Callaway and Sant'Anna (2020) and Sun and Abraham (2020). The main difference is that those papers do not assume random treatment timing, and thus consider ATTs rather than ATEs (in a sampling-based framework). In some cases, the $ATE(t, g)$ will be directly of interest and can be estimated in our framework.

When the dimension of t and g is large, however, it may be desirable to aggregate the $ATE(t, g)$ both for ease of interpretability and to increase precision. Our framework incorporates a variety of possible summary measures that aggregate the $ATE(t, g)$ across different cohorts and time periods. We briefly discuss a few possible aggregations which may be relevant in empirical work, mirroring proposals for aggregating the $ATT(t, g)$ in Callaway and Sant'Anna (2020).

When researchers are interested in how the treatment effect evolves with respect to the time elapsed since treatment started, they may want to consider "event-study" parameters that aggregate the ATEs at a given lag l since treatment ($l = 0, 1, ...$),

$$\theta_l^{ES} = \frac{1}{\sum_{g:g+l\leqslant T} N_g} \sum_{g:g+l\leqslant T} N_g ATE(g + l, g).$$

Note that the instantaneous parameter θ_0^{ES} is analogous to the estimand considered in

---

[6]This allows the possibility, for instance, that θ represents the difference between long-run and short-run effects, so that some of the $a_{t,gg'}$ are negative.

de Chaisemartin and D'Haultfœuille (2020) in settings like ours where treatment is an absorbing state (although their framework also extends to the more general setting where treatment turns on and off).

In other situations, it may be of interest to understand how the treatment effect differs over calendar time (e.g. during a boom or bust economy), or by the time that treatment began. In such cases, the summary parameters

$$\theta_t = \frac{1}{\sum_{g:g \leqslant t} N_g} \sum_{g:g \leqslant t} N_g ATE(t,g) \text{ and } \theta_g = \frac{1}{T-g+1} \sum_{t:t \geqslant g} ATE(t,g),$$

which respectively aggregate the ATEs for a particular calendar time or treatment adoption cohort, may be relevant.

Finally, researchers may be interested in a single summary parameter for the effect of a treatment. In this case, it may be instructive to consider a simple average of the $ATE(t,g)$ (weighted by cohort size),

$$\theta^{simple} = \frac{1}{\sum_t \sum_{g:g \leqslant t} N_g} \sum_t \sum_{g:g \leqslant t} N_g ATE(t,g),$$

or to consider a weighted average of the time or cohort effects,

$$\theta^{calendar} = \frac{1}{T} \sum_t \theta_t \qquad \text{or} \qquad \theta^{cohort} = \frac{1}{\sum_{g:g \neq \infty} N_g} \sum_{g:g \neq \infty} N_g \theta_g.$$

Since the most appropriate parameter will depend on context, we consider a broad framework that allows for efficient estimation of all of these (and other) parameters.[7]

2.3 Class of Estimators Considered

We now introduce the class of estimators we will consider. Intuitively, these estimators start with a sample analog to the target parameter and linearly adjust for differences in outcomes for units treated at different times in periods before either was treated.

Let $\bar{Y}_g = N_g^{-1} \sum_i D_{ig} Y_i$ be the vector of sample means of the outcome for treatment group g, and let $\hat{\tau}_{t,gg'} = \bar{Y}_{g,t} - \bar{Y}_{g',t}$ be the sample analog of $\tau_{t,gg'}$. We define

$$\hat{\theta}_0 = \sum_{t,g,g'} a_{t,gg'} \hat{\tau}_{t,gg'},$$

---

[7]We note that if $\infty \notin \mathcal{G}$, then $ATE(t,g)$ is only identified for $t < \max \mathcal{G}$. In this case, all of the sums above should be taken only over the (t,g) pairs for which $ATE(t,g)$ is identified.

which replaces the population means in the definition of θ with their sample analogues.

We will consider estimators of the form

$$\hat{\theta}_\beta = \hat{\theta}_0 - \hat{X}'\beta, \tag{3}$$

where, intuitively, $\hat{X}$ is a vector of differences-in-means that are guaranteed to be mean-zero under the assumptions of random treatment timing and no anticipation. Formally, we consider M-dimensional vectors $\hat{X}$ where each element of $\hat{X}$ takes the form

$$\hat{X}_j = \sum_{(t,g,g'):g,g'>t} b^j_{t,gg'} \hat{\tau}_{t,gg'},$$

where the $b^j_{t,gg'} \in \mathbb{R}$ are arbitrary weights. There are many possible choices for the vector $\hat{X}$ that satisfy these assumptions. For example $\hat{X}$ could be a vector where each component equals $\hat{\tau}_{t,gg'}$ for a different combination of (t,g,g') with $t < g, g'$. Alternatively, $\hat{X}$ could be a scalar that takes a weighted average of such differences. The choice of $\hat{X}$ is analogous to the choice of which variables to control for in a cross-sectional randomized experiment. In principle, including more covariates (higher-dimensional $\hat{X}$) will improve asymptotic precision, yet including "too many" covariates may lead to over-fitting, leading to poor performance in practice.[8] For now, we suppose the researcher has chosen a fixed $\hat{X}$, and will consider the optimal choice of β for a given $\hat{X}$. We will return to the choice of $\hat{X}$ in the discussion of our Monte Carlo results in Section 3 below.

Several estimators proposed in the literature can be viewed as special cases of the class of estimators we consider for an appropriately-defined estimand and $\hat{X}$, often with $\beta = 1$.

Example 1 (continued). In our running two-period example, $\hat{X} = \hat{\tau}_{1,2\infty}$ corresponds with the pre-treatment difference in sample means between the units first treated at period two and the never-treated units. Thus,

$$\hat{\theta}_1 = \hat{\tau}_{2,2\infty} - \hat{\tau}_{1,2\infty} = (\bar{Y}_{22} - \bar{Y}_{2\infty}) - (\bar{Y}_{12} - \bar{Y}_{1\infty})$$

is the canonical difference-in-differences estimator, where $\bar{Y}_{tg}$ represents the sample mean of Y_{it} for units with $G_i = g$. Likewise, $\hat{\theta}_0$ is the simple difference-in-means (DiM) in period two, $(\bar{Y}_{22} - \bar{Y}_{2\infty})$. More generally, the estimator $\hat{\theta}_\beta$ takes the simple difference-in-means in period two and adjusts by β times the difference-in-means in period one. Thus, for $\beta \in (0,1)$, $\hat{\theta}_\beta$ is a

---

[8]Lei and Ding (2020) study covariate adjustment in randomized experiments with a diverging number of covariates. In principle the vector $\hat{X}$ could also include pre-treatment differences in means of non-linear transformations of the outcome as well; see Guo and Basse (2020) for related results on non-linear covariate adjustments in randomized experiments.

weighted average of the DiM and DiD estimators. In this special case, the set of estimators of the form $\hat{\theta}_\beta$ is equivalent to the set of linear covariate-adjusted estimators for cross-sectional experiments considered in Lin (2013); Li and Ding (2017).[9]

Example 2 (Callaway and Sant'Anna (2020)). For settings where there is a never-treated group ($\infty \in \mathcal{G}$), Callaway and Sant'Anna (2020) consider the estimator

$$\hat{\tau}_{tg}^{CS} = \hat{\tau}_{t,g\infty} - \hat{\tau}_{g-1,g\infty},$$

i.e. a difference-in-differences that compares outcomes between periods t and $g-1$ for the cohort first treated in period g relative to the never-treated cohort. Observe that $\hat{\tau}_{tg}^{CS}$ can be viewed as an estimator of $ATE(t,g)$ of the form given in (3), with $\hat{X} = \hat{\tau}_{g-1,g\infty}$ and $\beta = 1$. Likewise, Callaway and Sant'Anna (2020) consider an estimator that aggregates the $\hat{\tau}_{tg}^{CS}$, say $\hat{\tau}_w^{CS} = \sum_{t,g} w_{t,g} \hat{\tau}_{t,g\infty}$, which can be viewed as an estimator of the parameter $\theta_w = \sum_{t,g} w_{t,g} ATE(t,g)$ of the form (3) with $\hat{X} = \sum_{t,g} w_{t,g} \hat{\tau}_{g-1,g\infty}$ and $\beta = 1$.[10] Similarly, Callaway and Sant'Anna (2020) consider an estimator that replaces the never-treated group with an average over cohorts not yet treated in period t,

$$\hat{\tau}_{tg}^{CS2} = \frac{1}{\sum_{g':g'>t} N_{g'}} \sum_{g':g'>t} N_{g'} \hat{\tau}_{t,gg'} - \frac{1}{\sum_{g':g'>t} N_{g'}} \sum_{g':g'>t} N_{g'} \hat{\tau}_{g-1,gg'}, \text{ for } t \geqslant g.$$

It is again apparent that this estimator can be written as an estimator of $ATE(t,g)$ of the form in (3), with $\hat{X}$ now corresponding with a weighted average of $\hat{\tau}_{g-1,gg'}$ and β again equal to 1.

Example 3 (Sun and Abraham (2020)). Sun and Abraham (2020) consider an estimator that is equivalent to that in Callaway and Sant'Anna (2020) in the case where there is a never-treated cohort. When there is no never-treated group, Sun and Abraham (2020) propose using the last cohort to be treated as the comparison. Formally, they consider the estimator of $ATE(t,g)$ of the form

$$\hat{\tau}_{tg}^{SA} = \hat{\tau}_{t,gg_{max}} - \hat{\tau}_{g-1,gg_{max}},$$

where $g_{max} = \max \mathcal{G}$ is the last period in which units receive treatment. It is clear that $\hat{\tau}_{tg}^{SA}$

---

[9]Lin (2013) and Li and Ding (2017) consider estimators of the form $\tau(\beta_0, \beta_1) = (\bar{Y}_1 - \beta_1'(\bar{X}_1 - \bar{X})) - (\bar{Y}_0 - \beta_0'(\bar{X}_0 - \bar{X}))$, where $\bar{Y}_d$ is the sample mean of the outcome Y_i for units with treatment $D_i = d$, $\bar{X}_d$ is defined analogously, and $\bar{X}$ is the unconditional mean of X_i. Setting $Y_i = Y_{i,2}$, $X_i = Y_{i,1}$, and $D_i = 1[G_i = 2]$, it is straightforward to show that the estimator $\tau(\beta_0, \beta_1)$ is equivalent to $\hat{\theta}_\beta$ for $\beta = \frac{N_2}{N}\beta_0 + \frac{N_\infty}{N}\beta_1$.

[10]This could also be viewed as an estimator of the form (3) if $\hat{X}$ were a vector with each element corresponding with $\hat{\tau}_{t,g\infty}$ and the vector β was a vector with elements corresponding with $w_{t,g\infty}$.

takes the form (3), with $\hat{X} = \hat{\tau}_{g-1,gg_{max}}$ and $\beta = 1$. Weighted averages of the $\hat{\tau}_{tg}^{SA}$ can likewise be expressed in the form (3), analogous to the Callaway and Sant'Anna (2020) estimators.

Example 4 (de Chaisemartin and D'Haultfœuille (2020)). de Chaisemartin and D'Haultfœuille (2020) propose an estimator of the instantaneous effect of a treatment. Although their estimator extends to settings where treatment turns on and off, in a setting like ours where treatment is an absorbing state, their estimator can be written as a linear combination of the $\hat{\tau}_{tg}^{CS2}$. In particular, their estimator is a weighted average of the Callaway and Sant'Anna (2020) estimates for the first period in which a unit was treated,

$$\hat{\tau}^{dCH} = \frac{1}{\sum_{g:g \leqslant T} N_g} \sum_{g:g \leqslant T} N_g \hat{\tau}_{gg}^{CS2}.$$

It is thus immediate from the previous examples that their estimator can also be written in the form (3).

Example 5 (TWFE Models). Athey and Imbens (2021) consider the setting with $\mathcal{G} = \{1, ...T, \infty\}$. Let $A_{it} = 1[G_i \leqslant t]$ be an indicator for whether unit i is already treated by period t. Athey and Imbens (2021, Lemma 5) show that the coefficient on A_{it} from the two-way fixed effects specification

$$Y_{it} = \alpha_i + \lambda_t + A_{it}\theta^{TWFE} + \epsilon_{it} \tag{4}$$

can be decomposed as

$$\hat{\theta}^{TWFE} = \sum_t \sum_{\substack{(g,g'): \\ min(g,g') \leqslant t}} \gamma_{t,gg'}\hat{\tau}_{t,gg'} + \sum_t \sum_{\substack{(g,g'): \\ min(g,g') > t}} \gamma_{t,gg'}\hat{\tau}_{t,gg'} \tag{5}$$

for weights $\gamma_{t,gg'}$ that depend only on the N_g and thus are non-stochastic in our framework. Thus, $\hat{\theta}^{TWFE}$ can be viewed as an estimator of the form (3) for the parameter $\theta^{TWFE} = \sum_t \sum_{(g,g'):min(g,g') \leqslant t} \gamma_{t,gg'}\tau_{t,gg'}$, with $X = -\sum_t \sum_{(g,g'):min(g,g') > t} \gamma_{t,gg'}\hat{\tau}_{t,gg'}$ and $\beta = 1$. As noted in Athey and Imbens (2021) and other papers, however, the parameter θ^{TWFE} may not have an intuitive causal interpretation under treatment effect heterogeneity, since the weights $\gamma_{t,gg'}$ may be negative.

2.4 Efficient "Oracle" Estimation

We now consider the problem of finding the best estimator $\hat{\theta}_\beta$ of the form introduced in (3). We first show that $\hat{\theta}_\beta$ is unbiased for all β, and then solve for the β^* that minimizes the variance.

13

Notation. We begin by introducing some notation that will be useful for presenting our results. Recall that the sample treatment effect estimates $\hat{\tau}_{t,gg'}$ are themselves differences in sample means, $\hat{\tau}_{t,gg'} = \bar{Y}_{t,g} - \bar{Y}_{t,g'}$. It follows that we can write

$$\hat{\theta}_0 = \sum_g A_{\theta,g} \bar{Y}_g \text{ and } \hat{X} = \sum_g A_{0,g} \bar{Y}_g$$

for appropriately defined matrices $A_{\theta,g}$ and $A_{0,g}$ of dimension $1 \times T$ and $M \times T$, respectively. Additionally, let $S_g = (N-1)^{-1} \sum_i (Y_i(g) - \mathbb{E}_f [Y_i(g)])(Y_i(g) - \mathbb{E}_f [Y_i(g)])'$ be the finite population variance of $Y_i(g)$ and $S_{gg'} = (N-1)^{-1} \sum_i (Y_i(g) - \mathbb{E}_f [Y_i(g)])(Y_i(g') - \mathbb{E}_f [Y_i(g')])'$ be the finite-population covariance between $Y_i(g)$ and $Y_i(g')$.

Our first result is that all estimators of the form $\hat{\theta}_\beta$ are unbiased, regardless of β.

Lemma 2.1 ($\hat{\theta}_\beta$ unbiased). *Under Assumptions 1 and 2, $\mathbb{E}\left[\hat{\theta}_\beta\right] = \theta$ for any $\beta \in \mathbb{R}^M$.*

We next turn our attention to finding the value β^* that minimizes the variance.

Proposition 2.1. *Under Assumptions 1 and 2, the variance of $\hat{\theta}_\beta$ is uniquely minimized at*

$$\beta^* = \mathbb{V}ar\left[\hat{X}\right]^{-1} \text{Cov}\left[\hat{X}, \hat{\theta}_0\right], \tag{6}$$

provided that $\mathbb{V}ar\left[\hat{X}\right]$ is positive definite. Further, the variances and covariances in the expression for β^* are given by

$$\mathbb{V}ar\left[\begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}\right] = \begin{pmatrix} \sum_g N_g^{-1} A_{\theta,g} S_g A'_{\theta,g} - N^{-1} S_\theta, & \sum_g N_g^{-1} A_{\theta,g} S_g A'_{0,g} \\ \sum_g N_g^{-1} A_{0,g} S_g A'_{\theta,g}, & \sum_g N_g^{-1} A_{0,g} S_g A'_{0,g} \end{pmatrix}$$

$$=: \begin{pmatrix} V_{\hat{\theta}_0} & V_{\hat{\theta}_0,\hat{X}} \\ V_{\hat{X},\hat{\theta}_0} & V_{\hat{X}} \end{pmatrix},$$

where $S_\theta = \mathbb{V}ar_f\left[\sum_g A_{\theta,g} Y_i(g)\right]$. The efficient estimator has variance given by $\mathbb{V}ar\left[\hat{\theta}_{\beta*}\right] = V_{\hat{\theta}_0} - (\beta^*)' V_{\hat{X}}^{-1}(\beta^*)$.

Equation (6) shows that the variance-minimizing β^* is the best linear predictor of $\hat{\theta}_0$ given $\hat{X}$. This formalizes the intuition that it is efficient to place more weight on pre-treatment differences in outcomes the more strongly they correlate with the post-treatment differences in outcomes.

Example 1 (continued). In our ongoing two-period example, the efficient estimator $\hat{\theta}_{\beta*}$ derived in Proposition 2.1 is equivalent to the efficient estimator for cross-sectional randomized experiments in Lin (2013) and Li and Ding (2017). The optimal coefficient β^* is equal to

$\frac{N_\infty}{N} \beta_2 + \frac{N_2}{N} \beta_\infty$, where β_g is the coefficient on Y_{i1} from a regression of $Y_{i2}(g)$ on Y_{i1} and a constant. Intuitively, this estimator puts more weight on the pre-treatment outcomes (i.e., β^* is larger) the more predictive is the first period outcome Y_{i1} of the second period potential outcomes. In the special case where the coefficients on lagged outcomes are equal to 1, the canonical difference-in-differences (DiD) estimator is optimal, whereas the simple difference-in-means (DiM) is optimal when the coefficients on lagged outcome are zero. For values of $\beta^* \in (0, 1)$, the efficient estimator can be viewed as a weighted average of the DiD and DiM estimators.

2.5 Properties of the plug-in estimator

Proposition 2.1 solves for the β^* that minimizes the variance of $\hat{\theta}_\beta$. However, the efficient estimator $\hat{\theta}_{\beta*}$ is not of practical use since the "oracle" coefficient β^* depends on the covariances of the potential outcomes, S_g, which are typically not known in practice. Mirroring Lin (2013) for cross-sectional randomized experiments, we now show that β^* can be approximated by a plug-in estimate $\hat{\beta}^*$, and the resulting estimator $\hat{\theta}_{\hat{\beta}*}$ has similar properties to the "oracle" estimator $\hat{\theta}_\beta$ in large populations.

2.5.1 Definition of the plug-in estimator

To formally define the plug-in estimator, let

$$\hat{S}_g = \frac{1}{N_g - 1} \sum_i D_{ig}(Y_i(g) - \bar{Y}_g)(Y_i(g) - \bar{Y}_g)'$$

be the sample analog to S_g, and let $\hat{V}_{\hat{X},\hat{\theta}_0}$ and $\hat{V}_{\hat{X}}$ be the analogs to $V_{\hat{X},\hat{\theta}_0}$ and $V_{\hat{X}}$ that replace S_g with $\hat{S}_g$ in the definitions. We then define the plug-in coefficient

$$\hat{\beta}^* = \hat{V}_{\hat{X}}^{-1} \hat{V}_{\hat{X},\hat{\theta}_0},$$

and will consider the properties of the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$.

Example 1 (continued). In our ongoing two-period example, which we have shown is analogous to a cross-sectional randomized experiment, the plug-in estimator $\hat{\theta}_{\hat{\beta}*}$ is equivalent to the efficient plug-in estimator for cross-sectional experiments considered in Lin (2013). As in Lin (2013), $\hat{\theta}_{\hat{\beta}*}$ can be represented as the coefficient on D_i in the interacted ordinary least squares (OLS) regression,

$$Y_{i2} = \beta_0 + \beta_1 D_i + \beta_2 \dot{Y}_{i1} + \beta_3 D_i \times \dot{Y}_{i1} + \epsilon_i, \tag{7}$$

where $\dot{Y}_{i1}$ is the demeaned value of Y_{i1}.[11] Intuitively, this fully-interacted specification fits one linear model to estimate the mean of $Y_{i2}(2)$ and another to estimate $Y_{i2}(\infty)$, and then computes the difference, and thus is analogous to an augmented inverse propensity weighted (AIPW) estimator (Glynn and Quinn, 2010).

Remark 3 (Connection to McKenzie (2012)). McKenzie (2012) proposes using an estimator similar to the plug-in efficient estimator in the two-period setting considered in our ongoing example. Building on results in Frison and Pocock (1992), he proposes using the coefficient γ_1 from the OLS regression

$$Y_{i2} = \gamma_0 + \gamma_1 D_i + \gamma_2 \dot{Y}_{i1} + \epsilon_i, \tag{8}$$

which is sometimes referred to as the Analysis of Covariance (ANCOVA I). This differs from the regression representation of the efficient plug-in estimator in (7), sometimes referred to as ANCOVA II, in that it omits the interaction term $D_i \dot{Y}_{i1}$. Treating $\dot{Y}_{i1}$ as a fixed pre-treatment covariate, the coefficient $\hat{\gamma}_1$ from (8) is equivalent to the estimator studied in Freedman (2008a,b). The results in Lin (2013) therefore imply that McKenzie (2012)'s estimator will have the same asymptotic efficiency as $\hat{\theta}_{\hat{\beta}*}$ under constant treatment effects. Intuitively, this is because the coefficient on the interaction term in (7) converges in probability to 0. However, the results in Freedman (2008a,b) imply that under heterogeneous treatment effects McKenzie (2012)'s estimator may even be less efficient than the simple difference-in-means $\hat{\theta}_0$, which in turn is (weakly) less efficient than $\hat{\theta}_{\hat{\beta}*}$.[12]

2.5.2 Asymptotic properties of the plug-in estimator

We will now show that in large populations, the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$ is asymptotically unbiased for θ and has the same asymptotic variance as the oracle estimator $\hat{\theta}_{\beta*}$. To derive the properties of the plug-in efficient estimator in large finite populations, we consider a sequence of finite populations of increasing sizes, as in Lin (2013) and Li and Ding (2017), among other papers. More formally, we consider sequences of populations indexed by m where the number of observations first treated at g, $N_{g,m}$, diverges for all $g \in \mathcal{G}$. For ease of notation, we leave the index m implicit in our notation for the remainder of the paper. We assume the sequence of populations satisfies the following regularity conditions.

Assumption 3. *(i) For all $g \in \mathcal{G}$, $N_g/N \to p_g \in (0,1)$.*

---

[11] We are not aware of a representation of the plug-in efficient estimator as the coefficient from an OLS regression in the more general, staggered case.

[12] Relatedly, Yang and Tsiatis (2001), Funatogawa et al. (2011), and Wan (2020) show that $\hat{\beta}_1$ from (7) is asymptotically at least as efficient as $\hat{\gamma}_1$ from (8) in sampling-based models similar to our ongoing example.

(ii) *For all g, g', S_g and $S_{gg'}$ have limiting values denoted S_g^* and $S_{gg'}^*$, respectively, with S_g^* positive definite.*

(iii) $\max_{i,g} ||Y_i(g) - \mathbb{E}_f [Y_i(g)] ||^2 / N \to 0.$

Part (i) imposes that the fraction of units first treated at N_g converges to a constant bounded between 0 and 1. Part (ii) requires the variances and covariances of the potential outcomes converge to a constant. Part (iii) requires that no single observation dominates the finite-population variance of the potential outcomes, and is thus analogous to the familiar Lindeberg condition in sampling contexts.

With these assumptions in hand, we are able to formally characterize the asymptotic distribution of the plug-in efficient estimator. The following result shows that $\hat{\theta}_{\hat{\beta}*}$ is asymptotically unbiased, with the same asympototic variance as the "oracle" efficient estimator $\hat{\theta}_{\beta*}$. The proof exploits the general finite population central limit theorem in Li and Ding (2017).

Proposition 2.2. *Under Assumptions 1, 2, and 3,*

$$\sqrt{N}(\hat{\theta}_{\hat{\beta}*} - \theta) \to_d \mathcal{N}\left(0, \sigma_*^2\right), \qquad where \qquad \sigma_*^2 = \lim_{N \to \infty} N \mathbb{V}ar\left[\hat{\theta}_{\beta*}\right].$$

2.6 Inference

We now introduce two methods for inference on θ, the first using conventional t-based confidence intervals, and the second using Fisher randomization tests.

2.6.1 t-based Confidence Intervals

To construct confidence intervals using the asymptotic normal distribution derived in Proposition 2.2, one requires an estimate of the variance σ_*^2. We first show that a simple Neyman-style variance estimator is conservative under treatment effect heterogeneity, as is common in finite population settings. We then introduce a refinement to this estimator that adjusts for the part of the heterogeneity explained by $\hat{X}$.

Recall that $\sigma_*^2 = \lim_{N \to \infty} N \mathbb{V}ar\left[\hat{\theta}_{\beta*}\right]$. Examining the expression for $\mathbb{V}ar\left[\hat{\theta}_{\beta*}\right]$ given in Proposition 2.1, we see that all of the components of the variance can be replaced with sample analogs except for the $-S_\theta$ term. This term corresponds with the variance of treatment effects, and is not consistently estimable since it depends on covariances between potential outcomes under treatments g and g' that are never observed simultaneously. This motivates the use of the Neyman-style variance that ignores the $-S_\theta$ term and replaces the variances

S_g with their sample analogs $\hat{S}_g$,

$$\hat{\sigma}_*^2 = \left(\sum_g \frac{N}{N_g} A_{\theta,g} \hat{S}_g A'_{\theta,g} \right) - \left(\sum_g \frac{N}{N_g} A_{\theta,g} \hat{S}_g A'_{0,g} \right) \left(\sum_g \frac{N}{N_g} A_{0,g} \hat{S}_g A'_{0,g} \right)^{-1} \left(\sum_g \frac{N}{N_g} A_{\theta,g} \hat{S}_g A'_{0,g} \right).$$

Since $\hat{S}_g \to_p S_g^*$ (see Lemma A.2), it is immediate that the estimator $\hat{\sigma}_*^2$ converges to an upper bound on the asymptotic variance σ_*^2, although the upper bound is conservative if there are heterogeneous treatment effects such that $S_\theta^* = \lim_{N\to\infty} S_\theta > 0$.

Lemma 2.2. *Under Assumptions 1, 2, and 3, $\hat{\sigma}_*^2 \to_p \sigma_*^2 + S_\theta^* \geqslant \sigma_*^2$.*

When the estimand θ does not involve any treatment effects for the cohort treated in period one, the estimator $\hat{\sigma}_*^2$ can be improved by using outcomes from earlier periods. The refined estimator intuitively lower bounds the heterogeneity in treatment effects by the part of the heterogeneity that is explained by the outcomes in earlier periods. The construction of this refined estimator mirrors the refinements using fixed covariates in randomized experiments considered in Lin (2013) and Abadie et al. (2020), with lagged outcomes playing a similar role to the fixed covariates.[13] To avoid technical clutter, we defer the technical derivation of the refinement to Appendix A.1, and merely state the sense in which the refined estimator improves upon the Neyman-style estimator introduced above.

Lemma 2.3. *The refined estimator $\hat{\sigma}_{**}$, defined in Lemma A.4, satisfies $\hat{\sigma}_{**}^2 \to_p \sigma_*^2 + S_{\tilde{\theta}}^*$, where $0 \leqslant S_{\tilde{\theta}}^* \leqslant S_\theta^*$, so that $\hat{\sigma}_{**}$ is asymptotically (weakly) less conservative than $\hat{\sigma}_*$.*

It is then immediate that the confidence interval, $CI_{**} = \hat{\beta}^* \pm z_{1-\alpha/2} \hat{se}$ is a valid $1 - \alpha$ level confidence interval for θ, where $\hat{se} = \hat{\sigma}_{**}/\sqrt{n}$ is the standard error and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the normal distribution.

2.6.2 Fisher Randomization Tests

An alternative approach to inference uses Fisher randomization tests (FRTs), otherwise known as permutation tests. We will show that an FRT using a studentized version of the efficient estimator has the dual advantages that it 1) has exact size under the sharp null of no treatment effects for all units, and 2) is asymptotically valid for the weak null that $\theta = 0$.

To derive the FRT, recall that the observed data is (Y, G), where Y collects all of the Y_{it} and $G = (G_1, ..., G_N)'$. Let $\mathcal{T} = \mathcal{T}(Y, G)$ denote a statistic of the data, and let $\mathcal{T}_\pi = \mathcal{T}(Y, G_\pi)$

---

[13] Aronow et al. (2014) provide sharp bounds on the variance of the difference-in-means estimator in randomized experiments, although these bounds are difficult to extend to other estimators and settings like those considered here.

be the statistic using the transformed data in which G is replaced with a permutation G_π.[14]
A Fisher randomization test (FRT) computes the p-value

$$p_{FRT} = P_{\pi \sim U(\Pi)}(\mathcal{T}_\pi \geqslant \mathcal{T}(Y, G)),$$

where the probability is taken over the uniform distribution on the set of permutations Π.[15]
Under the sharp null hypothesis that $Y_i(g) = Y_i(g')$ for all i, g, g', the distribution of $\mathcal{T}_\pi$ is
the same as the distribution as $\mathcal{T}(Y, G)$, and thus by standard arguments the FRT is exact
in finite samples (see, e.g., Imbens and Rubin (2015)).

The sharp null hypothesis of no treatment effect will often be too restrictive in practice,
however, as we may be more interested in the hypothesis that the *average* effect is zero, i.e.
$H_0 : \theta = 0$. Unfortunately, in general FRTs may not have correct size for such weak null
hypotheses even asymptotically (Wu and Ding, 2020).

We now show, however, that when the FRT is based on the studentized statistic $\mathcal{T}(Y, G) = \hat{\theta}_{\hat{\beta}*}/\hat{se}$, it has asymptotically correct size under the weak null. In fact, we will show that
asymptotically the FRT is equivalent to testing that 0 falls within the t-based confidence
interval CI_{**} derived in the previous section. Thus, this FRT based on the studentized
statistic is in some sense the "best of both worlds" of Fisherian and Neymanian inference in
that it has exact size under the sharp null hypothesis while having asymptotically correct
size under the weak null.

The following regularity condition imposes that the means of the potential outcomes have
limits, and that their fourth moment is bounded.

Assumption 4. *Suppose that for all g, $\lim_{N\to\infty} \mathbb{E}_f[Y_i(g)] = \mu_g < \infty$, and there exists $L < \infty$
such that $N^{-1}\sum_i ||Y_i(g) - \mathbb{E}_f[Y_i(g)]||^4 < L$ for all N.*

With this assumption in hand, we can make precise the sense in which the FRT is
asymptotically valid under the weak null.

Proposition 2.3. *Suppose Assumptions 1-4 hold. Let $t_\pi = (\hat{\theta}^*/\hat{se})_\pi$ be the studentized
statistic under permutation π. Then $t_\pi \to_d \mathcal{N}(0, 1)$, P_G-almost surely. Hence, if p_{FRT} is
the p-value from the FRT associated with $|t_\pi|$, then under $H_0 : \theta = 0$,*

$$\lim_{N\to\infty} P(p_{FRT} \leqslant \alpha) \leqslant \alpha,$$

P_G-almost surely, with equality if and only if $S_\theta^* = 0$.

---

[14]Formally, a permutation π is a bijective map from $\{1, ..., N\}$ onto itself, and $G_\pi = (G_{\pi(1)}, ..., G_{\pi(N)})'$.

[15]It is often difficult to calculate the p-value over all permutations exactly, so the p-value is approximated
via simulation. We use 500 simulation draws in our implementation of the FRT below.

Proposition 2.3 implies that the FRT using the studentized version of the efficient estimator asymptotically controls size under the weak null of no average treatment effects. Indeed, the proposition implies that the FRT is asymptotically equivalent to the test that the t-based confidence interval CI_{**} includes 0. Proposition 2.3 extends the results in Wu and Ding (2020) and Zhao and Ding (2020), who consider permutation tests based on a studentized statistic in cross-sectional randomized experiments.[16] Given the desirable properties of the FRT under both the sharp and weak null hypotheses, we recommend that researchers report p-values from the FRT alongside the usual t-based confidence intervals.

2.7 Implications for existing estimators

We now discuss the implications of our results for estimators previously proposed in the literature. We have shown that in the simple two-period case considered in Example 1, the canonical difference-in-differences corresponds with $\hat{\theta}_1$. Likewise, in the staggered case, we showed in Examples 2-4 that the estimators of Callaway and Sant'Anna (2020), Sun and Abraham (2020), and de Chaisemartin and D'Haultfœuille (2020) correspond with the estimator $\hat{\theta}_1$ for an appropriately defined estimand and $\hat{X}$. Our results thus imply that, unless $\beta^* = 1$, the estimator $\hat{\theta}_{\beta*}$ is unbiased for the same estimand and has strictly lower variance under random treatment timing. Since the optimal β^* depends on the potential outcomes, we do not generically expect $\beta^* = 1$, and thus the previously-proposed estimators will generically be dominated in terms of efficiency. Although the optimal β^* will typically not be known, our results imply that the plug-in estimator $\hat{\theta}_{\hat{\beta}*}$ will have similar properties in large populations, and thus will be more efficient than the previously-proposed estimators in large populations under random treatment timing.

We note, however, that the estimators in the aforementioned papers are valid for the ATT in settings where only parallel trends holds but there is not random treatment timing, whereas the validity of the efficient estimator depends on random treatment timing.[17] Although parallel trends is often justified in practice by arguing that treatment is (quasi-) randomly assigned, in some observational settings the researcher may be more comfortable imposing parallel trends than quasi-random treatment timing. We thus view the the plug-in efficient estimator to be complementary to the estimators considered in previous work, since it is more efficient under stricter assumptions that will not hold in all cases of interest.

---

[16]Permutation tests based on a studentized statistic have been considered in other contexts as well, for example Janssen (1997); Chung and Romano (2013, 2016); DiCiccio and Romano (2017); Bugni et al. (2018); Bai et al. (2019); MacKinnon and Webb (2020).

[17]The estimator of de Chaisemartin and D'Haultfœuille (2020) can also be applied in settings where treatment turns on and off over time.

2.8 Extensions and practical considerations

We now discuss several extensions and practical considerations that may be useful for applying our methods.

Remark 4 (Conditional Random Treatment Timing)**.** For simplicity, we have considered the case of unconditional random treatment timing. In some experiments, the treatment timing may be randomized among units with some shared observable characteristics (e.g. counties within a state). In this case, the methodology described above can be applied within each randomization stratum, and the stratum-level estimates can be pooled to form aggregate estimates for the population. Likewise, in quasi-experimental contexts, the assumption of quasi-random treatment timing may be more plausible among sub-groups of the population (e.g. within units of the same gender and education status), or among groups of units that were treated at similar times (e.g. within a decade). The units can then be partitioned into strata based on observable characteristics, and the analysis we describe can be conducted within each stratum.[18]

Remark 5 (Clustered Treatment Assignment)**.** Likewise, in some settings there may be clustered assignment of treatment timing — e.g. treatment is assigned to families f, and all units i in family f are first treated at the same time. This violates Assumption 1, since not all vectors of treatment timing are equally likely. However, note that any average treatment contrast at the individual level, e.g. $\frac{1}{N}\sum_i Y_{it}(g) - Y_{it}(g')$, can be written as an average contrast of a transformed family-level outcome, e.g. $\frac{1}{F}\sum_f \tilde{Y}_{ft}(g) - \tilde{Y}_{ft}(g')$, where $\tilde{Y}_{ft}(g) = (F/N)\sum_{i \in f} Y_{it}(g)$. Thus, clustered assignment can easily be handled in our framework by analyzing the transformed data at the cluster level.

Remark 6 (Testing the randomization assumption)**.** In non-experimental settings, it may be desirable to test the assumption of quasi-random treatment timing. One intuitive approach which mirrors the common practice of testing for pre-existing trends is to estimate an event-study, treating the initial time of treatment as $G_i - k$ for some $k > 0$, and then test whether the treatment effects corresponding with the leads $1, ..., k$ are different from zero. Alternatively, a permutation test (as in Section 2.6) using any test statistic depending only on pre-treatment outcomes will have exact size in finite samples; for instance, one could use an F-statistic for the equality of the mean pre-treatment outcomes for all cohorts in periods prior to treatment adoption. We caution, however, that as with tests of pre-existing trends (cf. Roth, 2020), such tests may have limited power to detect violations of the randomization

---

[18]It also seems feasible to adapt some of our procedures to alternative randomization schemes such as the covariate-adaptive randomization described in Bugni et al. (2019), though a formal treatment of this extension is beyond the scope of this paper.

assumption, and thus it is best to additionally motivate the randomization assumption based on context-specific knowledge.

Remark 7 (Fixed pre-treatment covariates)**.** In some settings, researchers may also have access to fixed covariates W_i. Differences in the mean of W_i between adoption cohorts can then be added to the vector $\hat{X}$ to further increase precision.

3 Monte Carlo Results

We present two sets of Monte Carlo results. In Section 3.1, we conduct simulations in a stylized two-period setting matching our ongoing example to illustrate how the plug-in efficient estimator compares to the classical difference-in-differences and simple difference-in-means (DiM) estimators. Section 3.2 presents a more realistic set of simulations with staggered treatment timing that is calibrated to our application, comparing the plug-efficient estimator to recent DiD-based estimators proposed for the staggered treatment case.

3.1 Two-period Simulations.

Specification. We follow the model in Example 1 in which there are two periods ($t = 1, 2$) and units are treated in period two or never-treated ($\mathcal{G} = \{1, 2\}$). We first generate the potential outcomes as follows. For each unit i in the population, we draw the never-treated potential outcomes $Y_i(\infty) = (Y_{i1}(\infty), Y_{i2}(\infty))'$ from a $\mathcal{N}(0, \Sigma_\rho)$ distribution, where Σ_ρ has 1s on the diagonal and ρ on the off-diagonal. The parameter ρ is the correlation between the untreated potential outcomes in period $t = 1$ and period $t = 2$. We then set $Y_{i2}(2) = Y_{i2}(\infty) + \tau_i$, where $\tau_i = \gamma(Y_{i2}(\infty) - \mathbb{E}_f[Y_{i2}(\infty)])$. The parameter γ governs the degree of heterogeneity of treatment effects: if $\gamma = 0$, then there is no treatment effect heterogeneity, whereas if γ is positive then individuals with larger untreated outcomes in $t = 2$ have larger treatment effects. We center by $\mathbb{E}_f[Y_{i2}(\infty)]$ so that the treatment effects are 0 on average. We generate the potential outcomes once, and treat the population as fixed throughout our simulations. Our simulation draws then differ based on the draw of the treatment assignment vector. For simplicity, we set $N_2 = N_\infty = N/2$, and in each simulation draw, we randomly select which units are treated in $t = 1$ or not. We conduct 1000 simulations for all combinations of $N_2 \in \{25, 1000\}$, $\rho \in \{0, .5, .99\}$, and $\gamma \in \{0, 0.5\}$.

Results. Table 1 shows the bias, standard deviation, and coverage of 95% confidence intervals for the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$, difference-in-differences $\hat{\theta}^{DiD} = \hat{\theta}_1$, and simple differences-in-means $\hat{\theta}^{DiM} = \hat{\theta}_0$. It also shows the size (null rejection probability) of the FRT

using a studentized statistic introduced in Section 2.6. Confidence intervals are constructed as $\hat{\theta}_{\hat{\beta}*} \pm 1.96\hat{\sigma}_{**}$ for the plug-in efficient estimator, and analogously for the other estimators.[19] For all specifications and estimators, the estimated bias is small, and coverage is close to the nominal level. Table 1 facilitates comparison of the standard deviations of the different estimators by showing the ratio relative to the plug-in estimator. The standard deviation of the plug-in efficient estimator is weakly smaller than that of either DiD or DiM in nearly all cases, and is never more than 2% larger than that of either DiD or DiM. The standard deviation of the plug-in efficient estimator is similar to DiD when auto-correlation of $Y(0)$ is high ($\rho = 0.99$) and there is no heterogeneity of treatment effects ($\gamma = 0$), so that $\beta^* \approx 1$ and thus DiD is (nearly) optimal in the class we consider. Likewise, it is similar to DiM when there is no autocorrelation ($\rho = 0$) and there is no treatment effect heterogeneity ($\gamma = 0$), and thus $\beta^* \approx 0$ and so DiM is (nearly) optimal in the class we consider. The plug-in efficient estimator is substantially more precise than DiD and DiM in many other specifications: the standard deviation of DiD can be as much as 1.7 times larger than the plug-in efficient estimator, and the standard deviation of the DiM can be as much as 7 times larger. These simulations thus illustrate how the plug-in efficient estimator can improve on DiD or DiM in cases where they are suboptimal, while retaining nearly identical performance when the DiD or DiM model is optimal.

				Bias			SD			Coverage			FRT Size		
N_1	N_0	ρ	γ	PlugIn	DiD	DiM	PlugIn	DiD	DiM	PlugIn	DiD	DiM	PlugIn	DiD	DiM
1000	1000	0.99	0.0	0.00	0.00	−0.00	0.01	0.01	0.04	0.95	0.95	0.95	0.05	0.05	0.05
1000	1000	0.99	0.5	0.00	0.00	−0.00	0.01	0.01	0.06	0.95	0.95	0.95	0.04	0.06	0.05
1000	1000	0.50	0.0	0.00	0.00	0.00	0.04	0.04	0.05	0.94	0.95	0.94	0.06	0.05	0.05
1000	1000	0.50	0.5	0.00	0.00	0.00	0.05	0.05	0.06	0.95	0.95	0.95	0.06	0.05	0.05
1000	1000	0.00	0.0	−0.00	0.00	−0.00	0.04	0.07	0.04	0.95	0.94	0.95	0.05	0.06	0.05
1000	1000	0.00	0.5	−0.00	0.00	−0.00	0.06	0.07	0.06	0.95	0.95	0.95	0.04	0.05	0.05
25	25	0.99	0.0	0.00	0.00	−0.03	0.04	0.04	0.27	0.94	0.94	0.94	0.04	0.05	0.06
25	25	0.99	0.5	0.00	−0.01	−0.04	0.05	0.08	0.34	0.92	0.93	0.93	0.06	0.06	0.06
25	25	0.50	0.0	−0.01	0.02	−0.02	0.24	0.29	0.26	0.94	0.95	0.94	0.04	0.04	0.05
25	25	0.50	0.5	−0.01	0.01	−0.03	0.30	0.32	0.33	0.94	0.95	0.94	0.04	0.04	0.05
25	25	0.00	0.0	−0.03	−0.02	−0.03	0.28	0.38	0.27	0.93	0.95	0.93	0.06	0.04	0.06
25	25	0.00	0.5	−0.04	−0.02	−0.04	0.35	0.42	0.34	0.93	0.94	0.94	0.06	0.05	0.06

Table 1: Bias, Standard Deviation, and Coverage for $\hat{\theta}_{\hat{\beta}*}, \hat{\theta}^{DiD}, \hat{\theta}^{DiM}$ in 2-period simulations

---

[19]For $\hat{\theta}_\beta$, we use an analog to $\hat{\sigma}_{**}$, except the unrefined estimate $\hat{\sigma}_*$ is replaced with the sample analog to the expression for $\mathbb{V}\mathrm{ar}\left[\hat{\theta}_\beta\right]$ implied by Proposition 2.1 rather than $\mathbb{V}\mathrm{ar}\left[\hat{\theta}_{\beta*}\right]$.

					SD Relative to Plug-In		
N_1	N_0	ρ	γ	β^*	PlugIn	DiD	DiM
1000	1000	0.99	0.0	0.99	1.00	1.00	7.09
1000	1000	0.99	0.5	1.24	1.00	1.71	7.07
1000	1000	0.50	0.0	0.52	1.00	1.13	1.15
1000	1000	0.50	0.5	0.65	1.00	1.04	1.15
1000	1000	0.00	0.0	−0.03	1.00	1.45	1.00
1000	1000	0.00	0.5	−0.03	1.00	1.31	1.00
25	25	0.99	0.0	0.97	1.00	0.99	6.58
25	25	0.99	0.5	1.22	1.00	1.47	6.31
25	25	0.50	0.0	0.41	1.00	1.21	1.10
25	25	0.50	0.5	0.51	1.00	1.08	1.10
25	25	0.00	0.0	0.10	1.00	1.35	0.98
25	25	0.00	0.5	0.13	1.00	1.22	0.98

Table 1: Ratio of standard deviations for $\hat{\theta}^{DiD}$ and $\hat{\theta}^{DiM}$ relative to $\hat{\theta}_{\hat{\beta}*}$ in 2-period simulations

3.2 Simulations Based on Wood et al. (2020b)

To evaluate the performance of our proposed methods in a more realistic staggered setting, we conduct simulations calibrated to our application to Wood et al. (2020b) in Section 4. The outcome of interest Y_{it} is the number of complaints against police officer i in month t for police officers in Chicago. Police officers were randomly assigned to first receive a procedural justice training in period G_i. See Section 4 for more background on the application.

Simulation specification. We calibrate our baseline specification as follows. The number of observations and time periods in the data exactly matches the data from Wood et al. (2020b) used in our application. We set the untreated potential outcomes $Y_{it}(\infty)$ to match the observed outcomes in the data Y_i (which would exactly match the true potential outcomes if there were no treatment effect on any units). In our baseline simulation specification, there is no causal effect of treatment, so that $Y_{it}(g) = Y_{it}(\infty)$ for all g. (We describe an alternative simulation design with heterogeneous treatment effects in Appendix Section B.) In each simulation draw s, we randomly draw a vector of treatment dates $G_s = (G_1^s, ..., G_N^s)$ such that the number of units first treated in period g matches that observed in the data (i.e. $\sum 1[G_i^s = g] = N_g$ for all g). In total, there are 72 months of data on 7785 officers. There are 48 distinct values of g, with the cohort size N_g ranging from 6 to 642. In an alternative specification, we collapse the data to the yearly level, so that there are 6 time periods and 5

larger cohorts.

For each simulated data-set, we calculate the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$ for four estimands: the simple weighted average ATE (θ^{simple}); the calendar- and cohort-weighted average treatment effects ($\theta^{calendar}$ and θ^{cohort}), and the instantaneous event-study parameter (θ_0^{ES}).[20] (See Section 2.2 for the formal definition of these estimands). In our baseline specification, we use as $\hat{X}$ the scalar weighted combination of pre-treatment differences used by the Callaway and Sant'Anna (2020, CS) estimator for the appropriate estimand (see Example 2). In the appendix, we also present results for an alternative specification in which $\hat{X}$ is a vector containing $\hat{\tau}_{t,gg'}$ for all pairs $g, g' > t$. For comparison, we also compute the CS and Sun and Abraham (2020, SA) estimators for the same estimand. Recall that for θ_0^{ES}, the CS estimator coincides with the estimator proposed in de Chaisemartin and D'Haultfœuille (2020) in our setting, since treatment is an absorbing state. Confidence intervals are calculated as $\hat{\theta}_{\hat{\beta}*} \pm 1.96\hat{\sigma}_{**}$ for the plug-in efficient estimator and analogously for the CS and SA estimators.[21]

Baseline simulation results. The results for our baseline specification are shown in Tables 2 and 3. As seen in Table 2, the plug-in efficient estimator is approximately unbiased, and 95% confidence intervals based on our standard errors have coverage rates close to the nominal level for all of the estimands, with size distortions no larger than 3% for all of our specifications. The size for the FRT is also close to the nominal level, which is intuitive since our baseline specification imposes the sharp null hypothesis, and thus the FRT should be exact up to simulation error. The CS and SA estimators are also both approximately unbiased and have good coverage for all of the estimands as well.

Table 3 shows that there are large efficiency gains from using the plug-in efficient estimator relative to the CS or SA estimators. The table compares the standard deviation of the plug-in efficient estimator to that of the CS and SA estimators. Remarkably, using the plug-in efficient estimator reduces the standard deviation relative to the CS estimator by a factor of nearly two for the calendar-weighted average, and by a factor between 1.36 and 1.67 for

---

[20]We do not report results for the estimand of TWFE specifications, in light of the recent literature showing that these estimands do not have an intuitive causal interpretation in settings with staggered treatment timing (e.g. Borusyak and Jaravel (2017); Athey and Imbens (2021); Goodman-Bacon (2018); de Chaisemartin and D'Haultfœuille (2020); Sun and Abraham (2020)). The results for the DiD estimator in the previous section illustrate the performance of TWFE in a simple setting where it has an intuitive estimand.

[21]The variance estimator for the CS and SA estimators is adapted analogously to that for the DiD and DiM estimators, as discussed in footnote 19. We note that these design-based standard errors differ slightly from those proposed in the original CS and SA papers, which adopt a sampling-based framework; using design-based standard errors makes the CIs for these estimators the more directly comparable to those for the plug-in efficient estimator.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.00	0.93	0.06	0.27	0.29
PlugIn	cohort	0.00	0.92	0.06	0.24	0.24
PlugIn	ES0	0.01	0.94	0.05	0.26	0.27
PlugIn	simple	0.00	0.92	0.06	0.22	0.22
CS	calendar	0.00	0.94	0.05	0.55	0.55
CS	cohort	-0.01	0.95	0.05	0.41	0.41
CS/dCDH	ES0	0.01	0.94	0.07	0.36	0.36
CS	simple	-0.01	0.96	0.05	0.41	0.40
SA	calendar	0.06	0.93	0.04	1.30	1.30
SA	cohort	0.05	0.92	0.05	1.34	1.38
SA	ES0	0.03	0.94	0.06	0.83	0.89
SA	simple	0.06	0.92	0.04	1.46	1.49

Table 2: Results for Simulations Calibrated to Wood et al. (2020b)

Note: This table shows results for the plug-in efficient and CS and SA estimators in simulations calibrated to Wood et al. (2020b). The estimands considered are the calendar-, cohort-, and simple-weighted average treatment effects, as well as the instantaneous event-study effect (ES0). The CS estimator for ES0 corresponds with the estimator in de Chaisemartin and D'Haultfœuille (2020). Coverage refers to the fraction of the time a nominal 95% confidence interval includes the true parameter, and FRT size refers to the null rejection rate of a Fisher Randomization Test. Mean SE refers to the average estimated standard error, and SD refers to the actual standard deviation of the estimator. The bias, Mean SE, and SD are all multiplied by 100 for ease of readability.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.92	4.57
cohort	1.67	5.68
ES0	1.36	3.33
simple	1.82	6.76

Table 3: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator

Note: This table shows the ratio of the standard deviation of the CS and SA estimators relative to the plug-in efficient estimator, based on the simulation results in Table 2.

the other estimands. Since standard errors are proportional to the square root of the sample size, these results suggest that using the plug-in efficient estimator is roughly equivalent to multiplying the sample size by a factor of four for the calendar-weighted average. The gains of using the plug-in efficient estimator relative to the SA estimator are even larger, with

reductions in the standard deviation by a factor of three or more. The reason for this is that the SA estimator uses only the last-treated units (rather than not-yet-treated units) as a comparison, but in our setting less than 1% of units are treated in the final period, leading to an efficiency loss.

Extensions. Appendix B contains several extensions to the baseline simulation specification, such as incorporating heterogeneous effects, annualizing the monthly data, using other choices of $\hat{X}$, and considering the other two outcomes in our application.[22]

4 Application to Procedural Justice Training

4.1 Background

Reducing police misconduct and use of force is an important policy objective. Wood et al. (2020a) studied the Chicago Police Department's staggered rollout of a procedural justice training program, which taught police officers strategies for emphasizing respect, neutrality, and transparency in the exercise of authority. Officers were randomly assigned a date for training.[23] Wood et al. (2020a) found large and statistically significant impacts of the program on complaints and sustained complaints against police officers and on officer use of force. However, our re-analysis in Wood et al. (2020b) highlighted a statistical error in the original analysis of Wood et al. (2020a), which failed to normalize for the fact that groups of officers trained on different days were of varying sizes. In Wood et al. (2020b), we re-analyzed the data using the procedure proposed by Callaway and Sant'Anna (2020) to correct for the error. The re-analysis found no significant effect on complaints or sustained complaints, and borderline significant effects on use of force, although the confidence intervals for all three outcomes included both near-zero and meaningfully large effects. Owens et al. (2018) studied a small pilot study of a procedural justice training program in Seattle, with point estimates suggesting reductions in complaints but imprecisely estimated.

---

[22] As in the baseline specification, the plug-in efficient estimator has good coverage and offers efficiency gains relative to the other methods in nearly all specifications. The one exception is one specification (calendar aggregation for sustained complaints), which has a rare outcome and puts large weight on small cohorts, in which t-based CIs for the plug-in efficient estimator undercover (79%); see Appendix B for details.

[23] See the Supplement to Wood et al. (2020a) for discussion of some concerns regarding non-compliance, particularly towards the end of the sample. We explore robustness to dropping officers trained in the last year in Appendix Figure 2. The results are qualitatively similar, although with smaller estimated effects on use of force.

4.2   Data

We use the same data as in the re-analysis in Wood et al. (2020b), which extends the data used in the original analysis of Wood et al. (2020a) through December 2016. As in Wood et al. (2020b), we restrict attention to the balanced panel of 7,785 who remained in the police force throughout the study period. The data contain three outcome measures (complaints, sustained complaints, and use of force) at a monthly level for 72 months (6 years), with the first cohort trained in month 13 and the final cohort trained in the last month of the sample.

4.3   Estimation

We apply our proposed plug-in efficient estimator to estimate the effects of the procedural justice training program on the three outcomes of interest. We estimate the simple-, cohort-, and calendar-weighted average effects described in Section 2.2 and used in our Monte Carlo study. We also estimate the average event-study effects for the first 24 months after treatment, which includes the instantaneous event-study effect studied in our Monte Carlo as a special case (for event-time 0). For comparison, we also estimate the Callaway and Sant'Anna (2020) estimator as in Wood et al. (2020b).[24]

4.4   Results

Figure 1 shows the results of our analysis for the three aggregate summary parameters. Table 4 compares the magnitudes of these estimates and their 95% confidence intervals (CIs) to the mean of the outcome in the 12 months before treatment began. It also reports p-values from the FRT.

For all outcomes, the CIs for the plug-in efficient overlap with those of the Callaway and Sant'Anna (2020, CS) estimator but are substantially narrower. Indeed, the final column of Table 4 show that the standard errors range from 1.3 to 5.6 times smaller depending on the specification. As in Wood et al. (2020b), we find no significant impact on complaints using any of the aggregations. Our bounds on the magnitude of the treatment effect are substantially tighter than before, however. For instance, using the simple aggregation we can now rule out reductions in complaints of more than 11%, compared with a bound of 26% using the CS estimator. Using the simple aggregation scheme our standard errors for complaints are 1.9 times smaller than when using CS and over three times smaller than those in Owens et al. (2018) (normalizing both estimates as a fraction of the pre-treatment mean). For use of force, the point estimates are somewhat smaller in magnitude than when using

---

[24]For direct comparability, we calculate design-based standard errors for the CS estimator using the analog to $\hat{\sigma}_{**}$, and thus the reported SEs differ slightly from the sampling-based SEs reported in Wood et al. (2020b).

Figure 1: Effect of Procedural Justice Training Using the Plug-In Efficient and Callaway and Sant'Anna (2020) Estimators
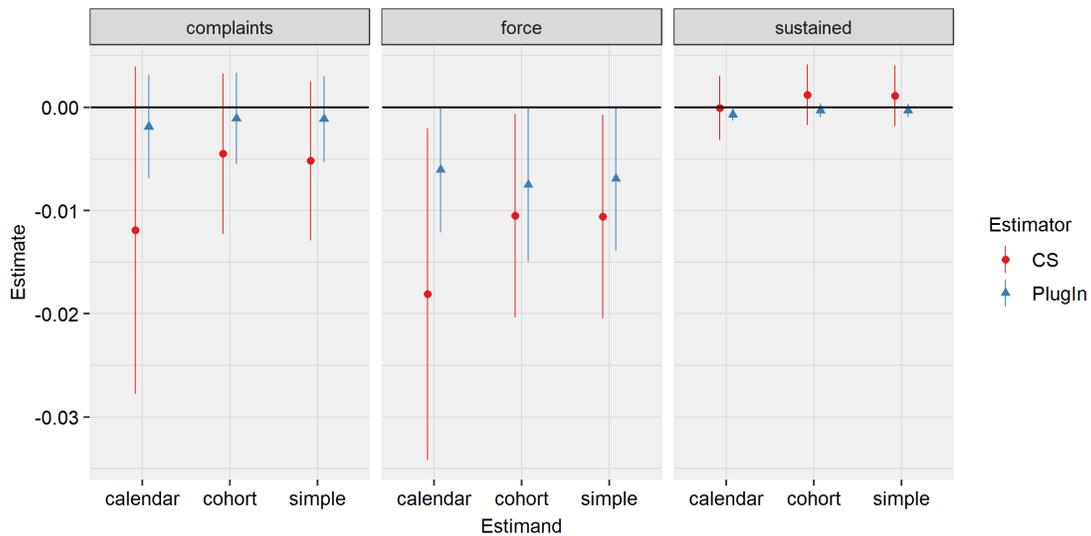
Note: this figure shows point estimates and 95% CIs for the effects of procedural justice training on complaints, force, and sustained complaints using the CS and plug-in efficient estimators. Results are shown for the calendar-, cohort-, and simple-weighted averages.

			Plug-In				CS				
Outcome	Estimand	Pre-treat Mean	Estimate	LB	UB	p-val (FRT)	Estimate	LB	UB	p-val (FRT)	CI Ratio
complaints	simple	0.049	−2%	−11%	6%	0.63	−10%	−26%	5%	0.21	1.9
complaints	calendar	0.049	−4%	−14%	6%	0.51	−24%	−56%	8%	0.20	3.2
complaints	cohort	0.049	−2%	−11%	7%	0.67	−9%	−25%	7%	0.28	1.8
sustained	simple	0.004	−7%	−23%	8%	0.38	27%	−44%	97%	0.51	4.5
sustained	calendar	0.004	−17%	−30%	−3%	0.10	−1%	−75%	73%	0.99	5.6
sustained	cohort	0.004	−7%	−22%	9%	0.43	29%	−41%	99%	0.47	4.4
force	simple	0.048	−15%	−29%	0%	0.10	−22%	−43%	−2%	0.03	1.4
force	calendar	0.048	−13%	−26%	0%	0.06	−38%	−72%	−4%	0.03	2.6
force	cohort	0.048	−16%	−31%	−0%	0.09	−22%	−43%	−1%	0.03	1.3

Table 4: Estimates and 95% CIs as a Percentage of Pre-treatment Means

Note: This table shows the pre-treatment means for the three outcomes. It also displays the estimates and 95% CIs in Figure 1 as percentages of these means, as well as the p-value from a Fisher Randomization Test (FRT). The final column shows the ratio of the length of the CI for the CS estimator relative to that for the plug-in efficient estimator.

the CS estimator and the upper bounds of the confidence intervals are all nearly exactly 0; p-values using the FRT are all between 0.05 and 0.1. Although precision is substantially higher than when using the CS estimator, the CIs for force still include effects between near-zero and 29% of the pre-treatment mean. For sustained complaints, all of the point estimates are near zero and the CIs are substantially narrower than when using the CS estimator, although the plug-in efficient estimate using the calendar aggregation is marginally significant (FRT $p = 0.1$). If we were to Bonferroni-adjust all of the CIs in Figure 1 for testing nine hypotheses (three outcomes times three aggregations), none of the confidence intervals would rule out zero.

Figure 2 shows event-study estimates for the first two years after treatment using the plug-in efficient estimator, and Appendix Figure 1 shows the analogous results using the CS estimator. In dark blue, we present point estimates and pointwise confidence intervals, and in light blue we present sup-t simultaneous confidence bands (Olea and Plagborg-Møller, 2019).[25] It has been argued that simultaneous confidence bands are more appropriate for event-study analyses since they control size over the full dynamic path of treatment effects (Freyaldenhoven et al., 2019; Callaway and Sant'Anna, 2020). Figure 2 shows that the simultaneous confidence bands include zero for nearly all periods for all three outcomes. Inspecting the results for force more closely, we see that the point estimates are positive (although typically not significant) for most of the first year after treatment, but become consistently negative around the start of the second year from treatment. This suggests that the negative point estimates in the aggregate summary statistics are driven mainly by months after the first year. Although it is possible that the treatment effects grow over time, this runs counter to the common finding of fadeout in educational programs in general (Bailey et al., 2020) and anti-bias training in particular (Forscher and Devine, 2017). Appendix C contains additional results for our application, including the dynamic event-study results for CS and other robustness checks.

Our analysis provides the most precise estimates to date on the effectiveness of procedural justice training for police officers. Unfortunately, our results are consistent with the hypothesis of near-zero effects on all outcomes, and provide substantially tighter bounds on the maximum effectiveness of such programs relative to prior work. Given the low cost of such training programs, however, these programs may nevertheless clear a cost-benefit analysis at the upper bounds of our CIs for their effectiveness (e.g. an 11% reduction in complaints). Thus, while our results suggest that these programs are likely not as effective as was initially hoped, there may yet be a role for procedural justice training in reducing police misconduct.

---

[25]We use the `suptCriticalValue` R package developed by Ryan Kessler.

Figure 2: Event-Time Average Effects Using the Plug-In Efficient Estimator
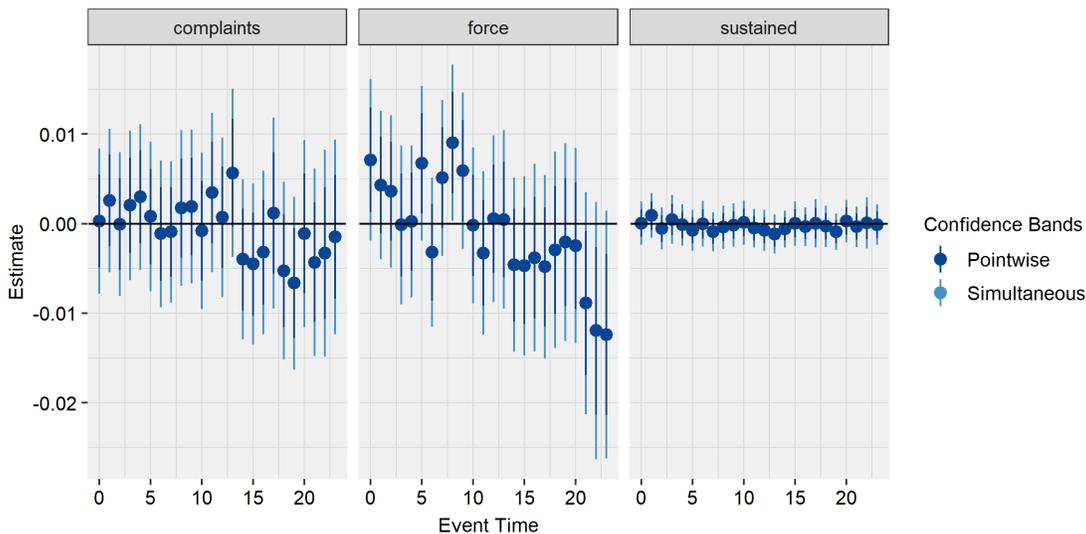
5   Conclusion

This paper considers efficient estimation in settings with staggered adoption and (quasi-) random treatment timing. The assumption of (quasi-) random treatment timing is stronger than parallel trends, but is often the justification given for the parallel trends assumption, and it can be ensured by design in experimental contexts like our application. We derive the most efficient estimator in a large class of estimators that nests many existing approaches. The "oracle" efficient estimator is not known in practice, but we show that a plug-in sample analog has similar properties in large populations, and we derive both t-based and permutation-based approaches to inference. We find in simulations that the proposed plug-in efficient estimator is approximately unbiased, yields reliable inference, and substantially increases precision relative to existing methods. We apply our proposed methodology to obtain the most precise estimates to date of the causal effects of procedural justice training programs for police officers.

References

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

Abbring, Jaap H. and Gerard J. van den Berg, "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 2003, *71* (5), 1491–1517. Publisher: [Wiley,

Econometric Society].

Aronow, Peter M., Donald P. Green, and Donald K. K. Lee, "Sharp bounds on the variance in randomized experiments," *The Annals of Statistics*, June 2014, *42* (3), 850–871.

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2021, *Forthcoming.*

Bai, Yuehao, Azeem Shaikh, and Joseph P. Romano, "Inference in Experiments with Matched Pairs," SSRN Scholarly Paper ID 3379977, Social Science Research Network, Rochester, NY April 2019.

Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman, and David S. Yeager, "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions:," *Psychological Science in the Public Interest*, October 2020.

Basse, Guillaume, Yi Ding, and Panos Toulis, "Minimax designs for causal effects in temporal experiments with treatment habituation," *arXiv:1908.03531 [stat]*, June 2020. arXiv: 1908.03531.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY 2017.

Brown, Celia A. and Richard J. Lilford, "The stepped wedge trial design: A systematic review," *BMC Medical Research Methodology*, 2006, *6*, 1–9.

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh, "Inference Under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, October 2018, *113* (524), 1784–1796. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1375934.

_ , _ , and _ , "Inference under covariate-adaptive randomization with multiple treatments," *Quantitative Economics*, 2019, *10* (4), 1747–1785.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, December 2020.

Chung, Eun Yi and Joseph P. Romano, "Multivariate and multiple permutation tests," *Journal of Econometrics*, July 2016, *193* (1), 76–91. Publisher: Elsevier BV.

Chung, EunYi and Joseph P. Romano, "Exact and asymptotically robust permutation tests," *The Annals of Statistics*, April 2013, *41* (2), 484–507. Publisher: Institute of Mathematical Statistics.

Davey, Calum, James Hargreaves, Jennifer A. Thompson, Andrew J. Copas, Emma Beard, James J. Lewis, and Katherine L. Fielding, "Analysis and reporting of stepped wedge randomised controlled trials: Synthesis and critical appraisal of published studies, 2010 to 2014," *Trials*, 2015, *16* (1).

de Chaisemartin, Clément and Xavier D'Haultfoeuille, "Difference-in-Differences Estimators of Intertemporal Treatment Effects," *arXiv:2007.04267 [econ]*, May 2021. arXiv: 2007.04267.

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

Denning, Jeffrey T., "College on the Cheap: Consequences of Community College Tuition Reductions," *American Economic Journal: Economic Policy*, May 2017, *9* (2), 155–188.

DiCiccio, Cyrus J. and Joseph P. Romano, "Robust Permutation Tests For Correlation And Regression Coefficients," *Journal of the American Statistical Association*, July 2017, *112* (519), 1211–1220. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1202117.

Doleac, Jennifer, "How to Fix Policing," *Neskanen Center*, 2020.

Forscher, Patrick S and Patricia G Devine, "Knowledge-based interventions are more likely to reduce legal disparities than are implicit bias interventions," 2017.

Freedman, David A., "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, 2008, *2* (1), 176–196.

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

Freyaldenhoven, Simon, Christian Hansen, and Jesse Shapiro, "Pre-event Trends in the Panel Event-study Design," *American Economic Review*, 2019, *109* (9), 3307–3338.

Frison, L. and S. J. Pocock, "Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design," *Statistics in Medicine*, September 1992, *11* (13), 1685–1704.

Funatogawa, Takashi, Ikuko Funatogawa, and Yu Shyr, "Analysis of covariance with pretreatment measurements in randomized trials under the cases that covariances and post-treatment variances differ between groups," *Biometrical Journal*, May 2011, *53* (3), 512–524.

Glynn, Adam N. and Kevin M. Quinn, "An Introduction to the Augmented Inverse Propensity Weighted Estimator," *Political Analysis*, 2010, *18* (1), 36–56. Publisher: [Oxford University Press, Society for Political Methodology].

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," Working Paper 25018, National Bureau of Economic Research September 2018.

Guo, Kevin and Guillaume Basse, "The Generalized Oaxaca-Blinder Estimator," *arXiv:2004.11615 [math, stat]*, April 2020. arXiv: 2004.11615.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 edition ed., New York: Cambridge University Press, April 2015.

Janssen, Arnold, "Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem," *Statistics & Probability Letters*, November 1997, *36* (1), 9–21.

Ji, Xinyao, Gunther Fink, Paul Jacob Robyn, and Dylan S. Small, "Randomization inference for stepped-wedge cluster-randomized trials: An application to community-based health insurance," *Annals of Applied Statistics*, 2017, *11* (1), 1–20.

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.

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference MethodsEstimation of Spatial Panels," *Foundations and Trends® in Econometrics*, 2010, *4* (3), 165–224.

Lei, Lihua and Peng Ding, "Regression adjustment in completely randomized experiments with a diverging number of covariates," *Biometrika*, December 2020, (Forthcoming).

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.

Lin, Winston, "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, March 2013, *7* (1), 295–318.

Lindner, Stephan and K John Mcconnell, "Heterogeneous treatment effects and bias in the analysis of the stepped wedge design," *Health Services and Outcomes Research Methodology*, 2021, (0123456789).

MacKinnon, James G and Matthew D Webb, "Randomization inference for difference-in-differences with few treated clusters," *Journal of Econometrics*, oct 2020, *218* (2), 435–450.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, April 2015, *124*, 1–17.

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.

McKenzie, David, "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 2012, *99* (2), 210–221.

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

Olea, José Luis Montiel and Mikkel Plagborg-Møller, "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 2019, *34* (1), 1–17. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jae.2656.

Owens, Emily, David Weisburd, Karen L. Amendola, and Geoffrey P. Alpert, "Can You Build a Better Cop?," *Criminology & Public Policy*, 2018, *17* (1), 41–87.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602 [econ.EM]*, 2020. arXiv: 2003.09915.

Rossin-Slater, Maya, "Signing Up New Fathers: Do Paternity Establishment Initiatives Increase Marriage, Parental Investment, and Child Well-Being?," *American Economic Journal: Applied Economics*, April 2017, *9* (2), 93–130.

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *Working paper*, 2020.

‗ **and Pedro H. C. Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form?," *arXiv:2010.04814 [econ, stat]*, January 2021. arXiv: 2010.04814.

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.

Shaikh, Azeem and Panos Toulis, "Randomization Tests in Observational Studies with Staggered Adoption of Treatment," *arXiv:1912.10610 [stat]*, December 2019. arXiv: 1912.10610.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, December 2020.

Thompson, Jennifer A., Katherine L. Fielding, Calum Davey, Alexander M. Aiken, James R. Hargreaves, and Richard J. Hayes, "Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis," *Statistics in Medicine*, 2017, *36* (23), 3670–3682.

Turner, Elizabeth L., Fan Li, John A. Gallis, Melanie Prague, and David M. Murray, "Review of recent methodological developments in group-randomized trials: Part 1 - Design," *American Journal of Public Health*, 2017, *107* (6), 907–915.

Wan, Fei, "Analyzing pre-post designs using the analysis of covariance models with and without the interaction term in a heterogeneous study population," *Statistical Methods in Medical Research*, January 2020, *29* (1), 189–204.

Wood, George, Tom R. Tyler, and Andrew V. Papachristos, "Procedural justice training reduces police use of force and complaints against officers," *Proceedings of the National Academy of Sciences*, May 2020, *117* (18), 9815–9821.

‗ , ‗ , ‗ , **Jonathan Roth, and Pedro H.C. Sant'Anna**, "Revised Findings for "Procedural justice training reduces police use of force and complaints against officers"," *Working Paper*, 2020.

Wu, Jason and Peng Ding, "Randomization Tests for Weak Null Hypotheses in Randomized Experiments," *Journal of the American Statistical Association*, May 2020, pp. 1–16. arXiv: 1809.07419.

Xiong, Ruoxuan, Susan Athey, Mohsen Bayati, and Guido Imbens, "Optimal Experimental Design for Staggered Rollouts," *arXiv:1911.03764 [econ, stat]*, November 2019. arXiv: 1911.03764.

Yang, Li and Anastasios A Tsiatis, "Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial," *The American Statistician*, November 2001, *55* (4), 314–321. Publisher: Taylor & Francis.

Zhao, Anqi and Peng Ding, "Covariate-adjusted Fisher randomization tests for the average treatment effect," *arXiv:2010.14555 [math, stat]*, November 2020. arXiv: 2010.14555.

Supplement to "Efficient Estimation for Staggered Rollout Designs"

A   Proofs

Proof of Lemma 2.1

Proof. By Assumption 1, $\mathbb{E}[D_{ig}] = (N_g/N)$. Hence,

$$\mathbb{E}\left[\hat{\theta}_0\right] = \mathbb{E}\left[\sum_g A_{\theta,g}\frac{1}{N_g}\sum_i D_{ig}Y_i\right] = \sum_g A_{\theta,g}\frac{1}{N_g}\sum_i \mathbb{E}[D_{ig}]Y_i(g) = \sum_g A_{\theta,g}\frac{1}{N_g}\sum_i \frac{N_g}{N}Y_i(g) = \theta.$$

Likewise,

$$\mathbb{E}\left[\hat{X}\right] = \mathbb{E}\left[\sum_g A_{0,g}\frac{1}{N_g}\sum_i D_{ig}Y_i\right] = \sum_g A_{0,g}\frac{1}{N}\sum_i Y_i(g) = \frac{1}{N}\sum_i\sum_g A_{0,g}Y_i(g) = 0,$$

since $\sum_g A_{0,g}Y_i(g) = 0$ by Assumption 2. The result follows immediately from the previous two displays. $\qquad\square$

Proof of Proposition 2.1

Proof. First, observe that

$$\min_\beta \mathbb{V}\mathrm{ar}\left[\hat{\theta}_\beta\right] = \min_\beta \mathbb{V}\mathrm{ar}\left[\hat{\theta}_0 - \hat{X}'\beta\right] = \min_\beta \mathbb{E}\left[\left((\hat{\theta}_0 - \theta) - (\hat{X} - \mathbb{E}\left[\hat{X}\right])'\beta)\right)^2\right].$$

From the usual least-squares formula, the unique solution is

$$\underbrace{\mathbb{E}\left[(\hat{X} - \mathbb{E}\left[\hat{X}\right])(\hat{X} - \mathbb{E}\left[\hat{X}\right])'\right]^{-1}}_{\mathbb{V}\mathrm{ar}[\hat{X}]^{-1}}\underbrace{\mathbb{E}\left[(\hat{X} - \mathbb{E}\left[\hat{X}\right])(\hat{\theta}_0 - \theta)\right]}_{\mathrm{Cov}[\hat{X},\hat{\theta}_0]},$$

which gives the first result.

To derive the form of the variance, let $A_{\tau,g} = \begin{pmatrix} A_{\theta,g} \\ A_{0,g} \end{pmatrix}$. Define

$$\hat{\tau} := \sum_g A_{\tau,g}\bar{Y}_g = \begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}.$$

Since Assumption 1 holds, we can appeal to Theorem 3 in Li and Ding (2017), which implies that $\mathbb{V}\text{ar}[\hat{\tau}] = \sum_g N_g^{-1} A_{\tau,g} S_g A'_{\tau,g} - N^{-1} S_\tau$, where $S_\tau = \mathbb{V}\text{ar}_f \left[\sum_g A_{\tau,g} Y_i(g) \right]$. The result then follows immediately from expanding this variance, as well as the observation that $S_\tau = \begin{pmatrix} S_\theta & 0 \\ 0 & 0 \end{pmatrix}$, where the 0 blocks are obtained by noting that $\sum_g A_{0,g} Y_i(g) = 0$ for all i by Assumption 2. $\qquad\square$

Proof of Proposition 2.2 To establish the proof, we first provide two lemmas that characterize the asymptotic joint distribution of $(\hat{\theta}_0, \hat{X}')'$, and show that $\hat{S}_g$ is consistent for S_g^* under Assumption 3. Both results are direct consequences of the general asymptotic results in Li and Ding (2017) for multi-valued treatments in randomized experiments.

Lemma A.1. *Under Assumptions 1, 2, and 3,*

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_0 - \theta \\ \hat{X} \end{pmatrix} \to_d \mathcal{N}(0, V^*),$$

where

$$V^* = \begin{pmatrix} \sum_g p_g^{-1} A_{\theta,g} S_g^* A'_{\theta,g} - S_\theta^* & \sum_g p_g^{-1} A_{\theta,g} S_g^* A'_{0,g} \\ \sum_g p_g^{-1} A_{0,g} S_g^* A'_{\theta,g} & \sum_g p_g^{-1} A_{0,g} S_g^* A'_{0,g} \end{pmatrix} =: \begin{pmatrix} V_{\hat{\theta}_0}^* & V_{\hat{\theta}_0,\hat{X}}^* \\ V_{\hat{X},\hat{\theta}_0}^* & V_{\hat{X}}^* \end{pmatrix},$$

and $S_\theta^* = \lim_{N \to \infty} S_\theta$ (where S_θ is defined in Proposition 2.1).

Proof. As in the proof to Proposition 2.1, we can write

$$\hat{\tau} = \sum_g A_{\tau,g} \bar{Y}_g = \begin{pmatrix} \hat{\theta}_0 \\ \hat{X} \end{pmatrix}.$$

The result then follows from Theorem 5 in Li and Ding (2017), combined with the observation noted in the proof to Proposition 2.1 that $S_\tau = \begin{pmatrix} S_\theta & 0 \\ 0 & 0 \end{pmatrix}$ and hence $S_\tau \to \begin{pmatrix} S_\theta^* & 0 \\ 0 & 0 \end{pmatrix}$. $\quad\square$

Lemma A.2. *Under Assumptions 1, 2, and 3, $\hat{S}_g \to_p S_g^*$ for all g.*

Proof. Follows immediately from Proposition 3 in Li and Ding (2017). $\qquad\square$

To complete the proof of Proposition 2.1, recall that $\hat{\beta}^* = \hat{V}_{\hat{X}}^{-1} \hat{V}_{\hat{X},\hat{\theta}_0}$. It is clear that $\hat{\beta}^*$ is a continuous function of $\hat{V}_{\hat{X}}$ and $\hat{V}_{\hat{X},\hat{\theta}_0}$, and that $\hat{V}_{\hat{X}}$ and $\hat{V}_{\hat{X},\hat{\theta}_0}$ are continuous functions of $\hat{S}_g$. From Lemma A.2 along with the continuous mapping theorem, we obtain that $\hat{\beta}^* \to_p (V_{\hat{X}}^*)^{-1} V_{\hat{X},\hat{\theta}_0}^*$. Lemma A.1 together with Slutsky's lemma then give that $\sqrt{N}(\hat{\theta}_{\hat{\beta}^*} - \theta) \to_d$

$\mathcal{N}\left(0, V_{\hat{\theta}_0}^* - V_{\hat{X},\hat{\theta}_0}^{*\prime}(V_{\hat{X}}^*)^{-1}V_{\hat{X},\hat{\theta}_0}^*\right)$. From Proposition 2.1, it is apparent that the asymptotic variance of $\hat{\theta}_{\hat{\beta}*}$ is equal to the limit of $N\mathbb{V}\mathrm{ar}\left[\hat{\theta}_{\beta*}\right]$, which completes the proof.

Proof of Lemma 2.2

Proof. Immediate from the fact that $\hat{S}_g \to_p S_g^*$ (see Lemma A.2) combined with the continuous mapping theorem. $\qquad\square$

Proof of Proposition 2.3

Proof. Note that, conditional on G, the distribution of t_π corresponds with the distribution of $\sqrt{N}\hat{\theta}^*/\sigma_{**}$ in a population with potential outcomes $Y^*(\cdot)$, where $Y_i^*(g) = Y_i(G_i)$ for all i, g. To prove the first assertion, it thus suffices to show that the populations defined by $Y^*(\cdot)$ satisfy Assumption 3, P_G-almost surely, in which case the result follows from Proposition 2.2 and Lemma A.4 applied to the population with potential outcomes $Y^*(\cdot)$.

Since the set of observations with $G_i = g$ is a simple random sample from a finite population, Lemma A5 in Wu and Ding (2020) implies that

$$\bar{Y}_g = \frac{1}{N_g}\sum_i 1[G_i = g]Y_i(g) \to_{a.s.} \lim_{N\to\infty}\mathbb{E}_f\left[Y_i(g)\right] =: \mu_g^*$$

$$\frac{1}{N_g - 1}\sum_i 1[G_i = g](Y_i(g) - \bar{Y}_g)^2 \to_{a.s.} \lim_{N\to\infty}\mathbb{V}\mathrm{ar}_f\left[Y_i(g)\right] =: S_g^*$$

In a slight abuse of notation, we will denote by $\mathbb{E}_f\left[Y_i^*(g)\right]$ the finite-population expectation in the population with potential outcomes $Y_i^*(g)$, where $Y_i^*(g) = Y_i(G_i)$. Now,

$$\mathbb{E}_f\left[Y_i^*(g)\right] = \frac{1}{N}\sum_i Y_i = \sum_g \frac{N_g}{N}\frac{1}{N_g}\sum_i 1[G_i = g]Y_i(g) \to_{a.s.} \sum_g p_g\mu_g^*$$

Similarly,

$$
\mathbb{V}\mathrm{ar}_f\left[Y_i^*(g)\right] = \frac{1}{N-1}\sum_i (Y_i - \bar{Y})^2
$$

$$
= \frac{N}{N-1}\left(\left(\sum_g \frac{N}{N_g}1[G_i = g]Y_i^2 - \frac{N}{N_g}\bar{Y}_g^2\right) + \left(\sum_g \frac{N_g}{N}\bar{Y}_g^2 - \left(\sum_g \frac{N_g}{N}\bar{Y}_g\right)^2\right)\right)
$$

$$
= \frac{N}{N-1}\left(\left(\sum_g \frac{N}{N_g}\frac{N_g-1}{N_g}\hat{S}_g\right) + \left(\sum_g \frac{N_g}{N}\bar{Y}_g^2 - \left(\sum_g \frac{N_g}{N}\bar{Y}_g\right)^2\right)\right)
$$

$$
\to_{a.s.} \sum_g p_g^{-1}S_g^* + \left(\sum_g p_g(\mu_g^*)^2 - \left(\sum_g p_g\mu_g^*\right)^2\right)
$$

where we obtain the convergence from the previous displays and the continuous mapping theorem (and we use the shorthand Y^2 for YY'). The first term in the limit is positive definite, since S_g^* is positive definite for each g by Assumption 3, and the second term is positive definite by Jensen's inequality. Hence, Assumption 3(ii) is satisfied for the population with potential outcomes Y_i^* P_g-almost surely. Finally, Assumption 3(iii) is satisfied P_g-almost surely by Lemma A6 in Wu and Ding (2020).

The second assertion then follows immediately from the fact that $\sqrt{N}(\hat{\theta}_{\hat{\beta}*} - \theta)/\hat{\sigma}_{**} \to_d \mathcal{N}(0, c)$, for $c = \sigma_*^2/(\sigma_*^2 + S_\theta^*) \leqslant 1$, by Proposition 2.2 and Lemma A.4. $\qquad\square$

A.1 Derivation of Variance Refinement

We now provide a derivation for the refined variance estimator discussed in Lemma 2.3, as well as a formal proof of its validity. First, recall that the Neyman-style variance estimator was conservative by $S_\theta^* = \lim_{N\to\infty} S_\theta$. We first provide a lemma which gives a consistently estimable lower bound on S_θ. Intuitively, this is the component of the treatment effect heterogeneity that is explained by lagged outcomes.

Lemma A.3. *Suppose that $A_{\theta,g} = 0$ for all $g < g_{min}$. If Assumption 2 holds, then*

$$
S_\theta = \mathbb{V}ar_f\left[\tilde{\theta}_i\right] + \frac{N+1}{N-1}\left(\sum_{g\geqslant g_{min}}\beta_g\right)'(MS_{g_{min}}M')\left(\sum_{g\geqslant g_{min}}\beta_g\right), \tag{9}
$$

where M is the matrix that selects the rows of Y_i corresponding with $t < g_{min}$; $\beta_g = (MS_gM')^{-1}MS_gA'_{\theta,g}$ is the coefficient from projecting $A_{\theta,g}Y_i(g)$ on $MY_i(g)$ (and a constant); and $\tilde{\theta}_i = \sum_{g\geqslant g_{min}} A_{\theta,g}Y_i(g) - \sum_{g\geqslant g_{min}}(MY_i(g))'\beta_g$.

Proof. For any g and functions of the potential outcomes $X_i \in \mathbb{R}^K$ and $Z_i \in \mathbb{R}$, let $\dot{X}_i = X_i - \mathbb{E}_f[X_i]$, $\dot{Z}_i = Z_i - \mathbb{E}_f[Z_i]$, and $\beta_{XZ} = \mathbb{V}\mathrm{ar}_f[X_i]^{-1}\mathbb{E}_f\left[\dot{X}_i\dot{Z}_i\right]$. Observe that

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}_f[Z_i - \beta'_{XZ}X_i] &= \frac{1}{N-1}\sum_i \left(\dot{Z}_i - \beta'_{XZ}\dot{X}_i\right)^2 \\
&= \frac{1}{N-1}\sum_i \dot{Z}_i^2 + \beta'_{XZ}\left(\frac{1}{N-1}\sum_i \dot{X}_i\dot{X}_i'\right)\beta_{XZ} - \beta'_{XZ}\frac{2}{N-1}\sum_i \dot{X}_i\dot{Z}_i \\
&= \mathbb{V}\mathrm{ar}_f[Z_i] + \beta'_{XZ}\mathbb{V}\mathrm{ar}_f[X_i]\beta_{XZ} - 2\frac{N}{N-1}\beta'_{XZ}\mathbb{V}\mathrm{ar}_f[X_i]\beta_{XZ} \\
&= \mathbb{V}\mathrm{ar}_f[Z_i] - \frac{N+1}{N-1}\beta'_{XZ}\mathbb{V}\mathrm{ar}_f[X_i]\beta_{XZ}.
\end{aligned}
$$

The result then follows from setting $Z_i = \sum_{g \geqslant g_{min}} A_{\theta,g}Y_i(g) = \theta_i$ and $X_i = MY_i(g_{min})$, and noting that under Assumption 2, $MY_i(g_{min}) = MY_i(g)$ for all $g \geqslant g_{min}$, and hence $\mathbb{V}\mathrm{ar}_f[MY_i(g_{min})] = M\hat{S}_{g_{min}}M' = MS_gM' = \mathbb{V}\mathrm{ar}_f[MY_i(g)]$. $\qquad\square$

We now formally define the refined estimator $\hat{\sigma}_{**}$ and give a more detailed statement of Lemma 2.3.

Lemma A.4. *Suppose that $A_{\theta,g} = 0$ for all $g < g_{min}$ and Assumptions 1-3 hold. Let M be the matrix that selects the rows of Y_i corresponding with periods $t < g_{min}$. Define*

$$
\hat{\sigma}_{**}^2 = \hat{\sigma}_*^2 - \left(\sum_{g>g_{min}}\hat{\beta}_g\right)'\left(M\hat{S}_{g_{min}}M'\right)\left(\sum_{g>g_{min}}\hat{\beta}_g\right), \tag{10}
$$

where $\hat{\beta}_g = (M\hat{S}_gM')^{-1}M\hat{S}_gA'_{\theta,g}$. Then $\hat{\sigma}_{**}^2 \to_p \sigma_*^2 + S_{\tilde{\theta}}^*$, where $0 \leqslant S_{\tilde{\theta}}^* \leqslant S_\theta^*$, so that $\hat{\sigma}_{**}$ is asymptotically (weakly) less conservative than $\hat{\sigma}_*$. (See Lemma A.3 for a closed-form expression for $S_{\tilde{\theta}}^* = \mathbb{V}\mathrm{ar}_f\left[\tilde{\theta}_i\right]$.)

Proof of Lemma A.4

Proof. Note that $\hat{\beta}_g$ is a continuous function of $\hat{S}_g$. Lemma A.2 together with the continuous mapping theorem thus imply that

$$
\left(\sum_{g>g_{min}}\hat{\beta}_g\right)'\left(M\hat{S}_{g_{min}}M'\right)\left(\sum_{g>g_{min}}\hat{\beta}_g\right) - \left(\sum_{g>g_{min}}\beta_g\right)'(MS_{g_{min}}M')\left(\sum_{g>g_{min}}\beta_g\right) \to_p 0.
$$

From Lemmas 2.2 and A.3, it is then immediate that $\sigma_{**}^2 \to_p \sigma_*^2 + S_{\tilde{\theta}}^*$, where $S_{\tilde{\theta}}^* = \lim_{N\to\infty}\mathbb{V}\mathrm{ar}_f\left[\tilde{\theta}_i\right] \leqslant \lim_{N\to\infty}S_\theta = S_\theta^*$. $\qquad\square$

B   Additional Simulation Results

This section presents results from extensions to the simulations in Section 3.

Other outcomes.   Appendix Tables 1-4 show results analogous to those in the main text, except using the other two outcomes considered in our application (use of force and sustained complaints). We again find that the plug-in efficient estimator has minimal bias and is substantially more precise than the CS and SA estimators for nearly all specifications (with reductions in standard deviations relative to CS by a factor of over 3 for some specifications). Likewise, both t-based and FRT-based approaches yield reliable inference in nearly all specifications. The one exception to the good performance of the plug-in efficient estimator is the calendar-weighted average for sustained complaints when using the monthly data: the coverage of t-based CIs for the plug-in efficient estimator is only 79% in this specification (although the FRT has size of only 6%). Two distinguishing features of this specification are that the outcome is very rare (pre-treatment mean 0.004) and the aggregation scheme places the largest weight on the earliest three cohorts, which were small (sizes 17,15,26). This finding aligns with the well-known fact that the central limit theorem may be a poor approximation in finite samples with a binary outcome that is very rare. The t-based CI for the plug-in efficient estimator again has good coverage (94%) when considering the annualized data where the cohort sizes are larger (see below). We thus urge some caution in using the plug-in efficient estimator with t-based CIs (or any procedure based on a normal approximation) when cohort sizes are small (<30) and the outcome is rare (mean < 0.01). If average effects are of interest in such settings, we recommend collapsing the data to a higher level of aggregation so that the cohorts are larger before using the plug-in estimator. Alternatively, if the sharp null hypotheses is of interest, one can rely on the FRT, which is exact (up to simulation error) under the sharp null.

Annualized data.   Appendix Tables 5-10 present simulations from an alternative specification where the monthly data is collapsed to the yearly level, so that there are six total time periods and five (larger) cohorts. The plug-in efficient estimator has minimal bias and both t-based and FRT-based methods yield reliable inference for all specifications, including the calendar-weighted average for sustained complaints. The plug-in efficient estimator again dominates the other estimators in efficiency, although the gains are smaller (e.g. 24 to 30% reductions in standard deviation relative to CS for complaints). The smaller efficiency gains in this specification are intuitive: the CS and SA estimators over-weight the pre-treatment periods (relative to the plug-in efficient estimator) in our setting, but the penalty for doing this is smaller in the collapsed data, where the pre-treatment outcomes are averaged over

more months and thus have lower variance.

Augmented $\hat{X}$.   Appendix Table 11 shows results for an alternative version of the plug-in efficient estimator where $\hat{X}$ is now a vector that contains the difference in means between cohort g and g' in all periods $t < min(g, g')$.[26] We find poor coverage of t-based CIs for this estimator in the monthly specification, where the dimension of $\hat{X}$ is large relative to the sample size (1987, compared with $N = 7785$), and thus the normal approximation derived in Proposition 2.2 is poor.[27] By contrast, when the data is collapsed to the yearly level, and thus the dimension of $\hat{X}$ constructed in this way is more modest (10), the coverage for this estimator is good, and it offers small efficiency gains (up to 3%) over the scalar $\hat{X}$ considered in the main text. These findings align with the results in Lei and Ding (2020), who show (under certain regularity conditions) that covariate-adjustment in cross-sectional experiments yields asymptotically normal estimators when the dimensions of the covariates is $o(N^{-\frac{1}{2}})$. We thus recommend using the version of $\hat{X}$ with all potential comparisons only when its dimension is small relative to the square root of the sample size.

Heterogeneous Treatment Effects.   Appendix Tables 12 and 13 show simulation results for a modification of our baseline specification in which there are heterogeneous treatment effects. In the baseline specification, $Y_i(g) = Y_i(\infty)$ for all g. In the modification, we set $Y_i(g) = Y_i(\infty) + 1[t >= g] \cdot u_i$. The u_i are mean-zero draws drawn from a normal distribution with standard deviation equal to the standard deviation of the untreated potential outcomes. We draw the u_i once and hold them fixed throughout the simulations, which differ only in the assignment of treatment timing. The relative efficiency of the estimators is similar to those for the main specification, although as expected, both t-based and FRT-based approaches to inference tend to be conservative.

---

[26]Calculation is more intensive using the longer $\hat{X}$, so we use 50 simulated permutations for the FRT, instead of the 500 used for the other specifications.

[27]The reported standard deviations of the estimator with augmented $\hat{X}$ are extremely large in the top panel of Appendix Table 11. However, this is mainly the result of a small number of outliers, possibly driven by numerical precision errors in computing $\hat{\beta}^*$ (which is more difficult in larger dimensions). For example, the first row the table reports an SD of 23.89, but the same SD would be only 4.5 if we excluded the most extreme 1% in either direction. Nevertheless, the poor coverage of t-based confidence intervals cannot be driven only by outliers.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.03	0.94	0.06	0.30	0.32
PlugIn	cohort	0.02	0.92	0.05	0.28	0.29
PlugIn	ES0	0.01	0.96	0.05	0.28	0.28
PlugIn	simple	0.01	0.93	0.06	0.26	0.27
CS	calendar	0.03	0.95	0.04	0.59	0.60
CS	cohort	0.01	0.96	0.04	0.45	0.44
CS/dCDH	ES0	0.01	0.96	0.04	0.37	0.37
CS	simple	0.01	0.96	0.04	0.45	0.44
SA	calendar	0.05	0.92	0.05	1.39	1.50
SA	cohort	0.03	0.90	0.05	1.43	1.54
SA	ES0	0.02	0.96	0.05	0.84	0.89
SA	simple	0.04	0.89	0.05	1.54	1.68

Appendix Table 1: Results for Simulations Calibrated to Wood et al. (2020b) – Use of Force

Note: This table shows results analogous to Table 2, except using Use of Force rather than Complaints as the outcome.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.88	4.72
cohort	1.51	5.25
ES0	1.34	3.23
simple	1.65	6.26

Appendix Table 2: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Use of Force

Note: This table shows results analogous to Table 3, except using Use of Force rather than Complaints as the outcome.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.00	0.79	0.06	0.06	0.07
PlugIn	cohort	0.00	0.92	0.06	0.03	0.03
PlugIn	ES0	0.01	0.95	0.05	0.08	0.08
PlugIn	simple	0.00	0.92	0.06	0.03	0.03
CS	calendar	0.01	0.95	0.06	0.14	0.17
CS	cohort	0.01	0.95	0.05	0.11	0.11
CS/dCDH	ES0	0.01	0.94	0.06	0.11	0.12
CS	simple	0.01	0.96	0.04	0.11	0.12
SA	calendar	0.00	0.83	0.05	0.33	0.39
SA	cohort	0.00	0.61	0.05	0.33	0.41
SA	ES0	0.01	0.97	0.05	0.22	0.27
SA	simple	0.00	0.63	0.05	0.35	0.44

Appendix Table 3: Results for Simulations Calibrated to Wood et al. (2020b) – Sustained Complaints

Note: This table shows results analogous to Table 2, except using Sustained Complaints rather than Complaints as the outcome.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	2.58	5.82
cohort	3.58	13.24
ES0	1.42	3.35
simple	3.74	14.45

Appendix Table 4: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Sustained Complaints

Note: This table shows results analogous to Table 3, except using Sustained Complaints rather than Complaints as the outcome.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.11	0.95	0.05	1.99	1.96
PlugIn	cohort	0.15	0.95	0.04	2.53	2.49
PlugIn	ES0	0.03	0.96	0.05	1.65	1.60
PlugIn	simple	0.14	0.95	0.04	2.41	2.37
CS	calendar	0.20	0.96	0.05	2.65	2.56
CS	cohort	0.26	0.96	0.05	3.24	3.13
CS/dCDH	ES0	0.04	0.96	0.05	2.05	1.98
CS	simple	0.27	0.96	0.05	3.17	3.05
SA	calendar	0.32	0.95	0.05	4.10	3.95
SA	cohort	0.40	0.96	0.04	4.36	4.22
SA	ES0	0.21	0.96	0.04	3.31	3.26
SA	simple	0.41	0.95	0.05	4.58	4.43

Appendix Table 5: Results for Simulations Calibrated to Wood et al. (2020b) – Annualized Data

Note: This table shows results analogous to Table 2, except the data is collapsed to the annual level.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.30	2.01
cohort	1.26	1.69
ES0	1.24	2.04
simple	1.29	1.87

Appendix Table 6: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Annualized Data

Note: This table shows results analogous to Table 3, except the data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	-0.01	0.94	0.06	2.23	2.27
PlugIn	cohort	0.01	0.93	0.05	2.81	2.84
PlugIn	ES0	-0.01	0.95	0.04	1.76	1.78
PlugIn	simple	0.00	0.93	0.05	2.70	2.73
CS	calendar	-0.03	0.94	0.06	2.83	2.88
CS	cohort	-0.01	0.94	0.05	3.46	3.48
CS/dCDH	ES0	-0.05	0.94	0.06	2.10	2.11
CS	simple	0.00	0.95	0.05	3.41	3.42
SA	calendar	0.05	0.94	0.05	4.38	4.39
SA	cohort	0.09	0.95	0.05	4.76	4.72
SA	ES0	0.07	0.95	0.05	3.54	3.48
SA	simple	0.10	0.95	0.05	4.99	4.95

Appendix Table 7: Results for Simulations Calibrated to Wood et al. (2020b) – Use of Force & Annualized Data

Note: This table shows results analogous to Table 2, except using Use of Force rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.27	1.94
cohort	1.23	1.66
ES0	1.19	1.95
simple	1.25	1.81

Appendix Table 8: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Use of Force & Annualized Data

Note: This table shows results analogous to Table 3, except using Use of Force rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	0.00	0.95	0.05	0.43	0.44
PlugIn	cohort	-0.01	0.94	0.05	0.53	0.55
PlugIn	ES0	0.01	0.95	0.05	0.45	0.45
PlugIn	simple	-0.01	0.94	0.05	0.51	0.52
CS	calendar	0.02	0.96	0.04	0.69	0.66
CS	cohort	0.03	0.96	0.04	0.81	0.78
CS/dCDH	ES0	0.01	0.96	0.04	0.61	0.60
CS	simple	0.02	0.96	0.04	0.80	0.77
SA	calendar	0.01	0.95	0.05	1.08	1.06
SA	cohort	0.02	0.96	0.04	1.11	1.08
SA	ES0	0.00	0.95	0.05	0.96	0.99
SA	simple	0.02	0.96	0.04	1.18	1.16

Appendix Table 9: Results for Simulations Calibrated to Wood et al. (2020b) – Sustained Complaints & Annualized Data

Note: This table shows results analogous to Table 2, except using Sustained Complaints rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.51	2.42
cohort	1.42	1.97
ES0	1.34	2.19
simple	1.47	2.23

Appendix Table 10: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Sustained Complaints & Annualized Data

Note: This table shows results analogous to Table 3, except using Sustained Complaints rather than Complaints as the outcome, and in simulations where data is collapsed to the annual level.

(a) Monthly Data

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn - Long X	calendar	1.39	0.01	0.06	0.13	23.89
PlugIn - Long X	cohort	0.46	0.01	0.07	0.04	31.49
PlugIn - Long X	ES0	1.18	0.02	0.07	0.21	104.97
PlugIn - Long X	simple	1.07	0.00	0.07	0.04	33.23
PlugIn	calendar	0.00	0.93	0.06	0.27	0.29
PlugIn	cohort	0.00	0.92	0.06	0.24	0.24
PlugIn	ES0	0.01	0.94	0.05	0.26	0.27
PlugIn	simple	0.00	0.92	0.06	0.22	0.22

(b) Annual Data

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn - Long X	calendar	0.33	0.94	0.06	1.93	1.95
PlugIn - Long X	cohort	0.37	0.93	0.06	2.47	2.49
PlugIn - Long X	ES0	0.25	0.95	0.06	1.59	1.56
PlugIn - Long X	simple	0.38	0.94	0.06	2.33	2.36
PlugIn	calendar	0.11	0.95	0.05	1.99	1.96
PlugIn	cohort	0.15	0.95	0.04	2.53	2.49
PlugIn	ES0	0.03	0.96	0.05	1.65	1.60
PlugIn	simple	0.14	0.95	0.04	2.41	2.37

Appendix Table 11: Performance of Plug-In Efficient Estimator Using Augmented $\hat{X}$

Note: This table shows the bias, coverage, mean standard error, and standard deviation of two versions of the plug-efficient estimator. The estimator with the label "Long X" uses an augmented version of $\hat{X}$ that includes the difference in means between all cohorts g, g' in periods $t < min(g, g')$. The estimator labeled PlugIn uses a scalar $\hat{X}$ such that the CS estimator corresponds with $\beta = 1$, as in the main text. The simulation specification in panel (a) is the baseline specification considered in the main text; in panel (b), the data is collapsed to the annual level.

Estimator	Estimand	Bias	Coverage	FRT Size	Mean SE	SD
PlugIn	calendar	-0.03	0.96	0.03	0.55	0.52
PlugIn	cohort	0.00	1.00	0.00	0.33	0.24
PlugIn	ES0	0.01	0.99	0.01	0.35	0.27
PlugIn	simple	-0.01	1.00	0.00	0.33	0.23
CS	calendar	-0.03	0.96	0.03	0.74	0.68
CS	cohort	-0.01	0.97	0.03	0.47	0.41
CS/dCDH	ES0	0.01	0.98	0.02	0.43	0.36
CS	simple	-0.01	0.98	0.02	0.47	0.41
SA	calendar	0.03	0.96	0.03	1.39	1.36
SA	cohort	0.05	0.93	0.06	1.36	1.38
SA	ES0	0.03	0.95	0.04	0.86	0.89
SA	simple	0.06	0.92	0.06	1.48	1.49

Appendix Table 12: Results for Simulations Calibrated to Wood et al. (2020b) – Heterogeneous Treatment Effects

Note: This table shows results analogous to Table 2, except the DGP adds heterogeneous treatment effect as described in Section B.

|	Ratio of SD to Plug-In	
Estimand	CS	SA
calendar	1.88	4.72
cohort	1.51	5.25
ES0	1.34	3.23
simple	1.65	6.26

Appendix Table 13: Comparison of Standard Deviations – Callaway and Sant'Anna (2020) and Sun and Abraham (2020) versus Plug-in Efficient Estimator – Heterogeneous Treatment Effects

Note: This table shows results analogous to Table 3, except the DGP adds heterogeneous treatment effect as described in Section B.

# C Additional Application Results

This section contains additional results pertaining to our application in Section 4. Appendix Figure 1 shows an event-plot analogous to Figure 2 except using the CS estimator.

Appendix Figure 1: Event-Time Average Effects Using the CS Estimator
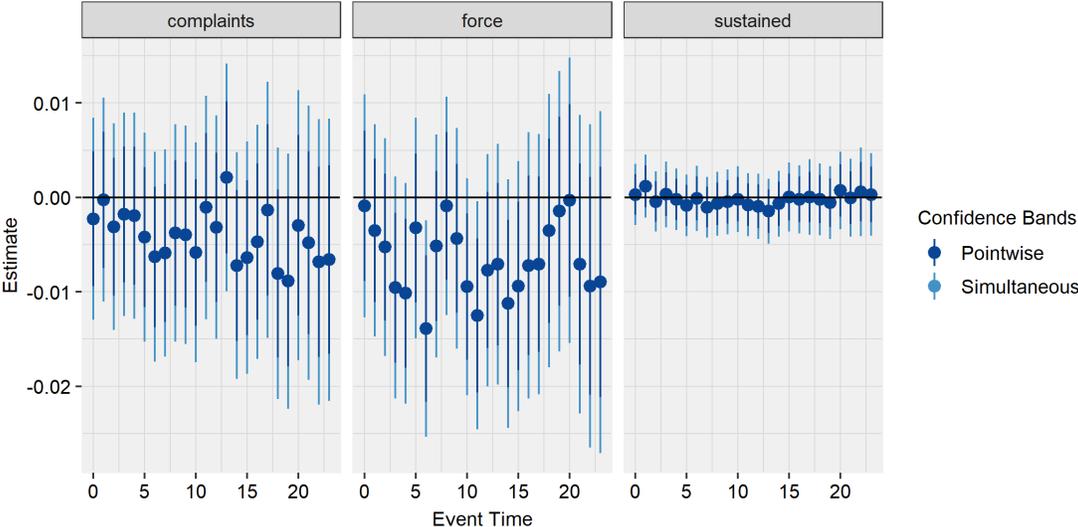
In Appendix Figure 2, we present results analogous to those in Figure 1 except removing officers who were treated in the last 12 months of the data. The reason for this is, as discussed in the supplement to Wood et al. (2020a), there was some non-compliance towards the end of the study period wherein officers who had not already been trained could volunteer to take the training at a particular date. The qualitative patterns after dropping these observations are similar, although the estimates for the effect on use of force are closer to zero and not statistically significant at conventional levels.

Appendix Figure 2: Effect of Procedural Justice Training Using the Plug-In Efficient and Callaway and Sant'Anna (2020) Estimators – Dropping Late-Trained Officers
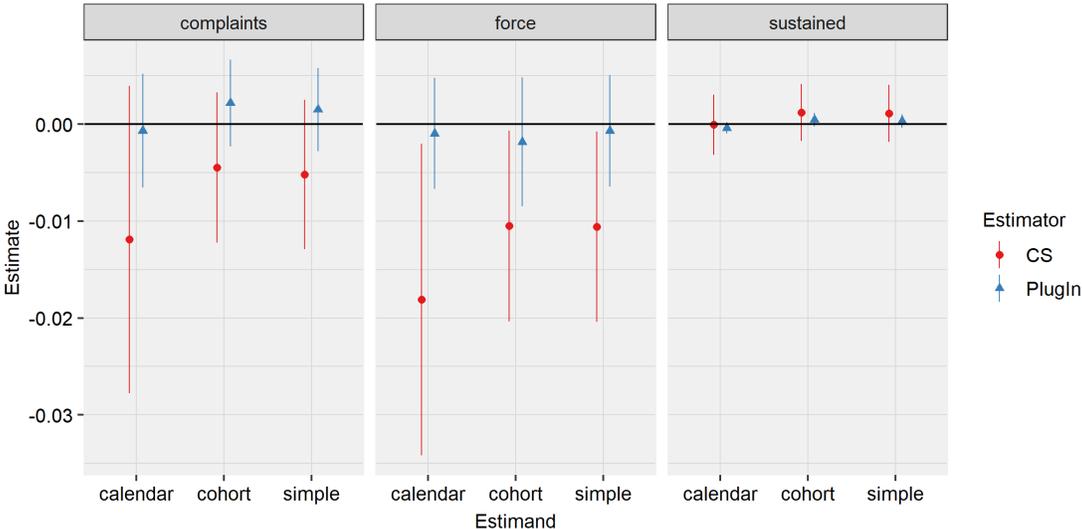
Note: This figure is analogous to Figure 1, except we remove from the data officers trained in the last 12 months of the data owing to concerns about treatment non-compliance.

Appendix References

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, December 2020.

Lei, Lihua and Peng Ding, "Regression adjustment in completely randomized experiments with a diverging number of covariates," *Biometrika*, December 2020, (Forthcoming).

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.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, December 2020.

Wood, George, Tom R. Tyler, and Andrew V. Papachristos, "Procedural justice training reduces police use of force and complaints against officers," *Proceedings of the National Academy of Sciences*, May 2020, *117* (18), 9815–9821.

— , — , — , **Jonathan Roth, and Pedro H.C. Sant'Anna**, "Revised Findings for "Procedural justice training reduces police use of force and complaints against officers"," *Working Paper*, 2020.

Wu, Jason and Peng Ding, "Randomization Tests for Weak Null Hypotheses in Randomized Experiments," *Journal of the American Statistical Association*, May 2020, pp. 1–16. arXiv: 1809.07419.