# Efficient Estimation for Staggered Rollout Designs[*]

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

March 23, 2021

**Abstract**

Researchers are frequently interested in the causal effect of a treatment that is (quasi-)randomly rolled out to different units at different points in time. This paper studies how to efficiently estimate a variety of causal parameters in a Neymanian-randomization based framework of random treatment timing. We solve for the most efficient estimator in a class of estimators that nests two-way fixed effects models as well as several popular generalized difference-in-differences methods. The efficient estimator is not feasible in practice because it requires knowledge of the optimal weights to be placed on pre-treatment outcomes. However, the optimal weights can be estimated from the data, and in large datasets the plug-in estimator that uses the estimated weights has similar properties to the "oracle" efficient estimator. We illustrate the performance of the plug-in efficient estimator in simulations and in an application to Wood et al. (2020a,b)'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 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.

[*] We are grateful to Brantly Callaway, Peng Ding, Avi Feller, Emily Owens, Ryan Hill, Lihua Lei, Ashesh Rambachan, Evan Rose, Adrienne Sabety, Jesse Shapiro, Yotam Shem-Tov, Dylan Small, Ariella Kahn-Lang Spitzer, and seminar participants at the Berkeley Causal Inference Reading Group, University of Cambridge, University of Delaware, and University of Florida for helpful comments and conversations. We thank Madison Perry for excellent research assistance.

[†] Microsoft. Jonathan.Roth@microsoft.com

[‡] Vanderbilt University. pedro.h.santanna@vanderbilt.edu

# 1 Introduction

Across a variety of domains, researchers are interested in the causal effect of a treatment that has a staggered rollout, meaning that it is first implemented for different units at different times. Social scientists frequently study the causal effect of a policy that is implemented in different locations at different times. Businesses may likewise be interested in the causal effect of a new feature or advertising campaign that is introduced to different customers over time. And clinical trials increasingly use a "stepped wedge" design in which a treatment is first given to patients at different points in time.

In many cases, the timing of the rollout is controlled by the researcher and can be explicitly randomized. Randomizing treatment timing is a natural way to learn about causal effects in settings where capacity or administrative constraints prevent treating everyone at once, while simultaneously allowing everyone to ultimately receive treatment. In other settings, the researcher cannot directly control the timing of treatment, but may argue that the timing of the treatment is as-if randomly assigned.[1]

Two common approaches to estimate treatment effects in such contexts are two-way fixed effects (TWFE) models that control for unit and time fixed effects (Xiong et al., 2019) and mixed-effects linear regression models (Hussey and Hughes, 2007). There are concerns, however, about how to interpret the estimates from such methods when the estimating model may be mis-specified, for example if treatment effects are dynamic or vary across individuals. A large recent literature in econometrics has highlighted that the estimand of TWFE models is difficult to interpret when there are heterogeneous treatment effects (Athey and Imbens, 2018; Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2018; Imai and Kim, 2020; Sun and Abraham, 2020). As a result, several recent papers have proposed methods that yield more easily interpretable estimands and effectively highlight treatment effect heterogeneity under a generalized parallel trends assumption (Callaway and Sant'Anna, 2020; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2020). Lindner and Mcconnell (2021) raise similar concerns about the interpretability of

[1] Over half (20 of 38) of the papers with staggered treatment timing in Roth (2020)'s survey of recent papers in leading economics journals using difference-in-differences and related methods refer to the timing of treatment as "quasi-random" or "quasi-experimental".

mixed-effects linear models under mis-specification, and instead recommend the use of Sun and Abraham (2020)'s estimator for stepped-wedge designs. However, these new estimators exploit a generalized parallel trends assumption, which is technically weaker than the assumption of random treatment timing. This suggests that it might be possible to obtain more precise estimates by more fully exploiting the random timing of treatment.

This paper studies the efficient estimation of treatment effects in a Neymanian randomization framework of random treatment timing. We consider the estimation of a variety of causal parameters that are easily interpretable under treatment effect heterogeneity, and solve for the most efficient estimator in a large class of estimators that nests many existing approaches as special cases. As in the literature on model-assisted estimation (Lin, 2013; Breidt and Opsomer, 2017), our proposed procedure is asymptotically valid under the assumption of random treatment timing, regardless of whether the model is mis-specified.

We begin by introducing a design-based framework that formalizes the notion that treatment timing is (as-if) 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. We make two key assumptions in this model. First, we assume that the treatment timing $G_i$ is (as-if) 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.

Under these assumptions, outcomes in periods before a unit is treated play a similar role to fixed pre-treatment covariates in a cross-sectional randomized experiment. In fact, we show that our setting is isomorphic to a cross-sectional randomized experiment in the special case with two periods ($T = 2$) when units are either treated in period two or never treated ($\mathcal{G} = \{2, \infty\}$). Our results thus nest previous results on covariate adjustment in randomized experiments (Freedman, 2008b,a; Lin, 2013; Li and Ding, 2017) as a special case. Our key theoretical contribution is extending these results to settings with staggered treatment timing, which poses technical challenges since a different number of pre-treatment outcomes are observed for units treated at different times. We repeatedly return to the special two-period case to build intuition and to connect our more general results to the previous literature.

In our staggered adoption setting, treatment effects may vary both over calendar time and time since treatment. We therefore consider a large class of possible causal parameters that highlight treatment effect heterogeneity across different dimensions. Specifically, 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, $\tau_{2,2\infty}$ 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 consider the large class of estimators that start with a sample analog to the target parameter and then 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 in $\hat{\theta}_\beta$ replaces the $\tau_{t,gg'}$ with their sample analogs in the definition of $\theta$, and the second term adjusts for a linear function of a vector $\hat{X}$, which compares outcomes for cohorts treated at different dates at points in time before either was treated. For example, in the simple two-period case, $\hat{X}$ corresponds with the average difference in outcomes at period one between units treated at period two and never-treated units. In this case, the estimator $\hat{\theta}_1$ corresponds with the canonical difference-in-differences estimator, whereas $\hat{\theta}_0$ corresponds with the simple difference-in-means. More generally, 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 $\theta$ under the assumptions of random treatment timing and no anticipation.

We then derive the most efficient estimator in this class. The optimal coefficient $\beta^*$ depends on covariances between the potential outcomes over time, and thus the estimators proposed in the literature will only be efficient for special covariance structures. Unfortunately, the covariances of the potential outcomes are generally not known ex ante, and so the efficient estimator is infeasible in practice. However, as in Lin (2013)'s analysis of covariate adjustment in cross-sectional randomized experiments, one can estimate a "plug-in" version of the efficient estimator that replaces the "oracle" coefficient $\beta^*$ 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 cross-sectional experiments. We also show how the covariance can be (conservatively) estimated. In a Monte Carlo study calibrated to our application, we find that confidence intervals based on the plug-in efficient estimator have good coverage and are substantially shorter than the procedures of Callaway and Sant'Anna (2020), Sun and Abraham (2020), and de Chaisemartin and D'Haultfœuille (2020).[2]

As an illustration of our method and standalone empirical contribution, we revisit the data from Wood et al. (2020a,b), who studied the randomized rollout of a procedural justice training program in Chicago. As in Wood et al. (2020b), we find limited evidence that the program reduced complaints against police officers and borderline significant effects on officer use of force. However, the use of our proposed methodology allows us to obtain substantially more precise estimates of the effect of the training program: 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).

**Related Literature.** Our work builds on results on covariate-adjustment in cross-sectional randomized experiments (Freedman, 2008a,b; Lin, 2013; Li and Ding, 2017) to develop efficient estimators of a variety of average causal parameters in a Neymanian-randomization framework of staggered treatment timing.

We contribute to an active literature on difference-in-differences and related methods with staggered treatment timing. As mentioned earlier, several recent papers have illustrated that the estimand of standard TWFE models may not have an intuitive causal interpretation when there are heterogeneous treatment effects, and new estimators for more sensible causal

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

estimands have been introduced. These new estimators typically rely on a generalized parallel trends assumption. By contrast, we consider the problem of efficient estimation under the stronger assumption of random treatment timing, and obtain an estimator that (under suitable regularity conditions) is asymptotically more precise than many of the proposals in the literature under this assumption. Unlike existing approaches, however, our approach need not be valid in observational settings where researchers are confident in parallel trends but not in random treatment timing.[3]

In contrast to much of the difference-in-differences literature, which takes a model-based perspective to uncertainty, our Neymanian randomization framework is design-based. Athey and Imbens (2018) adopt a design-based framework similar to ours, but consider the interpretation of the estimand of two-way fixed effects models rather than efficient estimation. Shaikh and Toulis (2019) consider inference on sharp null hypotheses in a design-based model where treatment timing is random conditional on observables and the probability of different units being treated at the same time is zero; by contrast, we consider inference on average causal effects under unconditional random treatment timing in a setting where multiple units begin treatment at the same time.

Several previous papers have analyzed the efficiency of difference-in-differences relative to other methods in a two-period setting similar to our ongoing example.[4] Frison and Pocock (1992) and McKenzie (2012) compare difference-in-differences to an estimator that has the same asymptotic efficiency as our proposed estimator under homogeneous treatment effects, but will generally be less efficient under treatment effect heterogeneity; see Remark 4 for more details and connections to the literature on the Analysis of Covariance. Neither of these papers considers a design-based framework, nor do they study the more general case of staggered treatment timing that is our primary focus.

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;

[3] Roth and Sant'Anna (2021) show that if treatment timing is not random, then the parallel trends assumption will be sensitive to functional form without strong assumptions on the full distribution of potential outcomes.

[4] Ding and Li (2019) show a bracketing relationship between the biases of difference-in-differences and other estimators in the class we consider when treatment timing is not random, but do not consider efficiency under random treatment timing.

Davey et al., 2015; Turner et al., 2017)). As discussed by Lindner and Mcconnell (2021), the dominant approach in this literature is to use mixed-effects linear regression models (Hussey and Hughes, 2007) to estimate a common post-treatment effect, but such approaches are susceptible to model mis-specification (Thompson et al., 2017) and are not suitable for disentangling treatment effect heterogeneity. By contrast, the efficient estimator we propose does not rely on distributional restrictions on the outcome, can be used to effectively highlight treatment effect heterogeneity, and is generally more efficient than the Sun and Abraham (2020) procedure recommended by Lindner and Mcconnell (2021). Our efficient estimator can be applied directly in stepped wedge designs with individual-level treatment assignment, and we discuss extensions to clustered assignment in Remark 2. Our approach is complementary to Ji et al. (2017), who propose using randomization-based inference procedures to test Fisher's sharp null hypothesis in stepped wedge designs, whereas we adopt a Neymanian randomization-based approach for inference on average causal effects. Our proposed efficient estimator differs from the one adopted by Ji et al. (2017), as well.

Our work is also 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 $G_i$ corresponds with the first period in which unit $i$ is treated (and $G_i = \infty$ denotes that a unit is never treated). We assume that treatment is an absorbing state.[5] 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

[5] 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.

$Y_i(g) = (Y_{i1}(g), ..., Y_{iT}(g))' \in \mathbb{R}^T$. 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 (2018) 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. All expectations ($\mathbb{E}[\cdot]$) and probability statements ($\mathbb{P}(\cdot)$) are taken over the distribution of $G$ conditional on the number of units treated at each period, $(N_g)_{g \in \mathcal{G}}$, and the potential outcomes, 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 $\mathbb{V}\text{ar}_f[W_i] = (N-1)^{-1} \sum_i (W_i - \mathbb{E}_f[W_i])(W_i - \mathbb{E}_f[W_i])'$ the finite-population expectation and variance of $W_i$.

Our first main assumption is that the treatment timing is (as-if) randomly assigned.

**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.*

**Remark 1** (Stratified Treatment Assignment)**.** For simplicity, we consider the case of unconditional random treatment timing. In some settings, the treatment timing may be randomized among units with some shared observable characteristics (e.g. counties within a state). In such cases, the methodology developed below can be applied to form efficient estimators for each stratum, and the stratum-level estimates can then be pooled to form aggregate estimates for the population.

**Remark 2** (Stepped Wedge Design)**.** The phrase "stepped wedge design" is used to refer to a clinical trial with a staggered rollout, typically in which all units are eventually treated ($\infty \notin \mathcal{G}$). This directly corresponds with our set-up if treatment is randomized at the individual level. Frequently, however, treatment timing may be clustered in the stepped wedge design — e.g. treatment is assigned to families $f$, and all units $i$ in family $f$ are first treated at the same time, which violates Assumption 1. 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.

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 (Malani and Reif, 2015).

**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 as well as the time that has elapsed since treatment. Rather, 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 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. We will therefore come back to this example throughout the paper to provide intuition and connect our results to the previous literature.

## 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 (2018), 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

$$\theta = \sum_{t,g,g'} a_{t,gg'} \tau_{t,gg'}, \tag{1}$$

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 general $a_{t,gg'}$.[6] The results extend easily to vector-valued $\theta$'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, which we have shown is analogous to a cross-sectional experiment in period two, 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

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

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 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 treatment effects on the treated population rather than average treatment effects (in a sampling-based framework). In some cases, the $ATE(t,g)$ will be directly of interest and can be estimated directly 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. For example, the following aggregation schemes mirror those proposed in Callaway and Sant'Anna (2020) for the $ATT(t,g)$, and may be intuitive in a variety of contexts. We define the simple-weighted ATE to be the simple weighted average of the $ATE(t,g)$, where each $ATE(t,g)$ is weighted by the cohort size $N_g$,

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

Likewise, we define the cohort- and time-specific weighted averages as

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

and introduce the summary parameters

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

Finally, we introduce "event-study" parameters that aggregate the treatment effects at a given lag $l$ since treatment

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

Note that the instantaneous parameter $\theta_0^{ES}$ is analogous to the estimand considered in 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).[7]

These different aggregate causal parameters can be to used to highlight different types of treatment effect heterogeneity. For instance, when researchers want to better understand how the average treatment effect evolves with respect to the time elapsed since treatment started, $l$, they can focus their attention on $\theta_l^{ES}$ $(l = 0, 1, ...)$. 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), in which case the $\theta_t$ may be of interest. Likewise, if one is interested in comparing the average effect of first being treated at different times, then comparing $\theta_g$ across $g$ is natural. When researchers are interested in a single summary parameter of the treatment effect, it is natural to further aggregate across times and treatment dates, and the parameters $\theta^{simple}, \theta^{calendar}, \theta^{cohort}$ provide aggregations that weight differently across both calendar time and time since treatment. 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.

## 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 sample mean 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

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

which replaces the population means in the definition of $\theta$ with their sample analogues.

[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.

We will consider estimators of the form

$$\hat{\theta}_\beta = \hat{\theta}_0 - \hat{X}'\beta \tag{2}$$

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 simple 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 $\beta$ 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 difference in sample means in period one between the units first treated at period two and the never-treated units. Thus,

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

[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.

is the canonical difference-in-differences estimator, where $\bar{Y}_{g,t}$ represents the sample mean of $Y_{it}$ for units with $G_i = g$. Likewise, $\hat{\theta}_0$ is the simple difference-in-means in period two, $(\bar{Y}_{2,2} - \bar{Y}_{2,\infty})$. More generally, the estimator $\hat{\theta}_\beta$ takes the simple difference-in-means in period two and adjusts by $\beta$ times the difference-in-means in period one. 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). In particular, 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 $\beta = \frac{N_2}{N}\beta_0 + \frac{N_\infty}{N}\beta_1$.[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. It is clear that $\hat{\tau}_{tg}^{CS}$ can be viewed as an estimator of $ATE(t,g)$ of the form given in (2), 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 (2) 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 $t$,

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

[9] In particular, the unconditional mean $\bar{X} = \frac{N_2}{N}\bar{X}_1 + \frac{N_\infty}{N}\bar{X}_0$. The result then follows from re-arranging terms in $\tau(\beta_0, \beta_1)$.

[10] This could also be viewed as an estimator of the form (2) if $\hat{X}$ were a vector with each element corresponding with $\hat{\tau}_{t,g\infty}$ and the vector $\beta$ was a vector with elements corresponding with $w_{t,g\infty}$.

It is again apparent that this estimator can be written as an estimator of $ATE(t,g)$ of the form in (2), with $\hat{X}$ now corresponding with a weighted average of $\hat{\tau}_{g-1,gg'}$ and $\beta$ 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

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

where $g_{max} = \max \mathcal{G}$ is the last period in which units receive treatment and $s < g$ is some reference period before $g$ (e.g. $g-1$). It is clear that $\hat{\tau}^{SA}_{tg}$ takes the form (2), with $\hat{X} = \hat{\tau}_{s,gg_{max}}$ and $\beta = 1$. Weighted averages of the $\hat{\tau}^{SA}_{tg}$ can likewise be expressed in the form (2), 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 $\hat{\tau}^{CS2}_{tg}$. 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,

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

It is thus immediate from the previous examples that their estimator can also be written in the form (2).

**Example 5** (TWFE Models)**.** Athey and Imbens (2018) consider the setting with $\mathcal{G} = \{1, ...T, \infty\}$. Let $D_{it} = 1[G_i \leqslant t]$ be an indicator for whether unit $i$ is treated by period $t$. Athey and Imbens (2018, Lemma 5) show that the coefficient on $D_{it}$ from the two-way fixed effects specification

$$Y_{it} = \alpha_i + \lambda_t + D_{it}\theta^{TWFE} + \epsilon_{it} \tag{3}$$

can be decomposed as

$$\hat{\theta}^{TWFE} = \sum_t \sum_{\substack{(g,g'):\\ min(g,g')\leqslant t}} \gamma_{t,gg'}\hat{\tau}_{t,gg'} + \sum_t \sum_{\substack{(g,g'):\\ min(g,g')>t}} \gamma_{t,gg'}\hat{\tau}_{t,gg'} \tag{4}$$

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 (2) 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 (2018) and other papers, however, the parameter $\theta^{TWFE}$ may not have an intuitive causal interpretation under treatment effect heterogeneity, since the weights $\gamma_{t,gg'}$ may be negative.

**Remark 3** (Covariate adjustment for multi-armed trials)**.** In a cross-sectional random experiment with multiple arms $g$ and a fixed covariate $X_i$, the natural extension of Lin (2013)'s approach for binary treatments is to estimate $\mathbb{E}_f[Y_i(g)]$ with $\bar{Y}_g - \beta_g'(\bar{X}_g - \mathbb{E}_f[X_i])$, where $\bar{Y}_g$ and $\bar{X}_g$ are the sample means of $Y_i$ and $X_i$ among units with $G_i = g$, and to then form contrasts by differencing the estimates for $\mathbb{E}_f[Y_i(g)]$. Our staggered setting is similar to this set-up with $X_i$ corresponding with outcomes before treatment begins. However, a key difference is that a different number of pre-treatment outcomes are observed for units treated at different times. For example, for units with $G_i = 1$, we do not observe any pre-treatment outcomes, whereas for units with $G_i = 4$, we observe $Y_{i1}(\infty), ..., Y_{i4}(\infty)$. It is thus not possible to directly apply this approach, since $X_i$ is not observed for all units and thus we cannot calculate $\mathbb{E}_f[X_i]$. However, the estimator of the form in (2) is based on a similar principle, since by construction, $\mathbb{E}\left[\hat{X}\right] = 0$, and likewise $\mathbb{E}\left[\bar{X}_g - \mathbb{E}_f[X_i]\right] = 0$ in the cross-sectional case. In fact, in the special case of our framework where all treated units begin treatment at the same time ($\mathcal{G} = \{T_0, \infty\}$), the covariate-adjustment estimator with $X_i$ a vector of pre-treatment outcomes can be represented in the form (2) for an appropriately defined $\hat{X}$.[11]

[11] This follows from the fact that $\bar{X}_g - \mathbb{E}_f[X_i]$ can be written as a linear combination of $\bar{X}_g - \bar{X}_{g'}$.

## 2.4 Efficient "Oracle" Estimation

We now consider the problem of finding the best estimator $\hat{\theta}_\beta$ of the form introduced in (2). We first show that $\hat{\theta}_\beta$ is unbiased for all $\beta$, and then solve for the $\beta^*$ that minimizes the variance.

We begin by introducing some notation that will be useful for presenting our results.

**Notation.** 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 $A_0$ 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 $\beta$.

**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 $\beta^*$ that minimizes the variance.

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

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

*provided that* $\mathbb{V}ar\left[\hat{X}\right]$ *is positive definite. Further, the variances and covariances in the expression for* $\beta^*$ *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* $\mathbb{V}ar\left[\hat{\theta}_{\beta*}\right] = V_{\hat{\theta}_0} - (\beta^*)'V_{\hat{X}}^{-1}(\beta^*)$.

**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 $\beta^*$ is equal to $\frac{N_\infty}{N}\beta_2 + \frac{N_2}{N}\beta_\infty$, where $\beta_g$ 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., $\beta^*$ 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 $\beta^*$ that minimizes the variance of $\hat{\theta}_\beta$. However, the efficient estimator $\hat{\theta}_{\beta *}$ is not of practical use since the "oracle" coefficient $\beta^*$ depends on the covariances of the potential outcomes, $S_g$, which are typically not known in practice. Mirroring Lin (2013) in the cross-sectional case, we now show that $\beta^*$ 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,

$$Y_{i2} = \beta_0 + \beta_1 D_i + \beta_2 \dot{Y}_{i1} + \beta_3 D_i \times \dot{Y}_{i1} + \epsilon_i, \tag{5}$$

where $\dot{Y}_{i1}$ is the demeaned value of $Y_{i1}$.[12]

**Remark 4** (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 $\gamma_1$ from the OLS regression

$$Y_{i2} = \gamma_0 + \gamma_1 D_i + \gamma_2 \dot{Y}_{i1} + \epsilon_i, \tag{6}$$

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 (5), 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 (6) is equivalent to the estimator studied in Freedman (2008b,a). 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 (5) converges in probability to 0. However, the results in Freedman (2008b,a) 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 $\hat{\theta}_{\hat{\beta}*}$. Relatedly, Yang and Tsiatis (2001), Funatogawa et al. (2011), and Wan (2020) show that $\hat{\beta}_1$ from (5) is asymptotically at least

[12] 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.

as efficient as $\hat{\gamma}_1$ from (6) in sampling-based models similar to our ongoing example.

### 2.5.2 Asymptotic properties of the plug-in estimator

We will now show that in large populations, $\hat{\theta}_{\hat{\beta}*}$ is asymptotically unbiased for $\theta$ and has the same asymptotic variance as the oracle estimator $\hat{\theta}_{\beta*}$. As in Lin (2013) and Li and Ding (2017) among other papers, 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 $g \in \mathcal{G}$, $N_g/N \to p_g \in (0,1)$.*

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

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,*

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

## 2.6 Covariance Estimation

To construct confidence intervals using Proposition 2.2, one requires an estimate of $\sigma_*^2$. 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 $\sigma_*^2 = \lim_{N\to\infty} N\mathbb{V}\text{ar}\left[\hat{\theta}_{\beta*}\right]$. Examining the expression for $\mathbb{V}\text{ar}\left[\hat{\theta}_{\beta*}\right]$ 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 $\sigma_*^2$, 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 $\theta$ 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); Abadie et al. (2020), with lagged outcomes playing a similar role to the fixed covariates.[13] To avoid technical clutter, we define the refined estimator here

[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.

and provide a more detailed derivation in Appendix A.1.

**Lemma 2.3.** *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*

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

*where $\hat{\beta}_g = (M\hat{S}_g M')^{-1} M \hat{S}_g A'_{\theta,g}$. Then $\hat{\sigma}^2_{**} \rightarrow_p \sigma^2_* + S^*_{\tilde{\theta}}$, 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 $S^*_{\tilde{\theta}}$.)*

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

**Remark 5** (Fisher Randomization Tests)**.** An alternative approach to inference would be to consider Fisher Randomization Tests (FRTs) based on the studentized statistic $\hat{\beta}^*/\hat{\sigma}_{**}$ (Wu and Ding, 2020; Zhao and Ding, 2020). By arguments analogous to those in Zhao and Ding (2020), the FRT based on the studentized statistic will be finite-sample exact under the sharp null hypothesis, and asymptotically equivalent to the test that $0 \in CI_{**}$ under the Neyman null that $\theta = 0$.

## 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 $\beta^*$ 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 $\beta^*$ 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.[14] We thus view the results on the efficient estimator as complementary to these estimators considered in previous work, since it is more efficient under stricter assumptions that will not hold in all cases of interest.

Similarly, in light of Example 5, our results imply that the TWFE estimator will generally not be the most efficient estimator for the TWFE estimand, $\theta^{TWFE}$. Previous work has argued that the estimand $\theta^{TWFE}$ may be difficult to interpret (e.g. Athey and Imbens (2018); Borusyak and Jaravel (2017); Goodman-Bacon (2018); de Chaisemartin and D'Haultfœuille (2020)). Our results provide a new and complementary critique of the TWFE specification: even if $\theta^{TWFE}$ is the target parameter, estimation via (3) will generally be inefficient in large populations under random treatment timing and no anticipation.

# 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 the data in Wood et al. (2020a) which we use in our application.

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

### 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 $Y_i(\infty) = (Y_{i1}(\infty), Y_{i2}(\infty))'$ from a $\mathcal{N}(0, \Sigma_\rho)$ distribution, where $\Sigma_\rho$ has 1s on the diagonal and $\rho$ on the off-diagonal. The parameter $\rho$ 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 $\gamma$ governs the degree of heterogeneity of treatment effects: if $\gamma = 0$, then there is no treatment effect heterogeneity, whereas if $\gamma$ 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 based on the plug-in efficient estimator $\hat{\theta}_{\hat{\beta}*}$, difference-in-differences $\hat{\theta}^{DiD} = \hat{\theta}_1$, and simple differences-in-means $\hat{\theta}^{DiM} = \hat{\theta}_0$. 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.[15] For all specifications and estimators, the estimated bias is small, and coverage is close to the nominal level. Table 2 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 ($\rho = 0.99$)

[15] 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 $\mathbb{V}\mathrm{ar}\left[\hat{\theta}_{\beta*}\right]$.

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 optimal in the class we consider. The plug-in efficient estimator is substantially more precise than DiD and DiM in many other specifications: in the worst specification, the standard deviation of DiD is 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 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N_\infty$ | $N_2$ | $\rho$ | $\gamma$ | 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 |
| 1000 | 1000 | 0.99 | 0.5 | 0.00 | 0.00 | −0.00 | 0.01 | 0.01 | 0.06 | 0.95 | 0.95 | 0.95 |
| 1000 | 1000 | 0.50 | 0.0 | 0.00 | 0.00 | 0.00 | 0.04 | 0.04 | 0.05 | 0.94 | 0.95 | 0.94 |
| 1000 | 1000 | 0.50 | 0.5 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.06 | 0.95 | 0.95 | 0.95 |
| 1000 | 1000 | 0.00 | 0.0 | −0.00 | 0.00 | −0.00 | 0.04 | 0.07 | 0.04 | 0.95 | 0.94 | 0.95 |
| 1000 | 1000 | 0.00 | 0.5 | −0.00 | 0.00 | −0.00 | 0.06 | 0.07 | 0.06 | 0.95 | 0.95 | 0.95 |
| 25 | 25 | 0.99 | 0.0 | 0.00 | 0.00 | −0.03 | 0.04 | 0.04 | 0.27 | 0.94 | 0.94 | 0.94 |
| 25 | 25 | 0.99 | 0.5 | 0.00 | −0.01 | −0.04 | 0.05 | 0.08 | 0.34 | 0.92 | 0.93 | 0.93 |
| 25 | 25 | 0.50 | 0.0 | −0.01 | 0.02 | −0.02 | 0.24 | 0.29 | 0.26 | 0.94 | 0.95 | 0.94 |
| 25 | 25 | 0.50 | 0.5 | −0.01 | 0.01 | −0.03 | 0.30 | 0.32 | 0.33 | 0.94 | 0.95 | 0.94 |
| 25 | 25 | 0.00 | 0.0 | −0.03 | −0.02 | −0.03 | 0.28 | 0.38 | 0.27 | 0.93 | 0.95 | 0.93 |
| 25 | 25 | 0.00 | 0.5 | −0.04 | −0.02 | −0.04 | 0.35 | 0.42 | 0.34 | 0.93 | 0.94 | 0.94 |

Table 1: Bias, Standard Deviation, and Coverage for $\hat{\theta}_{\hat{\beta}^*}$, $\hat{\theta}^{DiD}$, $\hat{\theta}^{DiM}$ 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. (2020a) in Section 4. The outcome of interest $Y_{it}$ is the number of complaints against police officer $i$ in month $t$ for

|  |  |  |  |  | SD Relative to Plug-In |  |  |
|---|---|---|---|---|---|---|---|
| $N_\infty$ | $N_2$ | $\rho$ | $\gamma$ | $\beta^*$ | 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 2: Ratio of standard deviations for $\hat{\theta}^{DiD}$ and $\hat{\theta}^{DiM}$ relative to $\hat{\theta}_{\hat{\beta}*}$ in 2-period simulations

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 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 ($\theta^{simple}$); the calendar- and cohort-weighted average treatment effects ($\theta^{calendar}$ and $\theta^{cohort}$), and the instantaneous event-study parameter ($\theta_0^{ES}$). (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 estimator for the same estimand, using the not-yet-treated as the control group (since all units are eventually treated). Recall that for $\theta_0^{ES}$, the CS estimator coincides with the estimator proposed in de Chaisemartin and D'Haultfœuille (2020) in our setting, since treatment is an absorbing state. We also compare to the Sun and Abraham (2020, SA) estimator that uses the last-to-be-treated units as the control group. 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.[16]

**Baseline simulation results.** The results for our baseline specification are shown in Tables 3 and 4. As seen in Table 3, 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 CS and SA estimators are also both approximately unbiased and have good coverage for all of the estimands as well.

Table 4 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 estimator. 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 the other estimands. Since standard errors are proportional to the square root of

[16] The variance estimator for the CS and SA estimators is adapted analogously to that for the DiD and DiM estimators, as discussed in footnote 15.

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. 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.

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

Table 3: Results for Simulations Calibrated to Wood et al. (2020a)

Note: This table shows results for the plug-in efficient and Callaway and Sant'Anna (2020) and Sun and Abraham (2020) estimators in simulations calibrated to Wood et al. (2020a). The estimands considered are the calendar-, cohort-, and simple-weighted average treatment effects, as well as the instantaneous event-study effect (ES0). The Callaway and Sant'Anna (2020) 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. 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.

**Extensions.** In Appendix B, we present simulations from an alternative specification where the monthly data is collapsed to the yearly level, leading to fewer time periods and fewer (but larger) cohorts. All three estimators again have good coverage and minimal bias. The plug-in efficient estimator again dominates the other estimators in efficiency, although the gains are smaller (24 to 30% reductions in standard deviation relative to CS). The smaller

|  | 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 4: 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 Callaway and Sant'Anna (2020) and Sun and Abraham (2020) estimators relative to the plug-in efficient estimator, based on the simulation results in Table 3.

efficiency gains in this specification are intuitive: the CS and SA estimators overweight 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.

In the appendix, we also present results from a modification of our baseline DGP with heterogeneous treatment effects. We again find that the plug-in efficient estimator performs well, with qualititative findings similar to those in the baseline specification, although the standard errors are somewhat conservative as expected.

In the appendix, we also conduct simulation results using a modified version of the plug-in efficient estimator in which $\hat{X}$ is a vector containing all possible comparisons of cohorts $g$ and $g'$ in periods $t < min(g, g')$. We find poor coverage of 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. 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 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.

Finally, we repeat the same exercise for the other outcomes used in our application (use of force and sustained complaints). We again find that the plug-in efficient estimator has minimal bias, good coverage properties, 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). 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 CIs based on the plug-in efficient estimator is only 79% in this specification. 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 plug-in efficient estimator again has good coverage (94%) when considering the annualized data where the cohort sizes are larger. We thus urge some caution in using the plug-in efficient estimator (or any procedure based on a normal approximation) when cohort sizes are small ($<$30) and the outcome is rare (mean $< 0.01$); 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.

# 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.[17] 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, Wood et al. (2020b) discovered 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. Wood et al. (2020b) 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.

## 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 the 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. The data also contain the date on which each officer was trained.

## 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 treat-

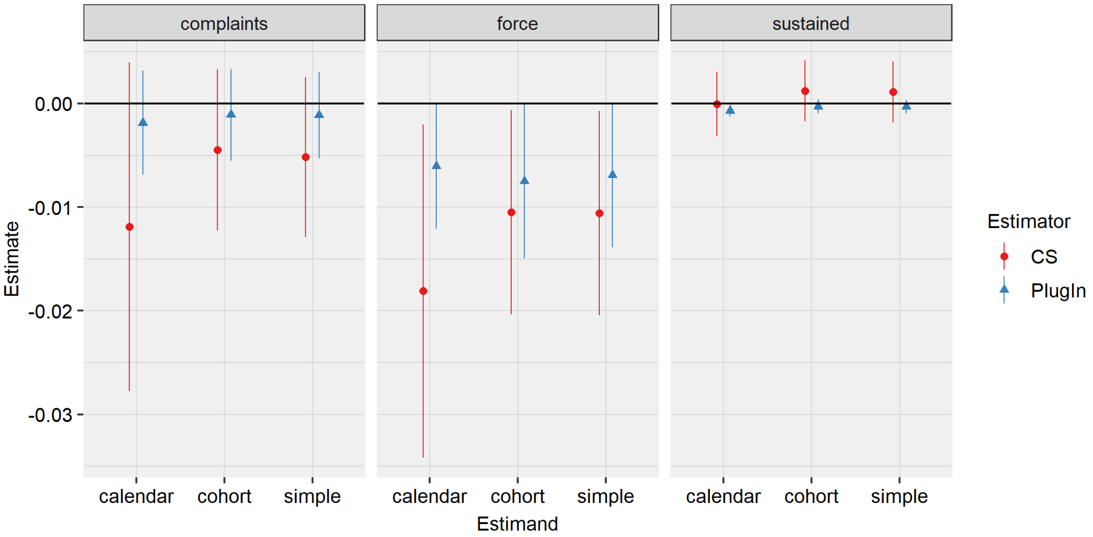
[17] See the Supplement to Wood et al. (2020a) for discussion of some concerns regarding non-compliance, particularly towards the end of the sample. We explore robustness to dropping officers trained in the last year in Appendix Figure 4. The results are qualitatively similar, although with smaller estimated effects on use of force.

ment, 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). (Recall that for the instantaneous event-study effect, the Callaway and Sant'Anna (2020) and de Chaisemartin and D'Haultfœuille (2020) estimators coincide.)

## 4.4 Results

Figure 2 shows the results of our analysis for the three aggregate summary parameters. Table 5 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. The estimates using the plug-in efficient estimator are substantially more precise than those using the Callaway and Sant'Anna (2020, CS) estimator, with the standard errors ranging from 1.3 to 5.6 times smaller (see final column of Table 5).

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.

As in Wood et al. (2020b), we find no significant impact on complaints using any of

|  |  |  | Plug-In |  |  | CS |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Outcome | Estimand | Pre-treat Mean | Estimate | LB | UB | Estimate | LB | UB | CI Ratio |
| complaints | simple | 0.049 | −2% | −11% | 6% | −10% | −26% | 5% | 1.9 |
| complaints | calendar | 0.049 | −4% | −14% | 6% | −24% | −56% | 8% | 3.2 |
| complaints | cohort | 0.049 | −2% | −11% | 7% | −9% | −25% | 7% | 1.8 |
| sustained | simple | 0.004 | −7% | −23% | 8% | 27% | −44% | 97% | 4.5 |
| sustained | calendar | 0.004 | −17% | −30% | −3% | −1% | −75% | 73% | 5.6 |
| sustained | cohort | 0.004 | −7% | −22% | 9% | 29% | −41% | 99% | 4.4 |
| force | simple | 0.048 | −15% | −29% | 0% | −22% | −43% | −2% | 1.4 |
| force | calendar | 0.048 | −13% | −26% | 0% | −38% | −72% | −4% | 2.6 |
| force | cohort | 0.048 | −16% | −31% | −0% | −22% | −43% | −1% | 1.3 |

Table 5: 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. The final columns shows the ratio of the CI length using the CS estimator relative to the plug-in efficient estimator.

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 than when using the CS estimator and the upper bounds of the confidence intervals are all nearly exactly 0. 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.[18] 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.

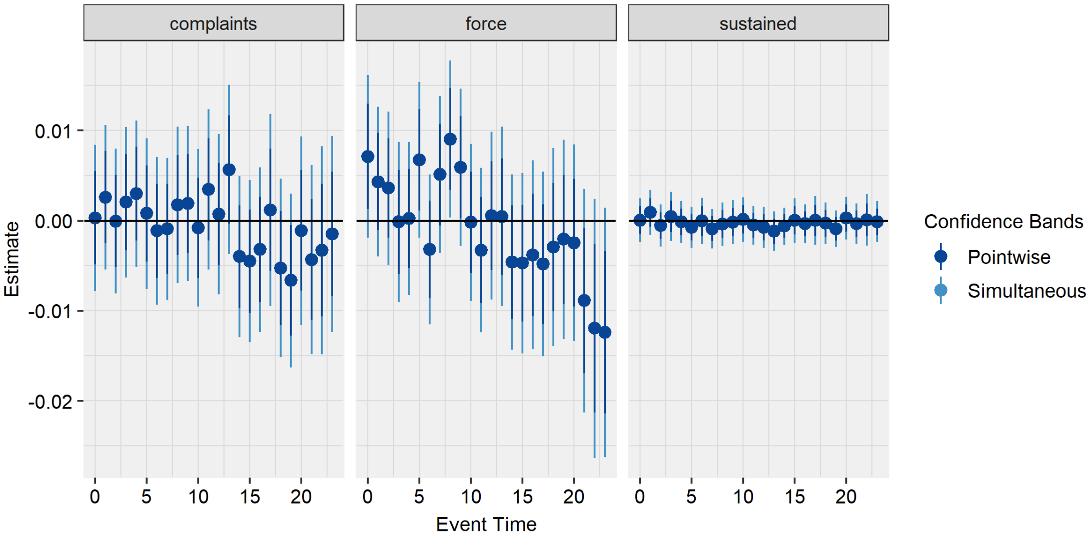
Figure 2 shows event-time estimates for the first two years using the plug-in efficient estimator. (To conserve space, we place the analogous results for the CS estimator in the appendix.) In dark blue, we present point estimates and pointwise confidence intervals, and in light blue we present simultaneous confidence bands calculated using sup-t confidence bands (Olea and Plagborg-Møller, 2019).[19] 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). The figure 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).

Finally, in Appendix Figure 4, 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 smaller and not statistically significant at conventional levels.

[18] Recall that the calendar aggregation for sustained complaints was the one specification for which CIs based on the plug-in efficient estimator substantially undercovered (79%), and thus the significant result should be interpreted with some caution.

[19] 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 a Neymanian randomization framework of random treatment timing. The assumption of random treatment timing is stronger than the typical parallel trends assumption, but can be ensured by design when the researcher controls the timing of treatment, and is often the justification given for parallel trends in quasi-experimental contexts. We then derive the most efficient estimator in a large class of estimators that nests many existing approaches. Although the "oracle" efficient estimator is not known in practice, we show that a plug-in sample analog has similar properties in large populations, and derive a valid variance estimator for construction of confidence intervals. We find in simulations that the proposed plug-in efficient estimator is approximately unbiased, yields CIs with good coverage, 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.

**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," *arXiv:1808.05293 [cs, econ, math, stat]*, August 2018.

**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.

**Breidt, F. Jay and Jean D. Opsomer**, "Model-Assisted Survey Estimation with Modern Prediction Techniques," *Statistical Science*, 2017, *32* (2), 190–205. Publisher: Institute of Mathematical Statistics.

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

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

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

**Ding, Peng and Fan Li**, "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment," *Political Analysis*, 2019, *27* (4), 605–615.

**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 pre-treatment measurements in randomized trials under the cases that covariances and post-treatment variances differ between groups," *Biometrical Journal*, May 2011, *53* (3), 512–524.

**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.

**Hussey, Michael A. and James P. Hughes**, "Design and analysis of stepped wedge cluster randomized trials," *Contemporary Clinical Trials*, 2007, *28* (2), 182–191.

**Imai, Kosuke and In Song Kim**, "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 2020, (Forthcoming).

**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.

**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).

**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.

**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.

**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.

**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.

# A Proofs

### Proof of Lemma 2.1

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

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

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. □

### Proof of Proposition 2.1

*Proof.* First, observe that

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

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

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

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 $\mathbb{V}\text{ar}\left[\hat{\tau}\right] = \sum_g {N_g}^{-1}A_{\tau,g}S_gA'_{\tau,g} - N^{-1}S_\tau$, 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. □

**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,*

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

*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 $S_\theta$ 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 $S_\tau \to \begin{pmatrix} S_\theta^* & 0 \\ 0 & 0 \end{pmatrix}$. □

**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). □

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. □

## A.1 Derivation of Variance Refinement

We now provide a derivation for the refined variance estimator introduced in Lemma 2.3, as well as a formal proof of the lemma. 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 $S_\theta$. 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*

$$S_\theta = \mathbb{V}ar_f\left[\tilde{\theta}_i\right] + \frac{N+1}{N-1}\left(\sum_{g\geqslant g_{min}} \beta_g\right)'(MS_{g_{min}}M')\left(\sum_{g\geqslant g_{min}} \beta_g\right), \tag{7}$$

*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 $\beta_{XZ} = \mathbb{V}\text{ar}_f[X_i]^{-1}\mathbb{E}_f\left[\dot{X}_i\dot{Z}_i\right]$. Observe that

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

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 $\mathbb{V}\text{ar}_f[MY_i(g_{min})] = MS_{g_{min}}M' = MS_gM' = \mathbb{V}\text{ar}_f[MY_i(g)]$. □

The proof of Lemma 2.3 then follows nearly immediately from A.3.

### Proof of Lemma 2.3

*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 $S^*_{\tilde{\theta}} = \lim_{N\to\infty} \mathbb{V}\mathrm{ar}_f \left[\tilde{\theta}_i\right] \leqslant \lim_{N\to\infty} S_\theta = S^*_\theta$. □

# B Additional Simulation Results

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

**Other outcomes.** Tables 6-9 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).

**Annualized data.** Tables 10-15 show versions of our simulations results (for all three outcomes) when the data is collapsed to the annual level, so that there are 6 total time periods and 5 cohorts.

**Augmented $\hat{X}$.** Table 16 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 $t < min(g, g')$. This vector is large relative to sample size in the monthly specification ($dim(\hat{X}) = 1987$, $N = 7785$), which leads to bias and severe undercoverage for the modified plug-in efficient estimator. In the annualized data, the dimension of the modified $\hat{X}$ is modest (10), and the modified efficient estimator has good coverage and yields small efficiency gains (up to 3%) relative to the plug-in efficient estimator considered in the main text.

**Heterogeneous Treatment Effects.** Tables 17 and 18 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 half 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 results are similar to those for the main specification, although as expected, the standard errors are somewhat conservative (i.e. the mean standard error exceeds the standard deviation of the estimator).

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

Table 6: Results for Simulations Calibrated to Wood et al. (2020a) – Use of Force

Note: This table shows results analogous to Table 3, 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 |

Table 7: 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 4, except using Use of Force rather than Complaints as the outcome.

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

Table 8: Results for Simulations Calibrated to Wood et al. (2020a) – Sustained Complaints

Note: This table shows results analogous to Table 3, 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 |

Table 9: 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 4, except using Sustained Complaints rather than Complaints as the outcome.

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

Table 10: Results for Simulations Calibrated to Wood et al. (2020a) – Annualized Data

Note: This table shows results analogous to Table 3, 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 |

Table 11: 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 4, except the data is collapsed to the annual level.

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

Table 12: Results for Simulations Calibrated to Wood et al. (2020a) – 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.

| | 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 |

Table 13: 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 4, 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 | Mean SE | SD |
|---|---|---|---|---|---|
| PlugIn | calendar | 0.00 | 0.95 | 0.43 | 0.44 |
| PlugIn | cohort | -0.01 | 0.94 | 0.53 | 0.55 |
| PlugIn | ES0 | 0.01 | 0.95 | 0.45 | 0.45 |
| PlugIn | simple | -0.01 | 0.94 | 0.51 | 0.52 |
| CS | calendar | 0.02 | 0.96 | 0.69 | 0.66 |
| CS | cohort | 0.03 | 0.96 | 0.81 | 0.78 |
| CS/CdH | ES0 | 0.01 | 0.96 | 0.61 | 0.60 |
| CS | simple | 0.02 | 0.96 | 0.80 | 0.77 |
| SA | calendar | 0.01 | 0.95 | 1.08 | 1.06 |
| SA | cohort | 0.02 | 0.96 | 1.11 | 1.08 |
| SA | ES0 | 0.00 | 0.95 | 0.96 | 0.99 |
| SA | simple | 0.02 | 0.96 | 1.18 | 1.16 |

Table 14: Results for Simulations Calibrated to Wood et al. (2020a) – 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.

| | 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 |

Table 15: 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 4, 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 | Mean SE | SD |
|---|---|---|---|---|---|
| PlugIn - Long X | calendar | 1.96 | 0.00 | 0.13 | 0.30 |
| PlugIn - Long X | cohort | 1.69 | 0.01 | 0.04 | 0.26 |
| PlugIn - Long X | ES0 | 1.93 | 0.01 | 0.21 | 0.38 |
| PlugIn - Long X | simple | 1.78 | 0.00 | 0.04 | 0.23 |
| PlugIn | calendar | 0.00 | 0.93 | 0.27 | 0.29 |
| PlugIn | cohort | 0.00 | 0.92 | 0.24 | 0.24 |
| PlugIn | ES0 | 0.01 | 0.94 | 0.26 | 0.27 |
| PlugIn | simple | 0.00 | 0.92 | 0.22 | 0.22 |

(b) Annual Data

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

Table 16: 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 | Mean SE | SD |
|---|---|---|---|---|---|
| PlugIn | calendar | 0.00 | 0.93 | 0.27 | 0.29 |
| PlugIn | cohort | 0.00 | 0.92 | 0.24 | 0.24 |
| PlugIn | ES0 | 0.01 | 0.94 | 0.26 | 0.27 |
| PlugIn | simple | 0.00 | 0.92 | 0.22 | 0.22 |
| CS | calendar | 0.00 | 0.94 | 0.55 | 0.55 |
| CS | cohort | -0.01 | 0.95 | 0.41 | 0.41 |
| CS/CdH | ES0 | 0.01 | 0.94 | 0.36 | 0.36 |
| CS | simple | -0.01 | 0.96 | 0.41 | 0.40 |
| SA | calendar | 0.06 | 0.93 | 1.30 | 1.30 |
| SA | cohort | 0.05 | 0.92 | 1.34 | 1.38 |
| SA | ES0 | 0.03 | 0.94 | 0.83 | 0.89 |
| SA | simple | 0.06 | 0.92 | 1.46 | 1.49 |

Table 17: Results for Simulations Calibrated to Wood et al. (2020a) – Heterogeneous Treatment Effects

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

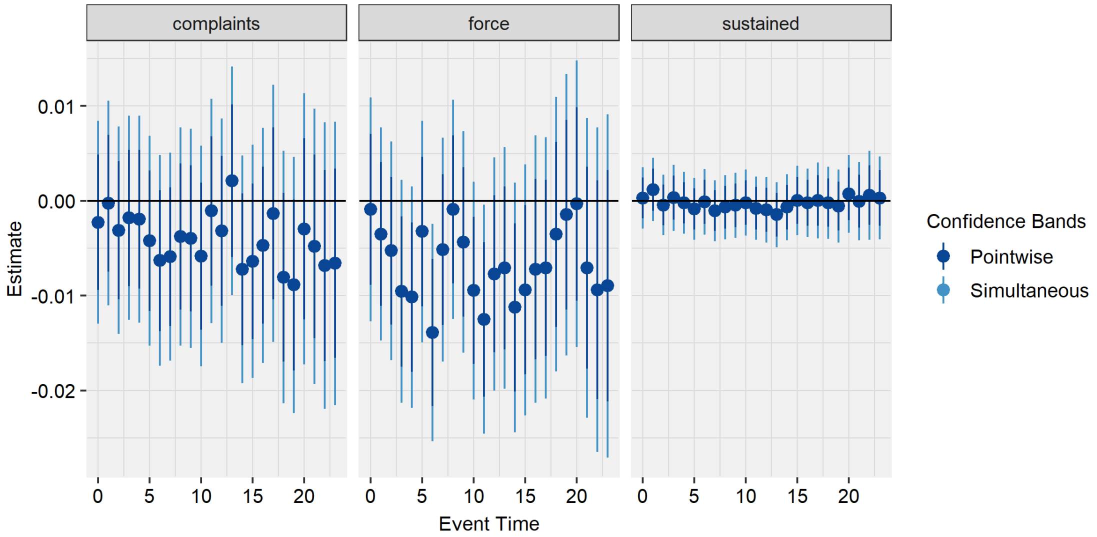
|  | Ratio of SD to Plug-In |  |
|---|---|---|
| Estimand | CS | SA |
| calendar | 1.88 | 4.72 |
| cohort | 1.51 | 5.25 |
| ES0 | 1.34 | 3.23 |
| simple | 1.65 | 6.26 |

Table 18: 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 4, except the DGP adds heterogeneous treatment effect as described in Section B.

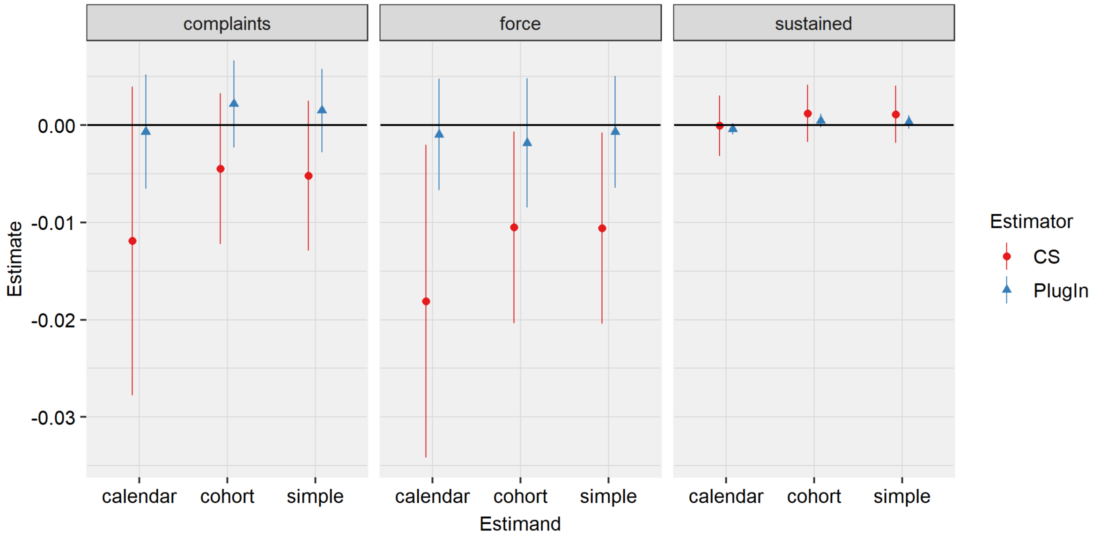
# C Additional Tables and Figures

Figure 3: Event-Time Average Effects Using the CS Estimator



Note: This figure is analogous to Figure 2 except it uses the CS estimator rather than the plug-in efficient.

Figure 4: 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.