Received: 21 December 2019 | Accepted: 19 May 2021

DOI: 10.1111/rssb.12448

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

# Synthetic controls with staggered adoption

**Eli Ben-Michael**[1] | **Avi Feller**[2] | **Jesse Rothstein**[2]

[1]Harvard University, Cambridge, Massachusetts, USA

[2]University of California, Berkeley, California, USA

**Correspondence**
Avi Feller, University of California, Berkeley, California, USA.
Email: afeller@berkeley.edu

**Funding information**
U.S. Department of Education, Grant/Award Number: R305D200010

**Abstract**

Staggered adoption of policies by different units at different times creates promising opportunities for observational causal inference. Estimation remains challenging, however, and common regression methods can give misleading results. A promising alternative is the synthetic control method (SCM), which finds a weighted average of control units that closely balances the treated unit's pre-treatment outcomes. In this paper, we generalize SCM, originally designed to study a single treated unit, to the staggered adoption setting. We first bound the error for the average effect and show that it depends on both the imbalance for each treated unit separately and the imbalance for the average of the treated units. We then propose 'partially pooled' SCM weights to minimize a weighted combination of these measures; approaches that focus only on balancing one of the two components can lead to bias. We extend this approach to incorporate unit-level intercept shifts and auxiliary covariates. We assess the performance of the proposed method via extensive simulations and apply our results to the question of whether teacher collective bargaining leads to higher school spending, finding minimal impacts. We implement the proposed method in the `augsynth` R package.

**KEYWORDS**

causal inference, panel data, synthetic control method

# 1 | INTRODUCTION

Jurisdictions often adopt policies at different times, creating promising opportunities for observational causal inference. In our motivating application, 33 states passed laws between 1964 and 1987 mandating that school districts bargain with teachers' unions (Hoxby, 1996; Paglayan, 2019); our goal is to estimate the impact of these laws on teacher salaries and school expenditures.

However, estimating causal effects under staggered adoption remains challenging. Workhorse methods, such as the regression-based two-way fixed effects model, rely on strong modelling assumptions and can give misleading estimates when treatment timing varies (Borusyak et al., 2021; Goodman-Bacon, 2021; Sun & Abraham, 2020). A promising alternative is the *synthetic control method* (SCM; Abadie et al., 2010, 2015). SCM estimates the counterfactual untreated outcome via a weighted average of untreated units, with weights chosen to match the treated unit's pre-treatment outcomes as closely as possible. SCM, however, was developed for settings where only a single unit is treated, and proposals for extending SCM to the staggered adoption case have been ad hoc. One common strategy is to estimate SCM weights separately for each treated unit and then average the estimates (see, e.g., Donohue et al., 2019; Dube & Zipperer, 2015). However, this relies on being able to find good synthetic controls for every treated unit, which is not possible in our application.

In this paper, we develop SCM for the staggered adoption setting. Under two common data generating processes for panel data, an autoregressive model and a linear factor model, we bound the error of a weighting estimator for the average effect and show that it depends on both the unit-specific imbalance for each treated unit and the imbalance for the average of the treated units. This leads to our main proposal, *partially pooled SCM*, which minimizes a weighted average of the two imbalances. This approach nests two special cases: *separate SCM*, which reflects the current practice of estimating weights that separately minimize the pre-treatment imbalance for each treated unit; and *pooled SCM*, which instead minimizes the average pre-treatment imbalance across all treated units. Both special cases have drawbacks. Separate SCM can lead to poor fit for the average, leading to possible bias when the average treatment effect is the estimand of interest. Pooled SCM, by contrast, can achieve nearly perfect fit for the average treated unit but can yield substantially worse unit-specific fits. This can lead to poor estimates of unit-level treatment effects and to bias for the average effect if the data generating process varies over time. Partially pooled SCM moves smoothly between these two extremes, with a hyperparameter denoting the relative weight of the two balance measures in the optimization problem. We discuss how to select weights to trade-off between these two quantities in practice.

We then explore several extensions. First, we incorporate an intercept shift into the SCM problem, following proposals by Doudchenko and Imbens (2017) and Ferman and Pinto (2021). The resulting treatment effect estimator has the form of a weighted difference-in-differences estimator, connecting our proposed approach to a large econometric literature (Callaway & Sant'Anna, 2020; Sun & Abraham, 2020). We recommend this approach as a reasonable default in practice; it amounts to applying our partially pooled SCM estimator to de-meaned outcome series. Second, we modify the SCM problem to incorporate auxiliary covariates alongside lagged outcomes. We also briefly address inference for SCM-like estimates in the staggered adoption setting. We implement the proposed methodology in the `augsynth` package for `R`, available at https://github.com/ebenmichael/augsynth.

We apply our methods to estimating the impact of mandatory teacher collective bargaining and show that they achieve better pre-treatment balance than existing approaches. We find no impact of teacher collective bargaining laws on either teacher salaries or student expenditures,

consistent with several recent papers (Frandsen, 2016; Paglayan, 2019) but counter to earlier claims (most notably Hoxby, 1996).

## 1.1 | Related work

Our paper contributes to several methodological literatures. First, there is a large and active applied econometrics literature on challenges and remedies for two-way fixed effects models with multiple treated units; see Borusyak et al. (2021); Sun and Abraham (2020); Athey and Imbens (2021); Goodman-Bacon (2021); Callaway and Sant'Anna (2020); Roth and Sant'Anna (2021). See also Xu (2017) and Athey et al. (2021) for recent generalizations of these models.

SCM has also attracted a great deal of attention; see Abadie (2019) for a review. Several recent papers have explored SCM with multiple treated units. In the case where all units adopt treatment at the same time, some propose to first average the units and then estimate SCM weights for the average, analogous to our fully pooled SCM estimate; for discussion, see Kreif et al. (2016); Robbins et al. (2017). An alternative is Abadie and L'Hour (2021), who instead propose to estimate separate SCM weights for each treated unit. In particular, they propose a penalized SCM approach that aims to reduce interpolation bias, allowing for weights that move continuously between standard SCM and nearest-neighbour matching. Our approach complements these papers by adapting some of these ideas to the staggered adoption setting. For some other examples of SCM under staggered adoption, see also Dube and Zipperer (2015); Shaikh and Toulis (2021); Donohue et al. (2019); Cao and Lu (2019).

## 1.2 | Motivating example: Teacher collective bargaining

The United States, like other developed countries, spends substantial resources on public education. Approximately 80% of education spending goes to teacher salaries and benefits (U.S. Department of Education, National Center for Education Statistics, 2018), and research points to teacher quality as a key determinant of student outcomes (Jackson et al., 2014). Over recent decades, the teacher employment relationship has changed dramatically via the introduction of unions and collective bargaining agreements (Goldstein, 2015). Critics identify these as a 'harmful anachronism' and 'the most daunting impediments' to education reform (Hess & West, 2006), while proponents argue that collective bargaining raises pay and thereby helps to attract and retain high-quality teachers. A major 2018 Supreme Court decision, *Janus v AFSCME*, is expected to weaken teachers' unions, bringing renewed attention to this area and raising interest in understanding the effects of teacher collective bargaining.

Since 1964, a number of states have passed laws mandating that school districts bargain with teachers' unions.[1] Given the strong criticism directed at teachers' unions, there is surprisingly little evidence that they, or the mandatory bargaining laws, have any effect at all. In a seminal study, Hoxby (1996) uses state-level changes in collective bargaining laws to argue that teacher collective bargaining raises teacher salaries and school expenditures but reduces student outcomes. However, several more recent papers have disputed Hoxby's conclusions. Using a panel of school districts, Lovenheim (2009) finds little effect of unionization on teacher pay or class size. Frandsen (2016)

[1]Another 10 states allow but do not require collective bargaining, while seven prohibit it. We focus on estimating the effects of mandates.

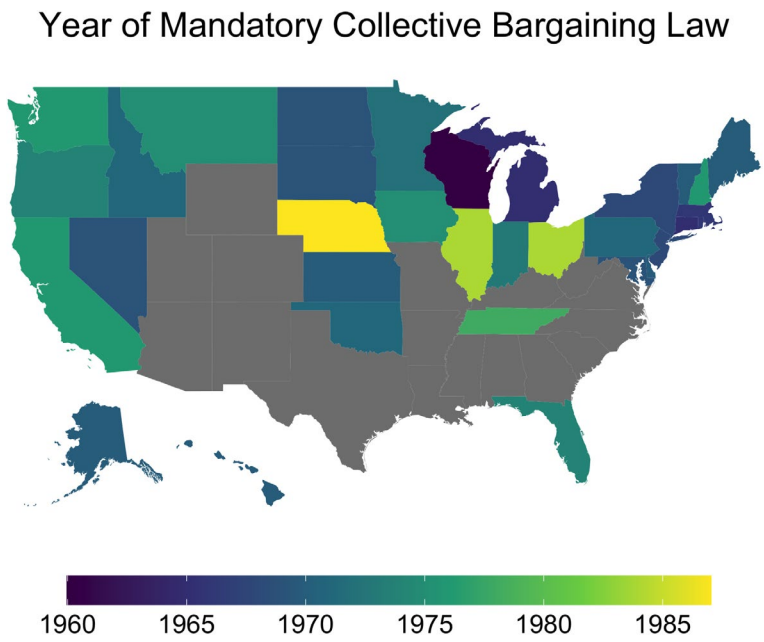


**FIGURE 1** Staggered adoption of mandatory collective bargaining laws from 1964 to 1990 [Colour figure can be viewed at wileyonlinelibrary.com]

similarly finds little effect of state unionization laws on teacher pay. Finally, Paglayan (2019) extends the historical state-level data set from Hoxby (1996). Using a variant of the two-way fixed effect model, she finds precisely estimated zero effects of mandatory bargaining laws on per-pupil school expenditures[2] and teacher salaries. Motivated in part by recent criticisms of such models (Goodman-Bacon, 2021), we revisit the Paglayan (2019) analysis using different methods.

Figure 1 shows adoption times of state mandatory bargaining laws between 1964 and 1990. Adoptions were spread across 14 separate years, although 16 states adopted laws between 1965 and 1970. Following Paglayan (2019), our main outcomes of interest are per-pupil student expenditures and teacher salaries, both measured in 2010 dollars and log transformed. We observe these outcomes back to 1959 for 49 states; we exclude Wisconsin, which adopted a mandatory bargaining law in 1960 and thus has only one year of pre-intervention data, as well as Washington, DC. This gives between 6 and 28 years of data before the adoption of mandatory bargaining, with an average of 13 years.

## 1.3 | Paper roadmap

Section 2 lays out the technical background and introduces the synthetic control estimator for a single treated unit. Section 3 bounds the estimation error for general weighting estimators under two families of data generating process, an autoregressive model and a linear factor model, with staggered adoption. Section 4 introduces partially pooled SCM as a solution to the problem of minimizing estimation error and considers two special cases, separate SCM and pooled SCM. Section 5 proposes several important extensions, including incorporating an intercept shift and

[2]Paglayan (2019) defines this as 'the total current operational expenditures (regardless of funding source) that are devoted to public schools in a state divided by the number of public school students in that state'.

auxiliary covariates, and briefly discusses inference. Section 6 describes a calibrated simulation study. Section 7 gives additional results for the teacher collective bargaining application. Finally, Section 8 discusses some directions for future work. The appendix includes further analyses and technical results. In particular, we provide an alternative motivation for our proposed partially pooled estimator, which we show is based on partially pooling parameters in the Lagrangian dual of the SCM constrained optimization problem.

## 2 | PRELIMINARIES

### 2.1 | Setup and notation

We consider a panel data setting where we observe outcomes $Y_{it}$ for $i = 1, ..., N$ units over $t = 1, ..., T$ time periods. In the teacher collective bargaining application, $N = 49$ and $T = 39$ years. Some but not all of the units adopt the treatment during the panel; once units adopt treatment, they stay treated for the remainder of the panel. Let $T_i$ represent the time period that unit $i$ receives treatment, with $T_i = \infty$ denoting never-treated units. Without loss of generality, we order units so that $T_1 \leq T_2 \leq \ldots \leq T_N$. We assume that there are a non-zero number of never-treated units, $N_0 \equiv \sum_i \mathbb{1}_{T_i = \infty}$, and we let $J = N - N_0 = \sum_i \mathbb{1}_{T_i \neq \infty}$. To clearly differentiate units that are eventually treated, we index them by $j = 1, ..., J$.

We adopt a potential outcomes framework to express causal quantities (Neyman, 1923; Rubin, 1974) and assume stable treatment and no interference between units (SUTVA; Rubin, 1980). In principle, each unit $i$ in each time $t$ might have a distinct potential outcome for each potential treatment time $s$, $Y_{it}(s)$, for $s = 1, ..., T, \infty$. Following Athey and Imbens (2021), we assume that prior to treatment, a unit's potential outcomes are equal to its never-treated potential outcome (see also Abbring & Van den Berg, 2003):

**Assumption 1** (No anticipation). $Y_{it}(s) = Y_{it}(\infty)$ for $t < s$, with treatment time $s$.

This assumption generalizes the consistency assumption typically employed in cross-sectional studies. We maintain it throughout. With it, the observed outcome is $Y_{it} = \mathbb{1}\{t < T_i\} Y_{it}(\infty) + \mathbb{1}\{t \geq T_i\} Y_{it}(T_i)$.

### 2.2 | Estimands

As is common in many panel data settings, we focus on effects a specified duration after treatment onset, known as *event time*. For treated unit $j$, we index event time relative to treatment time $T_j$ by $k = t - T_j$. The unit-level treatment effect for treated unit $j$ at event time $k$ is the difference between the potential outcome at time $T_j + k$ under treatment at time $T_j$ and under never treatment:

$$\tau_{jk} = Y_{jT_j+k}(T_j) - Y_{jT_j+k}(\infty).$$

By Assumption 1, $\tau_{jk} = 0$ for any $k < 0$.

The unit-specific effects, $\tau_{jk}$, are often the central quantities of interest in many synthetic controls analyses. In addition to these effects, we also focus on their average. Our primary averaged estimand is the average treatment effect on the treated (ATT) $k$ periods after treatment onset:

$$\text{ATT}_k \equiv \frac{1}{J}\sum_{j=1}^{J}\tau_{jk} = \frac{1}{J}\sum_{j=1}^{J} Y_{jT_j+k}(T_j) - Y_{jT_j+k}(\infty).$$

We are also interested in the average post-treatment effect, averaging across $k$: $\text{ATT} = \frac{1}{K+1}\sum_{k=0}^{K}\text{ATT}_k$. Our methods generalize to many other estimands; see Callaway and Sant'Anna (2020) for examples in this setting.

A challenge for staggered adoption analyses is that a panel that is balanced in calendar time is necessarily imbalanced in event time. That is, we observe outcomes $\ell$ periods before treatment only for units treated after period $\ell$, and we observe outcomes $k$ periods after treatment only for treated units treated before $T-k$. This means that populations of treated units over which one can average treatment effects vary with $k$, as do the possible donors. To minimize this problem, we assume that all treated units are observed for at least several periods before being treated (i.e. $T_1 \gg 1$) and for at least $K \geq 0$ periods after treatment ($T_J \leq T-K$). For treated unit $j$, we will consider outcomes up to $L_j \leq T_j - 1$ periods before treatment, with $L \equiv \max_{j\leq J} L_j$ denoting the maximum number of lagged outcomes.

With this, the challenge in estimating $\text{ATT}_k$ for $k \leq K$ is to impute the average of the missing never-treated potential outcomes. We define the set of possible 'donor units' for treated unit $j$ at event time $k$ as those units $i$ for which we observe $Y_{iT_j+k}(\infty)$, which we denote $\mathcal{D}_{jk} \equiv \{i\colon T_i > T_j + k\}$. The composition of $\mathcal{D}_{jk}$ varies with both treated unit $j$ and event time $k$; in particular, unit $i$ with $T_i < \infty$ is in $\mathcal{D}_{jk}$ for $k < T_i - T_j$ but not for $k \geq T_i - T_j$. We focus on fixed donor pools $\mathcal{D}_{jK}$ rather than allowing the donor pools to vary with $k$. This limits the number of potential donors, but ensures that estimated counterfactual outcomes do not vary spuriously across event time due to changing composition of the donor pool. Our proposed estimator does not require this restriction, but it greatly simplifies exposition. If $K \geq T_J - T_1$ then $\mathcal{D}_{jk}$ will only include never treated units as donors; otherwise $\mathcal{D}_{jk}$ will include *both* never treated and not-yet-treated units.

In our empirical application we exclude Wisconsin—which adopted a mandatory collective bargaining law in the second year of the sample—so the first treated state is Connecticut with $T_1 = 7$. We follow Paglayan (2019) in considering treatment effects only up to event time K = 10, and use as potential donors for treated state j any states that are not treated by $T_j + 10$.

## 2.3 | Restrictions on the data generating process

We now detail various restrictions on the data generating process that we will consider below. Because we are interested in treatment effects on treated units—and observe potential outcomes under treatment—we will place restrictions only on the potential outcomes under the never treated condition $Y_{it}(\infty)$ (see, e.g. Borusyak et al., 2021). Throughout, we follow Chernozhukov et al. (2021) and Ben-Michael et al. (2021) and write these potential outcomes as a model component plus additive noise.

We consider two alternative restrictions on the model terms and noise terms, corresponding to two common data generating processes for $Y_{it}(\infty)$: a time-varying autoregressive process and a linear factor model.

**Assumption 2** (Data generating processes). We consider the following:

(a) The untreated potential outcomes $Y_{it}(\infty)$ follow a time-varying AR($L$) process with coefficients at time $t$ $(\rho_{t1}, \ldots, \rho_{tL}) \in \mathbb{R}^L$:

$$Y_{it}(\infty) = \sum_{\ell=1}^{L} \rho_{t\ell} Y_{it-\ell}(\infty) + \varepsilon_{it}, \tag{1}$$

where $\varepsilon_{it}$ are mean zero and independent across units and time, with $\varepsilon_{is+k} \perp\!\!\!\perp \mathbb{1}\{T_i = s\}$ for $k \geq 0$ for all $i = 1, ..., N$.

(b) There are $F$ latent time-varying factors, where $F$ is typically small relative to both $N$ and $T$. The factors, $\mu_t \in \mathbb{R}^F$, are bounded, $\max_t \|\mu_t\|_\infty \leq M$. Each unit has a vector of time-invariant factor loadings $\phi_i \in \mathbb{R}^F$, and the untreated potential outcomes $Y_{it}(\infty)$ are generated as:

$$Y_{it}(\infty) = \phi_i \cdot \mu_t + \varepsilon_{it}, \tag{2}$$

where $\varepsilon_{it}$ are mean zero, independent across units and time and $\varepsilon_{it} \perp\!\!\!\perp T_i$ for all $i = 1, ..., N$, $t = 1, ..., T$.

Assumptions 2a and 2b impose different restrictions on the noise terms. Assumption 2b rules out correlation between treatment timing and the noise terms for any period while Assumption 2a only excludes correlation for noise terms *after* treatment. Therefore, under Assumption 2b treatment timing and pre-treatment outcomes are only dependent through the factor loadings, while under Assumption 2a there is no restriction on their dependence.

Finally, under each process, we assume that the noise terms do not have fat tails.

**Assumption 3.** $\varepsilon_{it}$ are sub-Gaussian random variables with scale parameter $\sigma$.

We use this restriction on the tail behaviour for the finite sample estimation error bounds we introduce in Section 3.

## 2.4 | The synthetic control method

In the synthetic control method (SCM), the counterfactual outcome under control is estimated from a weighted average, known as a *synthetic control*, of untreated units, where weights are chosen to minimize the squared imbalance between the lagged outcomes for the treated unit and the weighted control ('donor') units.

We consider a modified version of the original SCM estimator of Abadie et al. (2010, 2015) for a single treated unit $j$. In this version, the SCM weights $\widehat{\gamma}_j$ are the solution to a constrained optimization problem:

$$\min_{\gamma_j \in \Delta_j^{\text{scm}}} \underbrace{\frac{1}{L_j} \sum_{\ell=1}^{L_j} \left( Y_{jT_j-\ell} - \sum_{i=1}^{N} \gamma_{ij} Y_{iT_j-\ell} \right)^2}_{\text{objective}} + \underbrace{\lambda \sum_{i=1}^{N} \gamma_{ij}^2}_{\text{regularization}}, \tag{3}$$

where $\gamma_j \in \Delta_j^{\text{scm}}$ has elements $\{\gamma_{ij}\}$ that satisfy $\gamma_{ij} \geq 0$ for all $i$, $\sum_i \gamma_{ij} = 1$, and $\gamma_{ij} = 0$ whenever $i$ is not a possible donor, $i \notin \mathcal{D}_{jK}$.

Given an $N$-vector of weights $\widehat{\gamma}_j$ that solve Equation (3), the SCM estimate of the missing potential outcome for treated unit $j$ at event time $k$, $Y_{jT_j+k}(\infty)$, is:

$$\widehat{Y}_{jT_j+k}(\infty) = \sum_{i=1}^{N} \widehat{\gamma}_{ij} Y_{iT_j+k},$$

with estimated treatment effect $\widehat{\tau}_{jk} = Y_{jT_j+k} - \widehat{Y}_{jT_j+k}(\infty)$. This formulation can also be applied when $k < 0$, generating *placebo* treatment effect estimates, often referred to as 'gaps'. We denote the vector of placebo pre-treatment effect estimates as $\widehat{\tau}_j^{\text{pre}} = (\widehat{\tau}_{j(-L)}, \ldots, \widehat{\tau}_{j(-1)}) \in \mathbb{R}^L$, where we define $\widehat{\tau}_{j(-\ell)}$ to be zero for $\ell > L_j$. With this notation, the synthetic controls objective in Equation (3) is the mean squared placebo treatment effect on pre-treatment outcomes:

$$(q_j(\widehat{\gamma}_j))^2 \equiv \frac{1}{L_j} \left\| \widehat{\tau}_j^{\text{pre}} \right\|_2^2 = \frac{1}{L_j} \sum_{\ell=1}^{L_j} \left( Y_{jT_j-\ell} - \sum_{i=1}^{N} \widehat{\gamma}_{ij} Y_{iT_j-\ell} \right)^2. \tag{4}$$

The optimization problem in Equation (3) modifies the original SCM proposal in two key ways. First, where Abadie et al. (2010, 2015) balance auxiliary covariates, we focus exclusively on lagged outcomes; we re-introduce auxiliary covariates in Section 5.2. Second, following a suggestion in Abadie et al. (2015), we include a term that penalizes the weights towards uniformity, with hyperparameter $\lambda$. While we penalize the sum of the squared weights, there are many options, for example, an entropy or elastic net penalty (see Abadie & L'Hour, 2021; Doudchenko & Imbens, 2017). In settings where it is possible to achieve perfect balance, selecting $\lambda > 0$ ensures that Equation (3) has a unique solution. This is not the case in our setting, however, and so we largely view this term as a technical convenience.

Abadie (2019) gives several reasons for preferring SCM to outcome models such as linear regression or directly fitting the factor model. In particular, SCM weights are guaranteed to be non-negative, and are generally sparse and interpretable. By contrast, alternatives based on explicit models for $Y_{it}(\infty)$ often imply negative weights and thus unchecked extrapolation outside the support of the donor units. Outcome modelling can also be sensitive to model mis-specification, such as selecting an incorrect number of factors in a factor model. Finally, as we emphasize in our theoretical results in the next section, SCM can be appropriate under multiple data generating processes (e.g. both the autoregressive model and the linear factor model) so that it is not necessary for the applied researcher to take a strong stand on which is correct.

A central question for SCM is how to assess whether $\widehat{Y}_{jT_j+k}(\infty)$ is a reasonable estimate for $Y_{jT_j+k}(\infty)$. A minimal condition is that the SCM weights achieve a low root mean squared placebo treatment effect, that is, $q_j(\widehat{\gamma}_j)$ is close to zero. If it is not close to zero, there is a concern that estimated effects also capture systematic differences between $\widehat{Y}_{jT_j+k}(\infty)$ and $Y_{jT_j+k}(\infty)$. Under versions of either Assumptions 2a or 2b and for a single treated unit, Abadie et al. (2010) show that if $q_j(\widehat{\gamma}_j) = 0$ then the bias will tend to zero as $L_j \to \infty$; Ben-Michael et al. (2021) bound the estimation error of $\widehat{\tau}_{jk}$ in terms of $q_j(\widehat{\gamma}_j)$. Abadie et al. (2010, 2015) recommend that researchers only proceed with an SCM analysis if the pre-treatment fit is excellent, while Ben-Michael et al. (2021) propose an augmented SCM estimator that attempts to salvage cases where it is not.

## 3 | ESTIMATION ERROR UNDER STAGGERED ADOPTION

In order to extend SCM to the staggered adoption setting, we first develop appropriate balance measures for synthetic control-style weighting estimators under staggered adoption. We use

these to develop bounds on the estimation error for the ATT for our two example data generating processes. These bounds in turn motivate our proposal for partially pooled SCM as a way to choose weights under staggered adoption.

## 3.1 | Weights and measures of balance

With multiple treated units, we can generalize the above setup to allow for weights for each treated unit. For each $j \leq J$, let $\gamma_j \in \Delta_j^{\text{scm}}$ be an $N$-vector of weights on potential donor units, where $\gamma_{ij}$ is the weight on unit $i$ in the synthetic control for treated unit $j$. We collect the weights into an $N$-by-$J$ matrix $\Gamma = [\gamma_1, \ldots, \gamma_J] \in \Delta^{\text{scm}}$, where $\Delta^{\text{scm}} = \Delta_1^{\text{scm}} \times \ldots \times \Delta_J^{\text{scm}}$. The estimated treatment effect on unit $j$ at event time $k$ is then $\widehat{\tau}_{jk}$ as defined above, and the estimated ATT averages over the unit-level effect estimates:

$$\widehat{\text{ATT}}_k = \frac{1}{J}\sum_{j=1}^{J}\widehat{\tau}_{jk} = \frac{1}{J}\sum_{j=1}^{J}\left[Y_{jT_j+k} - \sum_{i=1}^{N}\widehat{\gamma}_{ij}Y_{iT_j+k}\right] = \frac{1}{J}\sum_{j=1}^{J}Y_{jT_j+k} - \sum_{i=1}^{N}\sum_{j=1}^{J}\frac{\widehat{\gamma}_{ij}}{J}Y_{iT_j+k}. \quad (5)$$

Equation (5) highlights two equivalent interpretations of the estimator: as the average of unit-specific SCM estimates and as an SCM estimate for the average treated unit.

Using the two interpretations of the ATT estimator in Equation (5), we construct goodness-of-fit measures for the ATT by aggregating $\widehat{\tau}_j^{\text{pre}}$ in two ways. First, we consider the root mean square of the pre-treatment fits across treated units,

$$q^{\text{sep}}(\widehat{\Gamma}) \equiv \sqrt{\frac{1}{J}\sum_{j=1}^{J}q_j^2(\widehat{\gamma}_j)} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{L_j}\|\widehat{\tau}_j^{\text{pre}}\|_2^2} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{L_j}\sum_{\ell=1}^{L_j}\left(Y_{jT_j-\ell} - \sum_{i=1}^{N}\widehat{\gamma}_{ij}Y_{iT_j-\ell}\right)^2}.$$

This is a useful measure of overall imbalance when SCM is estimated separately for each treated unit and generalizes the objective for the single synthetic control problem. Second, we consider the pre-treatment fit for the average of the treated units,

$$q^{\text{pool}}(\widehat{\Gamma}) \equiv \frac{1}{\sqrt{L}}\left\|\frac{1}{J}\sum_{j=1}^{J}\widehat{\tau}_j^{\text{pre}}\right\|_2 = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\left[\frac{1}{J}\sum_{T_j>\ell}Y_{jT_j-\ell} - \sum_{i=1}^{N}\widehat{\gamma}_{ij}Y_{iT_j-\ell}\right]^2}.$$

We refer to this interchangeably as the *pooled* or *global* fit.

Both $q^{\text{pool}}$ and $q^{\text{sep}}$ are on the same scale as the estimated treatment effect, $\widehat{\text{ATT}}_k$. However, the measures differ in whether they average *before* or *after* evaluating the pre-treatment fit. Thus, we typically expect $(q^{\text{pool}})^2 \ll (q^{\text{sep}})^2$, since the lagged outcomes for the *average* of the treated units are less extreme than the lagged outcomes for the units themselves. In practice, we therefore consider normalizing the imbalance measures by their values computed with weights $\widehat{\Gamma}^{\text{sep}}$, the set of solutions to Equation (3) applied separately to each treated unit. We define normalized measures $\tilde{q}^{\text{pool}}(\Gamma) \equiv q^{\text{pool}}(\Gamma)/q^{\text{pool}}(\widehat{\Gamma}^{\text{sep}})$ and $\tilde{q}^{\text{sep}}(\Gamma) \equiv q^{\text{sep}}(\Gamma)/q^{\text{sep}}(\widehat{\Gamma}^{\text{sep}})$, and use them in our proposed estimator in Section 4 below.

Ideally, both $q^{\text{sep}}$ and $q^{\text{pool}}$ would be close to zero; indeed if $q^{\text{sep}} = 0$ then $q^{\text{pool}}$ is also zero. When this is not possible, there is a trade-off between these two sources of imbalance.

Our proposed 'partially pooled' SCM estimator generalizes Equation (3) to minimize a weighted average of their normalized squares, $\nu(\tilde{q}^{\text{pool}})^2 + (1-\nu)(\tilde{q}^{\text{sep}})^2$, where $\nu$ is a hyperparameter selected by the researcher. To motivate this and to inform the choice of $\nu$, we develop error bounds for SCM-style weights under our two data generating models.

## 3.2 | Error bounds

### 3.2.1 | Autoregressive model

We first bound the estimation error for the ATT under the autoregressive process in Assumption 2a. To simplify notation and concepts, we initially focus on the ATT at event time $k = 0$, $\text{ATT}_0$. Two summaries of the autoregressive coefficients are important to our analysis: $\overline{\rho} = \frac{1}{J}\sum_{j=1}^{J}\rho_{T_j}$, the *average* autoregression coefficient across the $J$ treatment times, and $S_\rho^2 \equiv \frac{1}{J}\sum_{j=1}^{J}\|\rho_{T_j} - \overline{\rho}\|_2^2$, the corresponding *variance*; this variance is zero under simultaneous adoption, $S_\rho^2 = 0$.

**Theorem 1.** *Under Assumptions 2a and 3 with $L_j = L < T_1$ for* j = 1, ..., J, *for $\widehat{\Gamma} \in \Delta^{\text{scm}}$, where $\widehat{\gamma}_j$ is independent of $\epsilon_{\cdot T_j+k}$, and for any $\delta > 0$, the error for $\widehat{\text{ATT}}_0$ is*

$$\left|\widehat{\text{ATT}}_0 - \text{ATT}_0\right| \le \underbrace{\sqrt{L}\|\overline{\rho}\|_2 q^{\text{pool}}(\widehat{\Gamma})}_{\text{pooled fit}} + \underbrace{\sqrt{L}S_\rho\, q^{\text{sep}}(\widehat{\Gamma})}_{\text{unit-specific fit}} + \underbrace{\frac{\delta\sigma}{\sqrt{J}}\left(1 + \|\widehat{\Gamma}\|_F\right)}_{\text{noise}}$$

*with probability at least $1 - 2e^{-\frac{\delta^2}{2}}$, where for a matrix $A \in \mathbb{R}^{n\times m}$, $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}A_{ij}^2}$ is the Frobenius norm.*

Theorem 1 shows that the error for the ATT is bounded by several distinct terms, giving guidance for the choice of the weights $\Gamma$. First, error arises from the level of both the global fit and the unit-specific fits. The relative importance of these fits is governed by the ratio of the average coefficient value $\|\overline{\rho}\|_2$ and the standard deviation $S_\rho$ for the autoregressive coefficients over time.

Second, there is error due to post-treatment noise, inherent to any weighting method. Because the weights are independent of post-treatment outcomes, this term has mean zero and enters the finite sample bound above through the standard deviation, which is proportional to the Frobenius norm of the weight matrix, $\|\widehat{\Gamma}\|_F$. Thus, when selecting among weight matrices that yield similar unit-specific and pooled balance, we should prefer the one that minimizes $\|\widehat{\Gamma}\|_F$. This motivates a penalty term similar to that in Equation (3).

Finally, we can extend the bound in Theorem 1 to $\text{ATT}_k$ by noting that the autoregressive structure implies that $Y_{iT_i+k\infty} = \sum_{\ell=1}^{L}\rho_{t\ell}^{(k)}Y_{iT_i-\ell\infty} + \sum_{s=0}^{k}\eta_s^{(k)}\epsilon_{iT_i+s}$ for some set of coefficients $\rho_{t1}^{(k)}, \ldots \rho_{tL}^{(k)}$ and $\eta_0^{(k)}, \ldots, \eta_k^{(k)}$. We can then apply Theorem 1 to obtain bounds for $\left|\widehat{\text{ATT}}_k - \text{ATT}_k\right|$ by defining $\overline{\rho}$ and $S_\rho$ in terms of the new coefficients $\rho_{t\ell}^{(k)}$ and replacing $\sigma$ with $\sigma\sqrt{1 + \sum_s\left(\eta_s^{(k)}\right)^2}$. Similarly, we can obtain bounds for the overall $\text{ATT} = \frac{1}{K+1}\sum_{k=0}^{K}\text{ATT}_k$, by noting that the average outcome over $K$+1 periods following treatment can again be written as a weighted sum of the last $L$ outcomes before treatment plus a weighted sum of the $K+1$ errors following treatment. Thus, with suitable redefinition of the parameters, Theorem 1 continues to apply.

### 3.2.2 | Linear factor model

Next we consider the linear factor model in Assumption 2b and begin by defining additional notation. Let $\Omega_j \in \mathbb{R}^{L\times F}$ denote the matrix of factor values for time $T_j - L$ to $T_j - 1$, and denote $P^{(j)} = \sqrt{L}(\Omega_j'\Omega_j)^{-1}\Omega_j' \in \mathbb{R}^{F\times L}$ as the scaled projection matrix from outcomes to factors. Analogous to the autoregressive process above, the average (projected) factor value across the $J$ treatment times, $\overline{\mu}_k = \frac{1}{J}\sum_{j=1}^{J} P^{(j)\prime}\mu_{T_j+k}$, and the *variance*, $S_k^2 = \frac{1}{J}\sum_{j=1}^{J} \|P^{(j)\prime}\mu_{T_j+k} - \overline{\mu}_k\|_2^2$, determine the relative importance of the pooled and unit-specific fits respectively.

**Theorem 2.** *Assume that* $\Omega_j$ *is non-singular and* $\|\frac{1}{\sqrt{L}}\Omega_j\|_2 = 1$ *for* j = *1, ...,* J. *With* $L_j = L < T_1$ *for* j = *1, ...,* J, $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_J \in \Delta^{\text{scm}}$ *where* $\widehat{\gamma}_j$ *is independent of* $\epsilon_{\cdot T_j+k}$, k ≥ *0, and* δ > *0, under Assumptions 2b and 3 the error for* $\widehat{\text{ATT}}_k$ *is*

$$\left|\widehat{\text{ATT}}_k - \text{ATT}_k\right| \le \underbrace{\|\overline{\mu}_k\|_2\, q^{\text{pool}}(\widehat{\Gamma})}_{\text{pooled fit}} + \underbrace{S_k\, q^{\text{sep}}(\widehat{\Gamma})}_{\text{unit-specific fit}} + \underbrace{\frac{\sigma M^2 F}{\sqrt{L}}\left(3\delta + 2\sqrt{\log NJ}\right)}_{\text{approximation error}} + \underbrace{\frac{\delta\sigma}{\sqrt{J}}\left(1 + \|\widehat{\Gamma}\|_F\right)}_{\text{noise}}$$

*with probability at least* $1 - 6e^{-\frac{\delta^2}{2}}$, *where* $\max_t \|\mu_t\|_\infty \le M$.

Theorem 2 shows that under the linear factor model the error for the ATT can again be controlled by the level of pooled fit and unit-specific fits. As in Theorem 1, the relative importance of these fits is governed by the ratio of the average factor value $\overline{\mu}_k$ and the standard deviation $S_k$; similarly, under simultaneous adoption, $S_k = 0$ and $q^{\text{sep}}$ does not enter the bound.

Unlike in Theorem 1, this bound also includes an approximation error that arises due to balancing—and possibly over-fitting to—noisy outcomes rather than to the true underlying factor loadings. In the worst case, the $J$ synthetic controls match on the noise rather than the factors. Constraining the weights to lie in the simplex reduces the impact of this worst case, however, and the error decreases as more lagged outcomes are balanced; see Abadie et al. (2010); Ben-Michael et al. (2021); Arkhangelsky et al. (2019) for further discussion.

Finally, we can extend Theorem 2 to the estimation error of the overall post-treatment effect, $\text{ATT} = \frac{1}{K+1}\sum_{k=0}^{K}\text{ATT}_k$, by noting that the average post-treatment potential outcome also follows a linear factor structure with factor values $\frac{1}{K+1}\sum_{k=0}^{K}\mu_{T_j+k}$ and noise term $\frac{1}{K+1}\sum_{k=0}^{K}\epsilon_{iT_j+k}$. Thus the pooled- and unit-specific fit terms and the approximation error will depend, respectively, on the average, variance and maximum of the (projected) average post-treatment factor value, and the noise term will be reduced by a factor of $\frac{1}{\sqrt{K+1}}$.

## 4 | PARTIALLY POOLED SCM

We now turn to our main proposal, *partially pooled SCM*. Motivated by the finite sample error bounds in Theorems 1 and 2, this chooses SCM weights to minimize a weighted average of the (squared) pooled and unit-specific pre-treatment fits:

$$\min_{\Gamma\in\Delta^{\text{scm}}} \nu(\tilde{q}^{\text{pool}}(\Gamma))^2 + (1-\nu)(\tilde{q}^{\text{sep}}(\Gamma))^2 + \lambda\|\Gamma\|_F^2. \tag{6}$$

The hyperparameter $\nu \in [0, 1]$ governs the relative importance of the two objectives; higher values of $\nu$ correspond to more weight on the pooled fit relative to the separate fit. In Appendix A.3, we show that intermediate values of $\nu$ correspond to a partial pooling solution for the weights in the dual parameter space, motivating our choice of a name.

The optimization in Equation (6) differs from the bounds in Section 3 in two practical ways. First, we minimize the normalized imbalance measures (e.g. $\tilde{q}^{\text{pool}}$ rather than $q^{\text{pool}}$), so that the minimum with $\nu = 0$ and $\lambda = 0$ is indexed to 1. This ensures that the two objectives are on the same scale, regardless of the number of treated units, and makes it easier to form intuition about $\nu$. Second, we minimize the squared imbalances, which permits a computationally feasible quadratic program. As with the single synthetic controls problem in Equation (3), we penalize the sum of the squared weights, $\|\Gamma\|_F^2$.

## 4.1 | Special cases: Separate SCM ($\nu = 0$) and Pooled SCM ($\nu = 1$)

We first consider two special cases of Equation (6), which correspond to extreme values of the hyperparameter $\nu$, and then consider intermediate cases.

To date, common practice for staggered adoption applications of SCM is to estimate separate SCM fits for each treated unit, then estimate the ATT by averaging the unit-specific treatment effect estimates. This approach, which we refer to as separate SCM, minimizes $q^{\text{sep}}$ alone and is equivalent to our proposal in Equation (6) with $\nu = 0$. Since this separate SCM strategy prioritizes the unit-specific estimates, $\widehat{\tau}_{jk}$, an important question is when this approach will also give reasonable estimates of $\text{ATT}_k$. From Theorems 1 and 2, we can see that if the unit-specific fits are all excellent, then the estimation error $\left|\widehat{\text{ATT}}_k - \text{ATT}_k\right|$ will be small. However, this is not the case in our application. Figure 2a shows SCM 'gap plots' of $\widehat{\tau}_{j\ell}$ against $\ell$ for three illustrative treated states, taken one at a time. While Ohio shows relatively good pre-treatment fit, there are no synthetic controls that closely track Illinois or New York's pre-treatment outcomes. Thus, simply averaging the estimated treatment effects across these three states without attention to the overall fit does not yield a convincing estimate. Other recent applications also face the same issue where several treated units have poor pre-treatment fit (see e.g. Dube & Zipperer, 2015; Donohue et al., 2019).[3]

The other extreme case, which we refer to as *pooled SCM*, instead sets $\nu = 1$, finding weights that minimize $q^{\text{pool}}$, the root mean squared placebo estimate of the ATT. This ignores the unit-specific pre-treatment fits in the objective, resulting in poor unit-level synthetic controls and, in turn, leading to poor estimates of the unit-level treatment effects $\tau_{jk}$. Furthermore, even if the ATT is the only estimand of interest, Theorems 1 and 2 indicate that separate SCM is unlikely to control the error. In particular, if the pooled weights do a poor job of matching individual treated units, the pooled synthetic control may involve a great deal of interpolation and the component of the error bound due to separate imbalance can be large. In Section 6 we validate through simulation that pooled SCM leads to substantially worse unit-level estimates than separate SCM, and also that there are indeed settings where the bounds in Theorems 1 and 2 do bind, leading to large error in pooled SCM estimates of the ATT. See Abadie and L'Hour (2021) for further discussion on interpolation bias in synthetic control settings.

However, there are special cases where only controlling $q^{\text{pool}}$ with pooled SCM is sufficient. Theorems 1 and 2 indicate that only the across treated unit variation in $\rho_{T_j+k}$ and $\mu_{T_j+k}$ leads to

[3]One way to address this is to trim the sample and drop treated units with poor pre-treatment fit, noting that this changes the estimand.

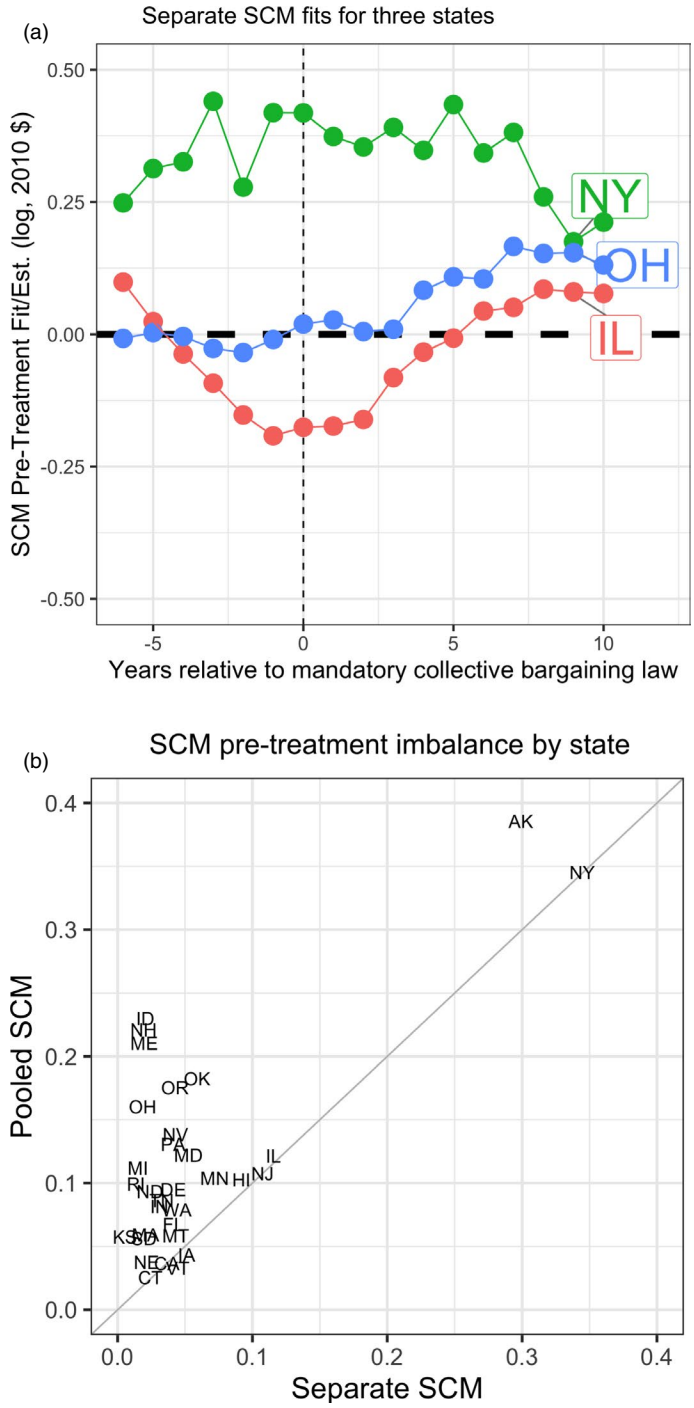


**FIGURE 2** (a) SCM pre-treatment fit for three states: (i) Ohio, with good overall fit, (ii) Illinois, where SCM fails to match an important pre-treatment trend, and (iii) New York, with pre-treatment imbalance roughly an order of magnitude larger than typical estimates for the impact of teacher mandatory bargaining. (b) SCM fits by state show that Separate SCM gives better pre-treatment fit than Pooled SCM for all treated states [Colour figure can be viewed at wileyonlinelibrary.com]

unit-specific fits contributing to the error bounds. Thus, when this variation is zero, the ATT error bound is minimized with $\nu = 1$. As we discuss above, under simultaneous adoption, with $T_1 = \ldots = T_J$, $S_\rho = 0$ in the autoregressive model and $S_k = 0$ in the linear factor model. The same arises in staggered adoption settings where the data generating process is homogeneous over time—for example, where $\rho_t \equiv \rho$ in the autoregressive model. It also holds approximately when the average autoregressive coefficient or factor values are large relative to the standard deviations—that is, $S_\rho \ll \overline{\rho}$ or $S_k \ll \overline{\mu}_k$, which could justify a choice of $\nu = 1$. Finally, when units are treated in cohorts (with $T_j = T_k$ for units in the same cohort), there is no variation in $\rho_t$ and $\mu_t$ across units in the same cohort. This suggests fully pooling (i.e. averaging) units that are treated at the same time, even if there is only partial pooling across treatment cohorts. We discuss this modification in Appendix A.2.

Figure 2b plots the state-level pre-treatment imbalances in our application for separate SCM versus pooled SCM. The separate SCM fit is better for all treated states, and so leads to more credible unit-level estimates. However, these fits are far from perfect and so the results from Section 3 imply that there is room for improvement by controlling the pooled fit. Figure 3a shows the implied placebo estimates for the overall ATT using the separate and pooled approaches: they are consistently positive for separate SCM weights and are all nearly zero for pooled SCM weights. At the same time, Figure 3b shows that pooled SCM has very poor unit-level fit, leading to the potential for error for both the overall ATT estimate and the unit-level estimates. This motivates choosing an intermediate choice of $\nu \in (0, 1)$.

## 4.2 | Intermediate choice of $\nu$

As we have seen, it is important to control both the pooled fit (for the ATT) and the unit-level fits (for both the ATT and the unit-level estimates). The hyper-parameter $\nu$ controls the relative weight of these in the objective.

One approach to choosing $\nu$ is to return to the error bounds in Theorems 1 and 2. The optimization problem in Equation (6) can be seen as a first-order approximation to the squares of the error bounds. Therefore, if the parameters of those bounds are known—and our only goal is to estimate the ATT—we can use these to choose an appropriate $\nu$.[4] Unfortunately, these will generally be infeasible as the analyst will not know these parameters, although in some applications it may be possible to obtain pilot estimates.

An alternative approach is to directly assess the implications of the choice of $\nu$ for the imbalance criteria for both the overall ATT and the unit-level effects. Figure 4 provides two views of this for the teacher collective bargaining application. Figure 4a shows the *balance possibility frontier*: the *y*-axis shows the pooled imbalance $q^{\text{pool}}$ and the *x*-axis shows the unit-level imbalance $q^{\text{sep}}$. The curve traces out how these change as we vary $\nu$ from the separate SCM solution at the upper left to the pooled solution at the lower right. The relationship is strongly convex, indicating that by accepting a very small increase in pooled imbalance from the fully pooled solution we can obtain large reductions in unit-level imbalance, and vice versa starting from the separate $\nu = 0$ solution. See King et al. (2017) and Pimentel and Kelz (2020) for other examples of balance frontiers in observational settings.

Figure 4b plots the two imbalances, here normalized as $\tilde{q}^{\text{pool}}$ and $\tilde{q}^{\text{sep}}$, to put them on comparable scales, against $\nu$. As $\nu$ rises, pooled imbalance falls while unit-level imbalance rises,

[4]For example, in the autoregressive model, letting $a = \|\overline{\rho}\|_2 q^{\text{pool}}(\widehat{\Gamma}^{\text{sep}})$ and $b = S_\rho q^{\text{sep}}(\widehat{\Gamma}^{\text{sep}})$, we could choose $\nu = \frac{a^2}{a^2 + b^2}$, with comparable quantities for the linear factor model.

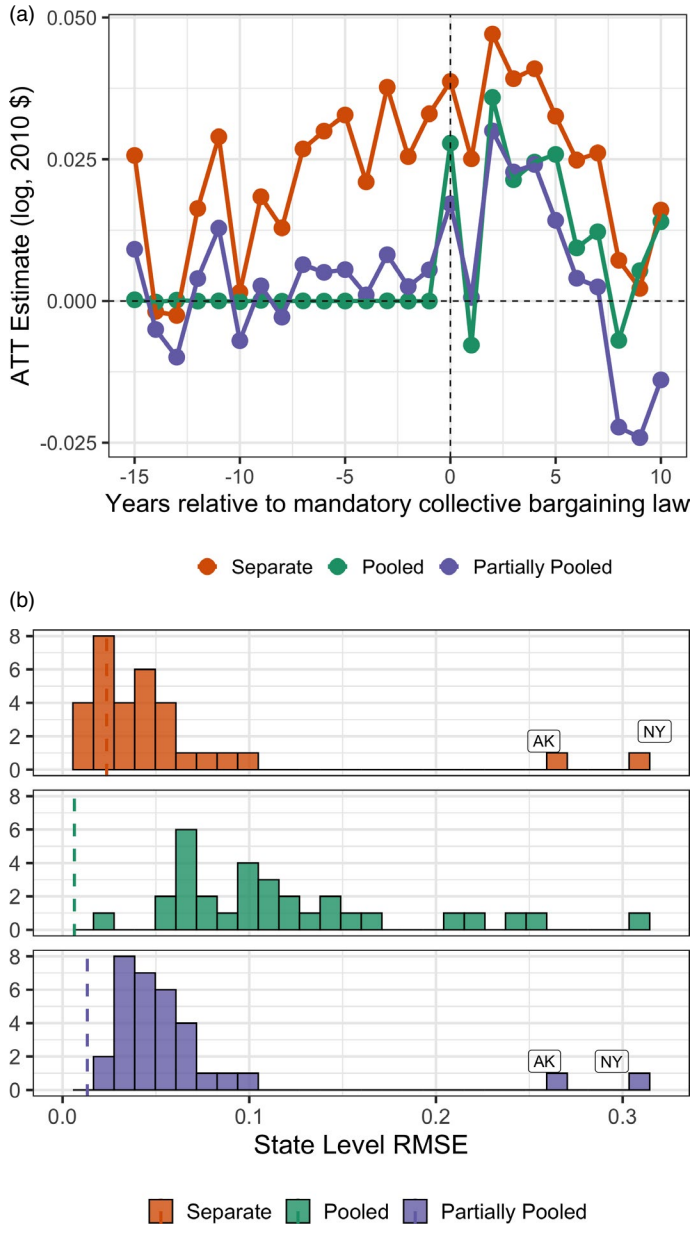


**FIGURE 3** (a) Series of estimated pre- and post-treatment effects $\widehat{\text{ATT}}_\ell$ and (b) state-level pre-treatment RMSE $\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\widehat{\tau}_{j\ell}^2}$ using separate, pooled, and partially pooled SCM [Colour figure can be viewed at wileyonlinelibrary.com]

although this is highly nonlinear, as the convex frontier in Figure 4a suggests. Moving from the separate SCM estimate of $\nu = 0$ to a partially pooled SCM estimate of $\nu = 0.5$ reduces the pooled imbalance by 80%, with more modest further reductions as $\nu \to 1$. Meanwhile, the unit-level imbalance declines quickly as $\nu$ falls from 1 to 0.9, then more slowly as $\nu$ declines further. Even a very small deviation from the pooled SCM solution, such as moving from $\nu = 1$ to $\nu = 0.99$, cuts the unit-level imbalance by 30% with essentially no change in the pooled fit. Due to the number of degrees of freedom involved, the pooled imbalance

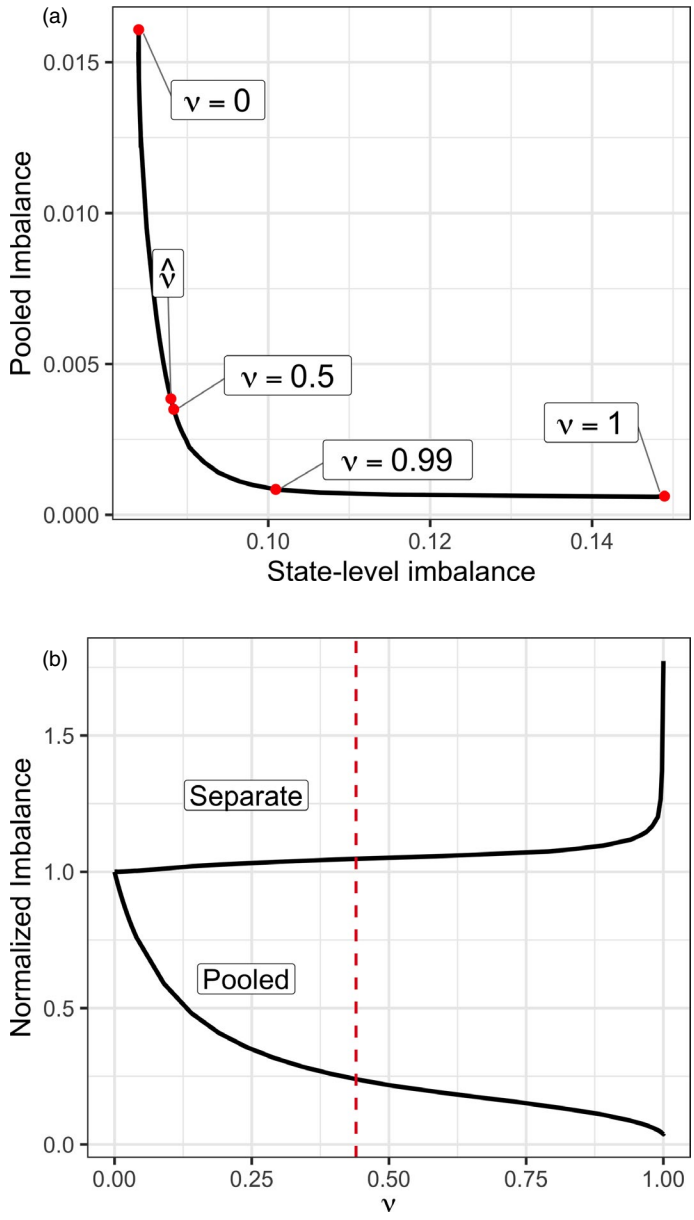


**FIGURE 4** (a) The trade-off between pooled imbalance ($q^{\text{pool}}$) and unit-specific imbalance ($q^{\text{sep}}$) as $\nu$ varies, where $\nu = 0$ is the separate SCM solution and $\nu = 1$ is the pooled SCM solution. (b) $q^{\text{sep}}$ and $q^{\text{pool}}$ versus $\nu$, each normalized by their values for separate SCM. The dashed red line indicates $\hat{\nu}$. The large distance in unit-level imbalance between $\nu = 0.99$ and $\nu = 1$ suggest meaningful gains in balance from deviating from the complete pooling estimate even by a small amount [Colour figure can be viewed at wileyonlinelibrary.com]

will often be near zero for $\nu = 1$, and the objective function $q^{\text{pool}}$ will be relatively flat in the neighbourhood of the pooled solution. Therefore we expect that in many cases it will be possible to trade-off a small increase in pooled imbalance for a large decrease in the unit-level imbalance, yielding a better estimator of both the overall ATT and the unit-level estimates at relatively little cost. We view the balance possibility frontier plot in Figure 4a as

an important tool for using partially pooled SCM in practice. By tracing out the curve, practitioners can see the trade-offs between the pooled and unit-level fit, and choose $\nu$ according to the trade-off they desire.

In our application, we use a simple heuristic to set $\nu$ based on the pooled fit of separate SCM, $q^{\text{pool}}(\widehat{\Gamma}^{\text{sep}})$, which we also use to normalize our objective function in Equation (6). We set $\nu$ to be the ratio of the pooled fit to the average unit-level fit: $\widehat{\nu} = \sqrt{L}\, q^{\text{pool}}(\widehat{\Gamma}^{\text{sep}}) / \frac{1}{J}\sum_{j=1}^{J} \sqrt{L_j}\, q_j(\widehat{\gamma}_j^{\text{sep}})$. This is bounded above by 1 due to the triangle inequality.[5] The key idea is that, if the separate SCM problem with $\nu = 0$ achieves good *pooled* fit on its own, then we want to select a small $\nu$, which will ensure both good unit-specific and pooled fit. Conversely, if the pooled fit of separate SCM is poor, then there can be substantial gains to giving $q^{\text{pool}}$ higher priority by setting $\nu$ to be large. In Section 6 we find through simulation that this heuristic results in weights that significantly reduce both the estimation error for the ATT relative to separate SCM and the estimation error of the unit-level effects relative to pooled SCM.

In the teacher bargaining example, our heuristic yields $\widehat{\nu} \approx 0.44$ for the per-pupil expenditure outcome, and we label this point in Figure 4a. The heuristic choice has similar global pre-treatment imbalance to the fully pooled estimator, $\nu = 1$, with only a modest increase in unit-level imbalance relative to the separate SCM estimate, $\nu = 0$. This is reflected in Figure 3, which also shows the placebo ATT estimates for partially pooled SCM. While the imbalance for the ATT is slightly larger than for pooled SCM, it is substantially better than for separate SCM.

There are many other potential choices for $\nu$, and, even if we focus solely on the ATT, this one is unlikely to be optimal. An alternative strategy when the balance possibility frontier exhibits a strong 'kink' shape is to choose $\nu$ to be the point after which small improvements to the pooled fit lead to substantially worse unit-level fits. Another heuristic is to choose $\nu$ to be the point where the tangent of the frontier is equal to the slope between the end points at $\nu = 0$ and $\nu = 1$ ($\nu = .84$ in the teacher bargaining application).

In the end, the nonlinear relationship between $\nu$ and $\{q^{\text{sep}}, q^{\text{pool}}\}$ in Figure 4b suggests that the loss from choosing a suboptimal $\nu$ is likely to be small, so long as we do not choose something too close to 0 or 1. We also recommend inspecting the sensitivity of estimates to the particular choice of $\nu$ in practice; we do this in Section 7.

## 5 | EXTENSIONS

We now add two elaborations to the basic setup. First, we incorporate an intercept shift into the SCM problem, following proposals by Doudchenko and Imbens (2017) and Ferman and Pinto (2021). Second, we incorporate auxiliary covariates alongside lagged outcomes. We conclude by briefly addressing inference in this setting.

### 5.1 | Incorporating intercept shifts

We have established that the partially pooled SCM estimator achieves nearly as good overall balance as the fully pooled estimator, while achieving much better balance for each unit.

[5] If the SCM fits with $\nu$=0 are perfect for each unit, $\frac{1}{J}\sum_{j=1}^{J} \sqrt{L_j}\, q_j = 0$, then the overall fit will also be perfect, $\sqrt{L}\, q^{\text{pool}} = 0$, and our heuristic sets $\widehat{\nu} = 0$. This is not a common situation.

Nevertheless, unit-level balance is often imperfect. Particularly when the scale of the outcome varies across units, it can be difficult to construct an adequate synthetic control, as one needs to match both the overall level and patterns over time. Several recent papers have proposed modifying SCM for a single treated unit by allowing for an *intercept shift* between the treated unit and its synthetic control (Abadie, 2019; Doudchenko & Imbens, 2017; Ferman & Pinto, 2021). We can adapt this approach to the staggered adoption setting by including an additional parameter vector $\alpha \in \mathbb{R}^J$, where $\alpha_j$ is an intercept term for unit $j$. We include this intercept in the counterfactual estimate as

$$\widehat{Y}_{jt}(\infty) = \alpha_j + \sum_{i=1}^{N} \gamma_{ij} Y_{it}$$

and in the separate and pooled imbalance measures as

$$(q^{\text{sep}}(\alpha, \Gamma))^2 = \frac{1}{2J} \sum_{j=1}^{J} \left[ \frac{1}{L_j} \sum_{\ell=1}^{L_j} \left( Y_{jT_j-\ell} - \alpha_j - \sum_{i=1}^{N} \gamma_{ij} Y_{iT_j-\ell} \right)^2 \right],$$

and

$$(q^{\text{pool}}(\alpha, \Gamma))^2 = \frac{1}{L} \sum_{\ell=1}^{L} \left[ \frac{1}{J} \sum_{T_j > \ell} \left( Y_{jT_j-\ell} - \alpha_j - \sum_{i=1}^{N} \gamma_{ij} Y_{iT_j-\ell} \right) \right]^2.$$

Again we can define normalized versions of these objectives, $\tilde{q}^{\text{pool}}(\alpha, \Gamma) \equiv q^{\text{pool}}(\alpha, \Gamma)/q^{\text{pool}}(\widehat{\alpha}^{\text{sep}}, \widehat{\Gamma}^{\text{sep}})$, where $\widehat{\alpha}^{\text{sep}}$ and $\widehat{\Gamma}^{\text{sep}}$ are the minimizers of $(q^{\text{sep}}(\alpha, \Gamma))^2$. As above, we then form an overall objective function as a convex combination of the normalized squares:

$$\min_{\alpha \in \mathbb{R}^J, \Gamma \in \Delta^{\text{scm}}} \quad \nu \, (\tilde{q}^{\text{pool}}(\alpha, \Gamma))^2 + (1 - \nu) \, (\tilde{q}^{\text{sep}}(\alpha, \Gamma))^2 \; + \; \lambda \|\Gamma\|_F^2. \tag{7}$$

The intercept $\widehat{\alpha}$ that solves Equation (7) has a closed form in terms of the solution for the weights, $\widehat{\Gamma}^*$; $\widehat{\alpha}_j$ is the average pre-treatment difference between treated unit $j$ and its synthetic control,

$$\widehat{\alpha}_j = \frac{1}{L_j} \sum_{\ell=1}^{L_j} Y_{jT_j-\ell} - \frac{1}{L_j} \sum_{i=1}^{N} \sum_{\ell=1}^{L_j} \widehat{\gamma}_{ij}^* Y_{jT_j-\ell}. \tag{8}$$

Plugging this value of $\widehat{\alpha}$ into Equation (7), we see that this procedure is equivalent to solving the partially pooled SCM problem (6) using the *residuals* $\dot{Y}_{iT_j-\ell} \equiv Y_{iT_j-\ell} - \frac{1}{L_j} \sum_{\ell=1}^{L_j} Y_{iT_j-\ell}$. The resulting treatment effect estimates have a particularly useful form:

$$\widehat{\tau}_{jk}^* = \frac{1}{L_j} \sum_{\ell=1}^{L_j} \left[ \left( Y_{jT_j+k} - Y_{jT_j-\ell} \right) - \sum_{i=1}^{N} \widehat{\gamma}_{ij}^* \left( Y_{iT_j+k} - Y_{iT_j-\ell} \right) \right], \tag{9}$$

and

$$\widehat{\text{ATT}}_k^* = \frac{1}{J}\hat{\tau}_{jk}^* = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{L_j}\sum_{\ell=1}^{L_j}\left[\left(Y_{jT_j+k} - Y_{jT_j-\ell}\right) - \sum_{i=1}^{N}\hat{\gamma}_{ij}^*\left(Y_{iT_j+k} - Y_{iT_j-\ell}\right)\right]\right]. \quad (10)$$

We can view this as a weighted difference-in-differences (DiD) estimator. In the special case with uniform weights over units, $\hat{\gamma}_{ij}^* = 1/\|\mathcal{D}_j\|$, Equation (9) is the simple average over all two-period, two-group DiD estimates, averaging over all pre-treatment lags $\ell$ and donor units $i$. This is equivalent to recent proposals for DiD estimators that allow for treatment effect heterogeneity with a fixed donor set per treatment time cohort (see Callaway & Sant'Anna, 2020; Sun & Abraham, 2020, among others). With non-uniform weights, $\hat{\tau}_{jk}^*$ compares the change in outcomes for treated unit $j$ to the change for the synthetic control, rather than the average change across all potential donors. Equation (10) averages these estimates across treated units $j$ to form $\widehat{\text{ATT}}_k^*$.

Figure 5 shows the value of including an intercept to improving pre-treatment fit in the teacher collective bargaining application. Figure 5a presents this as a balance possibility frontier for SCM with the weights alone and with the intercept, as well as the implied imbalance for the DiD estimator alone. Here, simple unweighted DiD achieves unit-level and pooled balance that improves on the no-intercept SCM possibility frontier. However, the intercept-shifted estimator dominates both DiD and no-intercept SCM estimates on both criteria, for all but the largest $\nu$. We see similar results when examining the state-specific fits. Figure 5b shows the unit-level fit for both partially pooled SCM and the intercept-augmented version. Two states, New York and Alaska, have especially bad pre-treatment fits without including an intercept because they have the highest per-pupil expenditures of all the states for many years (see Appendix Figure B.5). Accounting for the pre-treatment average through the intercept dramatically improves the fits for these states.

## 5.2 | Incorporating auxiliary covariates

We have focused thus far on matching pre-treatment values of the outcome variable. In practice, we typically observe a set of auxiliary covariates $X_i \in \mathbb{R}^d$ as well. In our collective bargaining application, we consider five covariates, measured as of the start of the sample in 1959–1960: income per capita, the student to teacher ratio, the per cent of the population with 12+ and 13+ years of education, and the female labour force participation rate.[6] We standardize all five covariates to have mean zero and variance one.

There are several ways to incorporate auxiliary covariates in the setting with a single treated unit. Here we directly include them into the optimization problem. Analogous to above, we define both the unit-level imbalance and pooled imbalance of $X$,

$$q_X^{\text{sep}}(\Gamma) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left\|X_j - \sum_{i=1}^{N}\gamma_{ij}X_i\right\|_2^2},$$

[6]Due to missing data for these auxiliary covariates, we restrict our analysis here to the contiguous United States. Note that this drops Alaska, which we have seen is far outside the convex hull of its donor units.

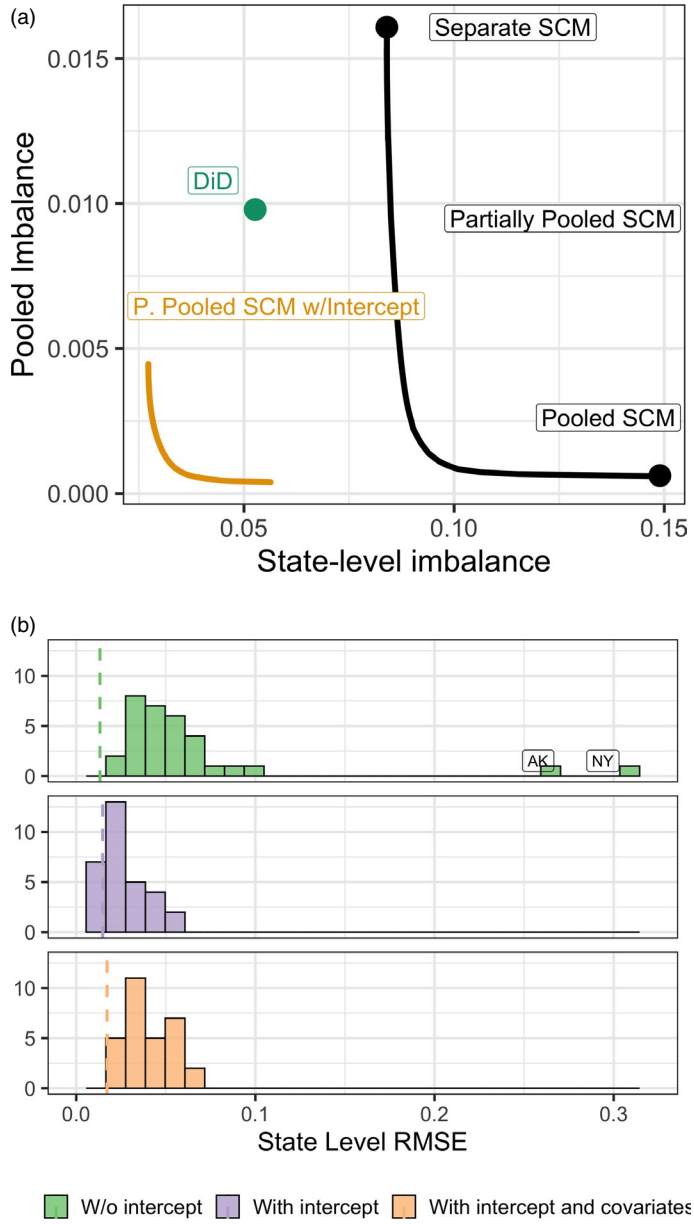


**FIGURE 5** (a) The balance possibility frontier for SCM with and without an intercept, as well as the implied imbalance for DiD. Incorporating unit-level fixed effects leads to substantial improvements in balance. For DiD, we compute the implied balance as $\sqrt{\sum_{\ell=1}^{L}\left(\widehat{\mathrm{ATT}}^{*}_{-\ell}\right)^{2}}$, the RMSE of the placebo estimates, from Equation (9) with uniform weights. (b) The distribution of state-level fits (in terms of RMSE) with and without an intercept and covariates; dashed lines show the pooled pre-treatment RMSE [Colour figure can be viewed at wileyonlinelibrary.com]

and another for the pooled synthetic control,

$$q_X^{\mathrm{pool}}(\Gamma) = \left\| \frac{1}{J}\sum_{j=1}^{J} X_j - \sum_{i=1}^{N} \gamma_{ij} X_i \right\|_2,$$

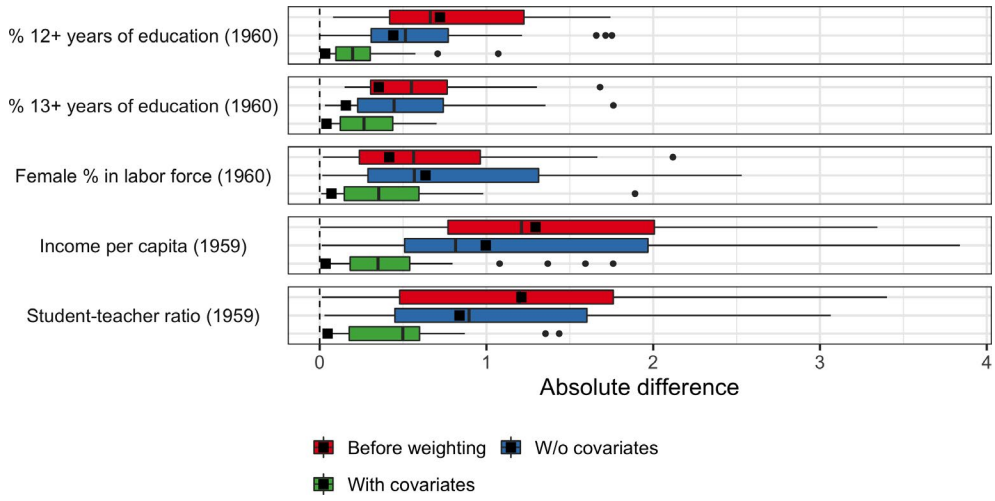


**FIGURE 6** Distribution of the absolute difference between each treated unit and its synthetic control for the (standardized) auxiliary covariates, before weighting and with/without including covariates in the optimization procedure. Black squares show the absolute average difference [Colour figure can be viewed at wileyonlinelibrary.com]

with normalized versions $\tilde{q}_X^{\text{sep}}(\Gamma)$ and $\tilde{q}_X^{\text{pool}}(\Gamma)$.[7] We then include these in our objective, with an additional hyper-parameter $\xi$:

$$\min_{\alpha \in \mathbb{R}^J, \Gamma \in \Delta^{\text{scm}}} \nu \left( (\tilde{q}^{\text{pool}}(\alpha, \Gamma))^2 + \xi (\tilde{q}_X^{\text{pool}}(\Gamma))^2 \right) + (1-\nu) \left( (\tilde{q}^{\text{sep}}(\alpha, \Gamma))^2 + \xi (\tilde{q}_X^{\text{sep}}(\Gamma))^2 \right) + \lambda \|\Gamma\|_F^2. \quad (11)$$

While we write this optimization problem with an intercept shift, we could also include auxiliary covariates but no intercept. The choice of $\xi$ determines the relative importance of the outcomes and the auxiliary covariates. Setting $\xi = 0$ recovers the optimization problem (7) without auxiliary covariates, while in the extreme case setting $\xi = \infty$ will, if feasible, enforce exact balance on the auxiliary covariates. We decide to give equal priority to both terms. Since the auxiliary covariates are standardized, we set $\xi$ to be the sample variance of the pre-$T_J$ outcomes for the never treated units. This equally weights both components in the objective functions, and reduces the number of hyper-parameters and specification choices. Finally, we can incorporate time-varying covariates by including the values at time periods before the first treatment time $T_1$ into the vector $X_i$.

Figure 6 shows the level of covariate balance between each treated unit and its synthetic control, as well as for the average across treated units. Before weighting there are large differences between the treated units and their donor sets, and weighting on the outcomes alone does little to alleviate these differences. Including the auxiliary covariates into the optimization procedure finds weights that give nearly perfect covariate balance for the pooled synthetic control (indicated as the black squares), while also significantly improving covariate balance for the individual treated units (indicated as boxplots). Figure 5b shows that this

[7]Specifically, let $\hat{\alpha}^{\text{sep}}$ and $\hat{\Gamma}^{\text{sep}}$ be the minimizers of $(q^{\text{sep}}(\alpha, \Gamma))^2 + \xi (q_X^{\text{sep}}(\Gamma))^2$, and $(C^{\text{sep}})^2 = (q^{\text{sep}}(\hat{\alpha}^{\text{sep}}, \hat{\Gamma}^{\text{sep}}))^2 + \xi (q_X^{\text{sep}}(\hat{\Gamma}^{\text{sep}}))^2$ and $(C^{\text{pool}})^2 = (q^{\text{pool}}(\hat{\alpha}^{\text{sep}}, \hat{\Gamma}^{\text{sep}}))^2 + \xi (q_X^{\text{pool}}(\hat{\Gamma}^{\text{sep}}))^2$ be the combined separate and pooled imbalances. We define the normalized objectives as $\tilde{q}_X^{\text{pool}}(\Gamma) = q_X^{\text{pool}}(\Gamma)/C^{\text{pool}}$, $\tilde{q}_X^{\text{sep}}(\Gamma) = q_X^{\text{sep}}(\Gamma)/C^{\text{sep}}$, and slightly abuse notation by re-defining $\tilde{q}^{\text{pool}}(\alpha, \Gamma) \equiv q^{\text{pool}}(\alpha, \Gamma)/C^{\text{pool}}$ and $\tilde{q}^{\text{sep}}(\alpha, \Gamma) \equiv q^{\text{sep}}(\alpha, \Gamma)/C^{\text{sep}}$.

improved covariate balance comes at a small cost to the fit on the pre-treatment outcomes: the distribution of unit-level pre-treatment root mean square error (RMSE) shifts slightly to the right.

## 5.3 | Inference

There is a growing literature on inference for SCM-type estimators, although no proposed approach is fully satisfactory for all cases. In settings where multiple units adopt treatment simultaneously, Abadie and L'Hour (2021) propose an extension of the original permutation procedure of Abadie et al. (2010), and Arkhangelsky et al. (2019) propose resampling-based approaches. In a staggered adoption setting, Shaikh and Toulis (2021) propose a weighted permutation approach based on a Cox proportional hazards model. This is not appropriate in our application, however, since multiple units have the same treatment time, which is incompatible with the Cox model. Finally, Cao and Lu (2019) propose an Andrews test for inference with intercept-shifted SCM under staggered adoption. Building on the existing literature, we consider constructing confidence intervals via the wild bootstrap. We briefly describe this method here; we address asymptotic Normality and inference via the jackknife in Appendix A.1.

The wild bootstrap approach we implement adapts the proposal from Otsu and Rai (2017) for bias-corrected matching estimators; see also Imai et al. (2019). First, we can re-write $\widehat{\text{ATT}}_k$ as the following average over units:

$$\widehat{\text{ATT}}_k = \frac{1}{J}\sum_{i=1}^{N}\sum_{g=T_1}^{T_J}\left(\mathbb{1}_{T_i=g} - \sum_{T_j=g}\widehat{\gamma}_{ij}\right)\left(Y_{ig+k} - \frac{1}{g-1}\sum_{\ell=1}^{g-1}Y_{ig-\ell}\right) = \frac{1}{J}\sum_{i=1}^{N}\tilde{\tau}_i. \tag{12}$$

This bootstrap procedure draws a sequence of random variables $W_1^{(b)}, \ldots, W_N^{(b)}$ independently with $P(W_i = -(\sqrt{5}-1)/2) = (\sqrt{5}+1)/2\sqrt{5}$ and $P(W_i = (\sqrt{5}+1)/2) = (\sqrt{5}-1)/2\sqrt{5}$ for $b = 1, \ldots, B$, and computes the boostrap statistic:

$$S^{(b)} = \frac{1}{J}\sum_{i=1}^{N} W_i^{(b)}\left(\tilde{\tau}_i - \widehat{\text{ATT}}_k\right), \tag{13}$$

for each draw. Letting $q_{\alpha/2}$ and $q_{1-\alpha/2}$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of $S^{(b)}$, we construct confidence intervals via $[\widehat{\text{ATT}}_k - q_{1-\alpha/2}, \widehat{\text{ATT}}_k + q_{\alpha/2}]$. Importantly, we keep the weights and outcomes fixed, and only re-sample the multiplier variables $W_i^{(b)}$.

In the next section, we evaluate the coverage of the wild bootstrap with a simulation study that mimics the structure of the collective bargaining application. In Appendix A.1, we take an alternative route and motivate the use of resampling methods via asymptotic Normality. In particular, we provide a set of sufficient conditions for $\widehat{\text{ATT}}_k - \text{ATT}_k$ to be asymptotically Normal. We consider an asymptotic regime in which $J, N_0 \to \infty$, with the number of lags $L$ fixed and the number of control units growing faster than the number of treated units $\frac{J}{N_0} \to \infty$. We also adapt a generalization of the conditional parallel trends assumption in Abadie (2005) to the staggered adoption setting. However, there are several ways such asymptotic results can be misleading. First, our result assumes that the synthetic control weights can achieve perfect fit within treatment time

cohorts, which ensures that the distribution of $\widehat{\text{ATT}}_k$ is centred around $\text{ATT}_k$. Poor fit, either overall or across time cohorts, can lead to under-coverage. Second, the asymptotic approximation can be poor when there are relatively few total units, and the use of resampling methods can exacerbate this. Thus, while we show that these approaches yield reasonable results in simulations, we suggest interpreting any confidence intervals for typical applications with caution.

## 6 | SIMULATION STUDY

We now consider the performance of different approaches in a simulation study calibrated to the collective bargaining data set; we turn to the impacts of mandatory teacher collective bargaining laws in the actual data in the next section. We evaluate performance with three different data generating processes. First, we generate never treated outcomes according to a two-way fixed effects model,

$$Y_{it}(\infty) = \text{int} + \text{unit}_i + \text{time}_t + \epsilon_{it}, \tag{14}$$

with both unit and time effects are normalized to have mean zero. This model satisfies the parallel trends assumption needed for the DiD estimator we consider below. We estimate (14) using only the never-treated observations, and extract the estimated variance of the unit effects, $\widehat{\Sigma}$, and of the error term, $\widehat{\sigma}^2_\epsilon$. We then generate $\text{unit}_i \overset{\text{iid}}{\sim} N(0, \widehat{\Sigma})$ and $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \widehat{\sigma}^2_\epsilon)$.

Second, we use a factor model with a two-dimensional latent time-varying factor $\mu_t \in \mathbb{R}^2$ and unit-specific coefficients $\phi_i \in \mathbb{R}^2$:

$$Y_{it}(\infty) = \text{int} + \text{unit}_i + \text{time}_t + \phi_i' \mu_t + \epsilon_{it}. \tag{15}$$

We estimate (15) using the `R` package `gsynth` (Xu, 2017) for the untreated units and time periods, then estimate the variance–covariance matrix of the unit fixed effects and factor loadings, $\widehat{\Sigma}$, and the variance of the error term $\widehat{\sigma}^2_\epsilon$. Here we use the estimated $\{\widehat{\text{time}}_t, \widehat{\mu}_t\}$, and draw $\{\text{unit}_i, \phi_i\} \overset{\text{iid}}{\sim} \text{MVN}(0, \widehat{\Sigma})$ and $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \widehat{\sigma}^2_\epsilon)$.

Finally, we have a random effects autoregressive model:

$$Y_{it}(\infty) = \sum_{\ell=1}^{3} \rho_\ell Y_{it-\ell}(\infty) + \epsilon_{it}, \quad \rho \sim N(\mu_\rho, \sigma^2_\rho), \tag{16}$$

that we fit using `lme4` (Bates et al., 2015) to obtain estimates $\widehat{\mu}_\rho$ and $\widehat{\sigma}_\rho$. In order to increase the level of heterogeneity across time, we simulate from this hierarchical model with eight times the standard deviation $8\widehat{\sigma}_\rho$. For all three outcome processes we generate simulated data sets with the same dimensions as the data, $N = 49$ and $T = 39$, and impose a sharp null of no treatment effect, $Y_{it}(s) = Y_{it}(\infty) = Y_{it}$.

A key component of the simulation model is selection into treatment. We fix the treatment times to be the same as in the teacher unionization application. For each treatment time, we assign treatment to those units not already treated with probability $\pi_i$, sweeping through the fixed set of treatment times. For the two-way fixed effects model, we set the probability that unit $i$ is treated at each treatment time to be $\pi_i = \text{logit}(\theta_0 + \theta_1 \cdot \text{unit}_i)$, with $\theta_0 = -2.7$ and $\theta_1 = -1$, yielding around 30 units that are eventually treated in each simulation draw. For the factor model

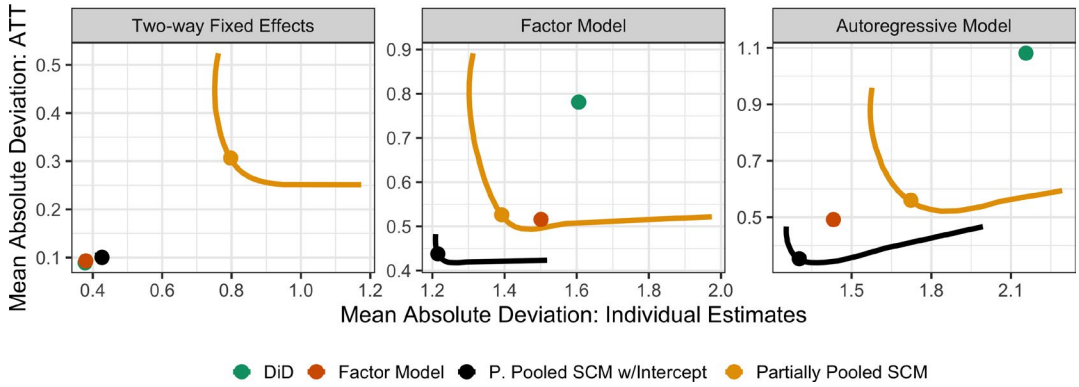


**FIGURE 7** Monte Carlo estimates of the MAD for the overall ATT vs the MAD for the individual ATT estimates. The lines trace out values for $\nu \in [0, 1]$, the solid points are the average value using the heuristic $\hat{\nu}$. In the two-way fixed effects and factor model simulations, the estimated factor model is the oracle estimator. Among the alternatives, the intercept-shifted partially pooled SCM has lowest MAD for both the overall ATT and the individual ATT estimates [Colour figure can be viewed at wileyonlinelibrary.com]

we choose $\pi_i = \text{logit}(\theta_0 + \theta_1(\text{unit}_i + \phi_{i1} + \phi_{i2}))$, and set $\theta_0 = -2.7$ and $\theta_1 = -1$ so that around 32 units are eventually treated in each simulation draw, following the distribution of the data. For the autoregressive process we allow selection to depend on the three lagged outcomes $\pi_i = \text{logit}\left(\theta_0 + \theta_1 \sum_{\ell=1}^{3} Y_{it-\ell}\right)$, where $\theta_0 = \log 0.04$ and $\theta_1 = -2$.

**Estimation.** We consider several estimators for the average post-treatment effect ATT. Figure 7 shows four: (1) A difference-in-differences estimator following Equation (9) with uniform weights, (2) the partially pooled SCM estimator, as we vary $\nu$ between 0 and 1, (3) partially pooled SCM with an intercept, again varying $\nu$ and (4) directly estimating the factor model. Solid points indicate the heuristic choice of $\hat{\nu}$ above. The vertical axis of each panel shows the mean absolute deviation (MAD) for the ATT, $\mathbb{E}\left[\left|\text{ATT} - \widehat{\text{ATT}}\right|\right]$, while the horizontal axis shows the average of the individual post-treatment effect estimates, $\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\tau_j - \hat{\tau}_j|\right]$. Appendix Figures B.1 and B.2 show the analogous results for the bias and RMSE.

There are several key takeaways from Figure 7. First, under each data generating process there is a trade-off between estimating the ATT and the individual effects, with $\nu = 1$ at the top left of the 'MAD frontier' and $\nu = 0$ at the bottom right. Partially pooled SCM significantly reduces the bias for the overall ATT relative to separate SCM, and a small amount of pooling also leads to slightly better individual ATT estimates. The gains to pooling, however, diminish for $\nu$ close to 1, with the fully pooled SCM yielding poor individual ATT estimates under all three models. Under a two-way fixed effects model there is no penalty to pooling in terms of MAD for the overall ATT. This comports with Theorem 2, which shows that targeting the pooled pre-treatment fit is sufficient under a two-way fixed effects model. However, under the factor model and AR process the fully pooled estimator leads to worse MAD for the overall ATT estimates than partially pooled SCM. Second, when mis-specified, the DiD estimator does not do particularly well at controlling the MAD for either overall ATT or the unit-level estimates. Third, the intercept-shifted estimator dominates either of the alternatives in terms of both overall and unit-level estimates. Here again there are gains to partially pooling SCM, albeit with the possibility for a large amount of error from over-pooling. Fourth, our heuristic choices of $\nu$ perform reasonably well at selecting a point close to the value that minimizes the MAD for the ATT, while also reducing the MAD for the

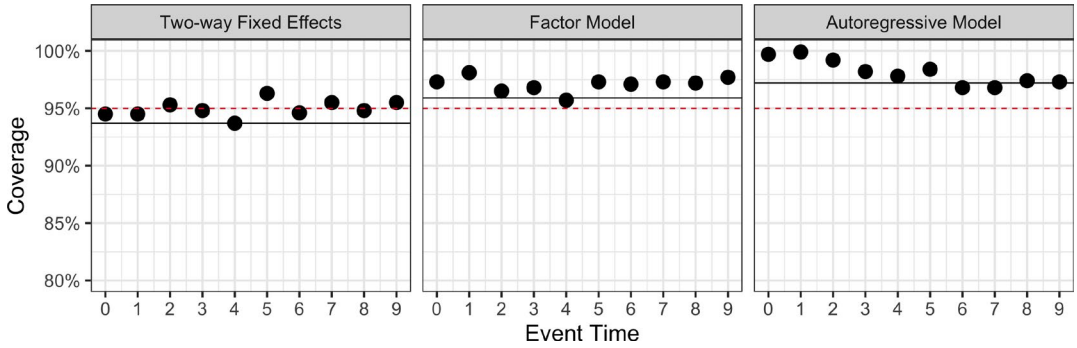


**FIGURE 8** Monte Carlo estimates of the coverage of approximate 95% confidence intervals $k = 0, ..., 9$ periods after treatment. The solid line indicates the coverage for the overall ATT estimate averaged across all post-treatment periods [Colour figure can be viewed at wileyonlinelibrary.com]

individual estimates. Finally, the partially pooled SCM estimator with an intercept shift performs as well as or better than fitting the factor model directly.

**Inference.** We conclude by examining the finite-sample coverage of approximate 95% confidence intervals from the wild bootstrap. Figure 8 shows the coverage of approximate confidence intervals for partially pooled SCM with an intercept shift, using the wild bootstrap to construct the intervals. Under the two-way fixed effects model, in which there is no bias from inexact fit, the wild bootstrap has close to 95% coverage. Under both the linear factor model and the autoregressive model, however, the wild bootstrap is somewhat conservative.[8] Overall, the wild bootstrap appears to be a reasonable, if conservative, choice.

## 7 | IMPACTS OF MANDATORY TEACHER COLLECTIVE BARGAINING LAWS

We now return to measuring the impact of mandatory teacher collective bargaining. The left of Figure 9a shows the placebo estimates from Equation (9), where $k < 0$.[9] We see that along with the good unit-specific fits shown in Figure 5b and the good covariate balance shown in Figure 6, the pooled synthetic control estimate is near zero for $k < 0$. The right side of the figure shows the estimated impact on per-pupil current expenditures, with approximate 95% confidence intervals computed via the wild bootstrap.

Consistent with Paglayan (2019), we find weakly negative effects of mandatory teacher collective bargaining laws on student expenditures. Pooled across the 11 years after treatment adoption, the overall estimate is $\widehat{\text{ATT}} = -0.03$, or a 3% decrease in per-pupil expenditures, with an approximate 95% confidence interval of [−0.06, +0.005]. In Appendix Figure B.7 we show the

[8]Appendix Figure B.3 shows the analogous results for partially pooled SCM without including an intercept. In this case, the wild bootstrap is extremely conservative.

[9]These placebo checks differ from those typically performed in traditional event studies, which test for the parallel trends assumption by comparing pre-treatment outcomes between treated and control units. These tests generally have low power, however; see, for example, Roth (2018); Bilinski and Hatfield (2018); Kahn-Lang and Lang (2019). In contrast, the intercept-shifted estimator uses pre-treatment outcomes to select donor units that best balance the treated units, in effect optimizing for the placebo test. It is still possible to inspect pre-treatment fit, as in standard SCM, but this is best seen as an assessment of the quality of the match rather than as a formal placebo test.

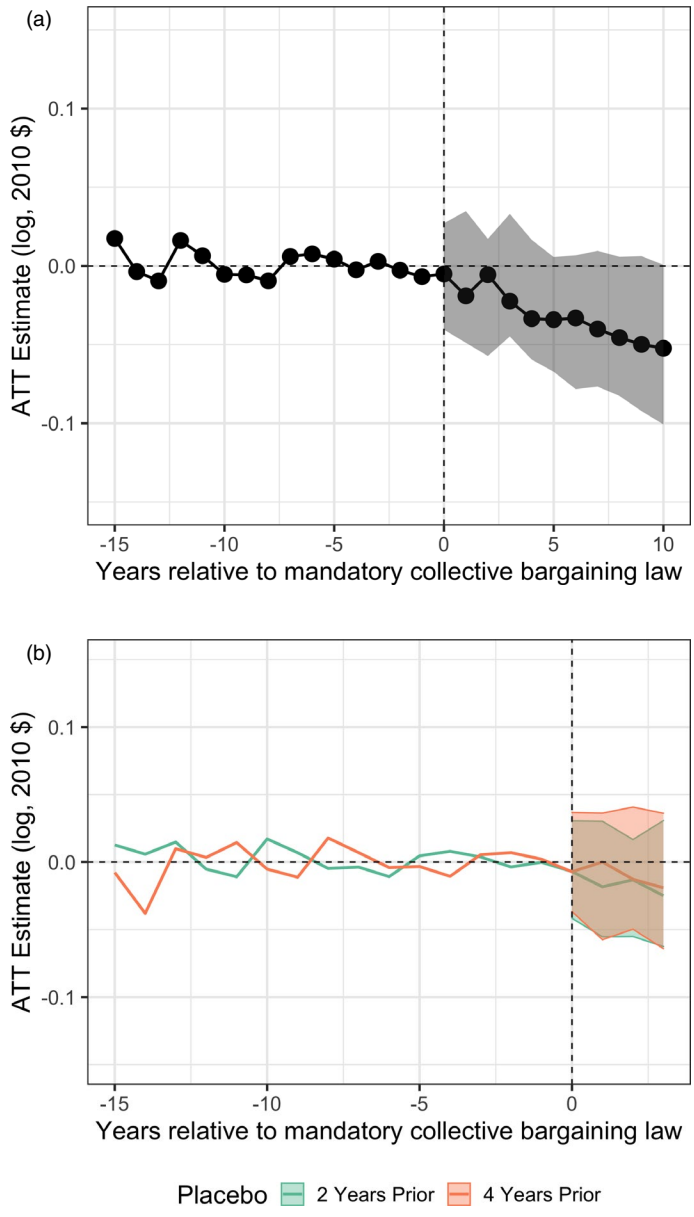


**FIGURE 9** Estimates of the ATT on per-pupil current expenditures (log, 2010 $) and placebo estimates re-indexing treatment time to 2 and 4 years before the true treatment time. The placebo effects are very close to zero and are indistinguishable from zero at this level of precision [Colour figure can be viewed at wileyonlinelibrary.com]

average post-treatment effect for each state and the unit-level fits. For those states with good pre-treatment fit, we find small positive and negative effects, while we estimate larger negative effects for those with worse fit. These estimates are in stark contrast to the results from Hoxby (1996), who argues for a 12% positive effect, although she gives a range of estimates. One possible explanation for this is that school districts are able to divert funds from other purposes to fund higher teacher salaries with minimal net effect on total expenditures. In Appendix Figure B.7 we show estimates of the effect on teacher salaries, finding evidence against a positive effect.

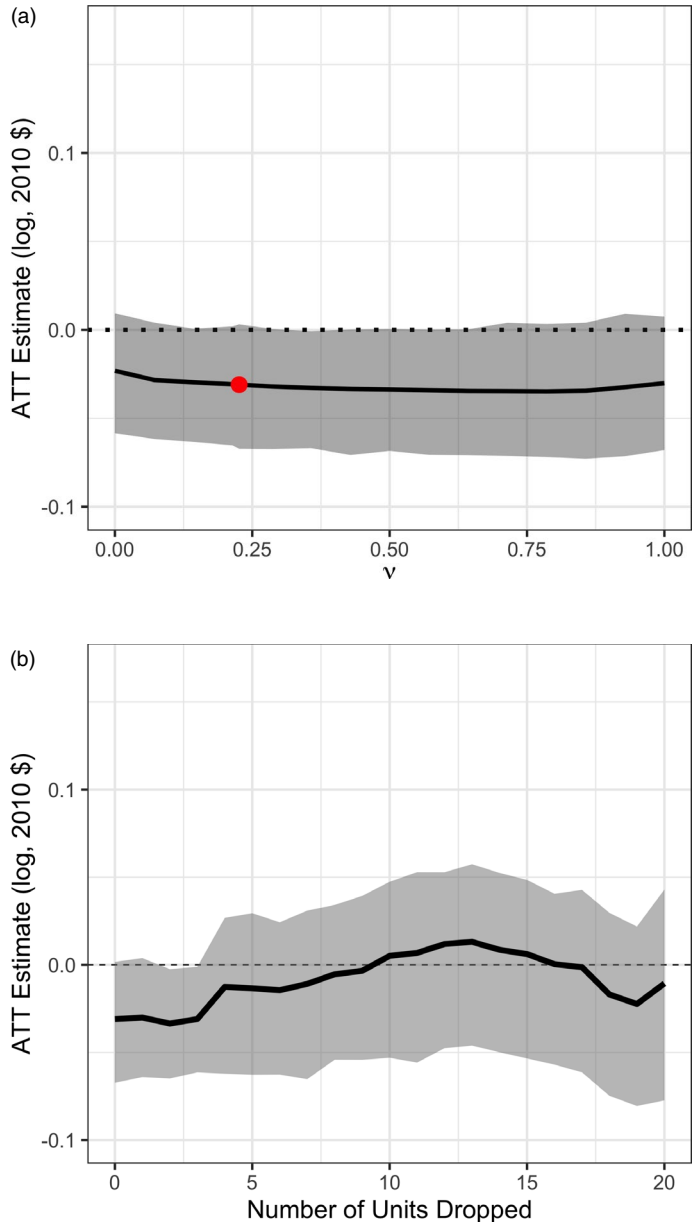


**FIGURE 10** (a) $\widehat{\text{ATT}}$ and approximate 95% confidence intervals as $\nu$ varies between 0 and 1, $\hat{\nu}$ highlighted. (b) Estimates are not especially sensitivity to dropping an increasing number of units (ranked by pre-treatment imbalance), although the uncertainty intervals are wider with fewer units in the analysis [Colour figure can be viewed at wileyonlinelibrary.com]

We can assess the strength of evidence by conducting robustness and placebo checks. First, following Abadie et al. (2015), we begin by assessing out-of-sample validity via *in time placebo checks*. These checks hold out some pre-treatment time periods by re-indexing treatment time to be earlier (i.e. setting $T_j' = T_j - x$ for some $x$), then estimate placebo effects for the held-out pre-intervention time periods. Figure 9b shows the placebo estimates for the intercept-shifted partially pooled SCM estimator with covariates using a placebo treatment time two and four periods

before the true treatment time. Both estimators achieve excellent pre-treatment fit and estimate placebo effects that are indistinguishable from zero.

Another important check that we recommend in practice is to gauge the sensitivity of the ATT estimates to the particular choice of pooling parameter $\nu$. Figure 10a shows the overall ATT estimates varying $\nu$ from separate SCM $\nu = 0$ to pooled SCM $\nu = 1$. No choice of $\nu$ substantively changes the conclusions, and each rules out large positive effects. Finally, we consider the result of trimming states with poor pre-treatment fit, following common practice in the matching and SCM literatures. Figure 10b shows the overall ATT estimates when removing an increasing number of treated units with poor fits, in order of decreasing unit-level fit. Overall, omitting the worst-fit states decreases the magnitude of the estimated effect, and increases the variability of the estimate. However, all estimates still rule out large positive effects.

An important feature of SCM-based methods over model-based methods is that we can directly inspect the weights, and that these weights are non-negative and sum to one. Appendix Figures B.8 and B.9 show the state-specific weights over donor states for each treated unit for partially pooled SCM without an intercept and with both an intercept and auxiliary covariates respectively. Without the intercept, both Illinois and Wyoming are consistently important donor states. Both states had relatively high levels of per-pupil expenditures throughout the study period and several synthetic controls place nearly all of the weight on these two states in order to match the level. However, after removing pre-treatment averages via an intercept, the weights are much more evenly distributed across the donor pool, suggesting that estimates are not overly reliant on a single control unit.

## 8 | DISCUSSION

In this paper, we develop a new framework for estimating the impact of a treatment adopted gradually by units over time. In our motivating example, 33 states have enacted laws mandating school districts to bargain with teachers' unions (Paglayan, 2019), and we seek to estimate the effects of these laws on educational expenditures. To do so, we adapt SCM to the staggered adoption setting. We argue that current practice of estimating separate SCM weights for each treated unit is unlikely to yield good results, but also that fully pooled SCM may over-correct; our preferred approach, partially pooled SCM, finds weights that balance both state-specific and overall pre-treatment fit. We then extend this basic approach to incorporate an intercept shift as well as auxiliary covariates. We apply this approach to the teacher bargaining example and, consistent with recent analyses, find weakly negative estimates on student expenditures.

We briefly note some directions for future work. First, we could extend these ideas to other settings with multiple treated units, such as where treatment can 'shut off' for some units (Imai & Kim, 2021), or where all units are eventually treated (Athey & Imbens, 2021). This would likely require additional assumptions. We could similarly incorporate other structure from our application. For example, in staggered adoption settings where multiple units adopt treatment at the same time, we could add a layer in the hierarchy and more closely pool units treated at the same time while still partially pooling different treatment cohorts. See Appendix A.2.

Second, many SCM analyses explore multiple outcomes. As in other SCM studies, we treat each outcome separately, choosing different synthetic control weights for each. In many settings, however, lagged values from one outcome may predict future values of another, suggesting that balancing multiple outcome variables would be useful. This seems especially important in settings like ours with relatively few units.

Finally, we could adapt recent proposals for bias correction and other 'doubly robust' estimators to this setting, which will be important for both estimation and inference (Abadie & L'Hour, 2021; Arkhangelsky et al., 2019; Ben-Michael et al., 2021). Existing approaches have largely been limited to the case with a single treated unit or, if multiple units are treated, to a single adoption time. More complex models are possible and may be desirable in the staggered adoption setting. For example, Fesler and Pender (2019) apply the Ridge Augmented SCM proposal in Ben-Michael et al. (2021) to a staggered adoption setting, modelling each treated unit separately. Partial pooling may be helpful here. In another direction, we might consider an outcome model that incorporates the time weights used in Arkhangelsky et al. (2019). We anticipate that, unlike in the simple case with unit fixed effects, these augmented approaches likely require more elaborate shrinkage estimation, such as via matrix penalties.

## ACKNOWLEDGEMENTS

We thank Alberto Abadie, Howard Bloom, Peng Ding, Arin Dube, Guido Imbens, Skip Hirshberg, Brian Jacob, Luke Keele, Luke Miratrix, Joe Ornstein, Agustina Paglayan, Sam Pimentel, Jake Soloff, Panos Toulis, Chelsea Zhang and Ben Zipperer for useful discussion and comments, as well as participants at the 2019 Atlantic Causal Inference Conference. We also thank the associate editor and reviewers for constructive feedback. This research was supported in part by the Opportunity Lab and the Institute for Research on Labor and Employment at UC Berkeley, as well as the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## REFERENCES

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

Abadie, A. (2019) Using synthetic controls: feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59, 391–425.

Abadie A. & L'Hour J. (2021) A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 1–34. http://dx.doi.org/10.1080/01621459.2021.1971535

Abadie, A., Diamond, A. & Hainmueller, J. (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.

Abadie, A., Diamond, A. & Hainmueller, J. (2015) Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), 495–510.

Abbring, J.H. & Van den Berg, G.J. (2003) The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5), 1491–1517.

Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W. & Wager, S. (2019) Synthetic differences in differences. *American Economic Review. Forthcoming*. https://www.aeaweb.org/articles?id=10.1257/aer.20190159

Athey, S. & Imbens, G.W. (2021) Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*. In press. https://doi.org/10.1016/j.jeconom.2020.10.012

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. & Khosravi, K. (2021) Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 1–41. Ahead of print. https://doi.org/10.1080/01621459.2021.1891924

Bates, D., Mächler, M., Bolker, B.M. & Walker, S.C. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Michael, E., Feller, A. & Rothstein, J. (2021) The augmented synthetic control method [just-accepted]. *Journal of the American Statistical Association*, 1–34. In press. https://doi.org/10.1080/01621459.2021.1929245

Bilinski, A. & Hatfield, L.A. (2018) Seeking evidence of absence: reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273*.

Borusyak, K., Jaravel, X. & Spiess, J. (2021) Revisiting event study designs: robust and efficient estimation.

Callaway, B. & Sant'Anna, P.H. (2020) Difference-in-differences with multiple time periods. *Journal of Econometrics*. In press. https://doi.org/10.1016/j.jeconom.2020.12.001

Cao, J. & Lu, S. (2019) Synthetic control inference for staggered adoption: estimating the dynamic effects of board gender diversity policies. *arXiv:1912.06320*.

Chernozhukov, V., Wüthrich, K. & Zhu, Y. (2021) An exact and robust conformal inference method for counterfactual and synthetic controls [just-accepted]. *Journal of the American Statistical Association*, 1–44. In press. https://doi.org/10.1080/01621459.2021.1920957

Donohue, J.J., Aneja, A. & Weber, K.D. (2019) Right-to-carry laws and violent crime: a comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies*, 16(2), 198–247.

Doudchenko, N. & Imbens, G.W. (2017) Difference-in-differences and synthetic control methods: a synthesis. *arxiv 1610.07748*.

Dube, A. & Zipperer, B. (2015) Pooling multiple case studies using synthetic controls: an application to minimum wage policies.

Ferman, B. & Pinto C. (2021) Synthetic controls with imperfect pre-treatment fit. *Quantitative Economics*.

Fesler, L. & Pender, M. (2019) Local promise programs: varying impacts on enrollment, graduation, and financial outcomes.

Frandsen, B.R. (2016) The effects of collective bargaining rights on public employee compensation: evidence from teachers, firefighters, and police. *ILR Review*, 69(1), 84–112.

Goldstein, D. (2015) *The teacher wars: a history of America's most embattled profession*. Mumbai: Anchor.

Goodman-Bacon, A. (2021) Difference-in-differences with variation in treatment timing. *Journal of Econometrics*. In press. https://doi.org/10.1016/j.jeconom.2021.03.014

Hess, F.M. & West, M.R. (2006) A better bargain: overhauling teacher collective bargaining for the 21st century. *Program on Education Policy and Governance, Harvard University*.

Hoxby, C.M. (1996) How teachers' unions affect education production. *The Quarterly Journal of Economics*, 111(3), 671–718.

Imai, K. & Kim, I.S. (2021) On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3), 405–415.

Imai, K., Kim, I.S. & Wang, E. (2019) Matching methods for causal inference with time-series cross-section data.

Jackson, C.K., Rockoff, J.E. & Staiger, D.O. (2014) Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1), 801–825.

Kahn-Lang, A. & Lang, K. (2019) The promise and pitfalls of differences-in-differences: reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38(3), 613–620.

King, G., Lucas, C. & Nielsen, R.A. (2017) The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, 61(2), 473–489.

Kreif, N., Grieve, R., Hangartner, D., Turner, A.J., Nikolova, S. & Sutton, M. (2016) Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, 25(12), 1514–1528.

Lovenheim, M.F. (2009) The effect of teachers' unions on education production: evidence from union election certifications in three midwestern states. *Journal of Labor Economics*, 27(4), 525–587.

Neyman, J. (1990 [1923]) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472.

Otsu, T. & Rai, Y. (2017) Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112, 1720–1732.

Paglayan, A.S. (2019) Public-sector unions and the size of government. *American Journal of Political Science*, 63(1), 21–36.

Pimentel, S.D. & Kelz R.R. (2020) Optimal tradeoffs in matched designs comparing US-trained and internationally trained surgeons. *Journal of the American Statistical Association*, 115(532), 1675–1688.

Robbins, M., Saunders, J. & Kilmer, B. (2017) A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *Journal of the American Statistical Association*, 112(517), 109–126.

Roth, J. (2018) Should we condition on the test for pre-trends in difference-in-difference designs?*arXiv preprint arXiv:1804.01208*.

Roth, J. & Sant'Anna, P.H. (2021) Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.

Rubin, D.B. (1980) Comment on "randomization analysis of experimental data: the fisher randomization test". *Journal of the American Statistical Association*, 75(371), 591–593.

Sun, L. & Abraham, S. (2020) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*. In press. https://doi.org/10.1016/j.jeconom.2020.09.006

Shaikh, A.M. & Toulis, P. (2021) Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 1–14. http://dx.doi.org/10.1080/01621459.2021.1974458

U.S. Department of Education, National Center for Education Statistics. (2018) Fast facts: expenditures. Technical report.

Xu, Y. (2017) Generalized synthetic control method: causal inference with interactive fixed effects models. *Political Analysis*, 25, 57–76.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Ben-Michael, E., Feller, A., & Rothstein, J. (2021) Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 351–381. https://doi.org/10.1111/rssb.12448