NBER WORKING PAPER SERIES

STACKED DIFFERENCE-IN-DIFFERENCES

Coady Wing
Seth M. Freedman
Alex Hollingsworth

Working Paper 32054
http://www.nber.org/papers/w32054

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2024

We thank Kosali Simon, Dan Sacks, Sunny Karim, and Matt Webb for feedback on early versions of the paper. We are also grateful to Austin Nichols, Adam Looney, Enrique Pinzon, Vivian Wong, Peter Steiner, and participants at the WHY Utah conference for providing a venue for early feedback and for many helpful comments. Madeline Yozwiak, Patrick Carlin, and Mallory Dreyer provided excellent research assistance. You can find example code to calculate and implement these weights here: https://github.com/hollina/stacked-did-weights The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Stacked Difference-in-Differences
Coady Wing, Seth M. Freedman, and Alex Hollingsworth
NBER Working Paper No. 32054
January 2024
JEL No. C01,C13,I0

ABSTRACT

This paper introduces the concept of a "trimmed aggregate ATT," which is a weighted average of a set of group-time average treatment effect on the treated (ATT) parameters identified in a staggered adoption difference-in-differences (DID) design. The set of identified group-time ATTs that contribute to the aggregate is trimmed to achieve compositional balance across an event window, ensuring that comparisons of the aggregate parameter over event time reveal dynamic treatment effects and differential pre-trends rather than compositional changes. Taking the trimmed aggregate ATT as a target parameter, we investigate the performance of stacked DID estimators. We show that the most basic stacked estimator does not identify the target aggregate or any other average causal effect because it applies different implicit weights to treatment and control trends. The bias can be eliminated using corrective sample weights. We present a weighted stacked DID estimator, and show that it correctly identifies the target aggregate, providing justification for using the estimator in applied work.

Coady Wing
Indiana University
1315 E 10th St
Bloomington, IN 47405
cwing@indiana.edu

Seth M. Freedman
School of Public and Environmental Affairs
1315 E. 10th St.
Bloomington, IN 47405
freedmas@indiana.edu

Alex Hollingsworth
Department of Agricultural, Environmental,
and Development Economics,
Department of Economics, and
the John Glenn College of Public Affairs
The Ohio State University
2120 Fyffe Road
Columbus, OH 43210
and NBER
hollingsworth.126@osu.edu

A code repository is available at https://github.com/hollina/stacked-did-weights

1 Introduction

In a staggered adoption difference-in-differences (DID) design, units are exposed to treatments at varying times. Until recently, two-way fixed effects (TWFE) regressions were the standard method of estimating causal effects in these designs. However, recent studies by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfoeuille (2020) have revealed threats to validity in the TWFE approach. The main problem is that the within-variation that identifies the TWFE coefficients includes comparisons between late and early treatment adopters. These comparisons may violate the common trend assumption unless treatment effects are constant over time. The upshot is that the conventional TWFE estimator does not identify a well-defined average causal effect in the staggered adoption setting, at least under the standard DID assumptions. In response, several new analytic methods have emerged to support causal inference in staggered adoption designs (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Dube et al., 2023; Gardner, 2022; Wooldridge, 2021).

The *stacked DID* is one approach to analyzing staggered adoption designs (Cengiz et al., 2019; Deshpande and Li, 2019; Butters et al., 2022; Callison and Kaestner, 2014). In stacked DID, researchers construct a separate data set for each *valid* sub-experiment, excluding the problematic late-early comparisons. These sub-experimental data sets are vertically concatenated to form a stacked analytic file. The goal is to estimate an average causal effect by fitting DID or event study regressions to the stacked dataset. Despite its appeal, the existing literature has not worked out the precise parameter estimated by stacked DID or determined whether stacked regressions have a causal interpretation.

In this paper, we clarify the causal estimand of alternative stacked DID estimators in settings where treatment effects may be heterogeneous across units and time periods. We state inclusion criteria for building a stacked data set that is trimmed to ensure balance in the number of pre- and post-periods for each sub-experiment. We then present a new method of estimating an aggregate average treatment effect on the treated (ATT) parameter using a single stacked regression that allows for conventional approaches to statistical inference. In event study form, our approach provides evidence related to both pre-trends and dynamic treatment effects. The weighted stacked regression we propose identifies an aggregate causal parameter that we call the trimmed aggregate ATT:

$$\theta_{\kappa}^{e} = \sum_{a \in \Omega_{\kappa}} ATT(a, a+e) \times \frac{N_{a}^{D}}{N_{\Omega_{\kappa}}^{D}}.$$

In the expression, $ATT(a, a+e)$ represents the average causal effect of adopting treatment in period a on outcomes experienced in period $a+e$, among units who are first treated in period a. Ω_κ is a trimmed set of treatment adoption events that *excludes* any adoption event where the group-time ATT is not identified for each event time period running from κ_{pre} periods before treatment to κ_{post} periods after treatment. N_a^D is the number of units that adopt treatment in period a, and $N_{\Omega_\kappa}^D$ is the number of units that ever adopt treatment in one of the adoption events in the trimmed set. The aggregate parameter, θ_κ^e, is a weighted average of group-time ATTs, with each group-time $ATT(a, a+e)$ weighted by the fraction of all trimmed treated units that adopt in period a. In other words, the aggregate is a kind of overall ATT.[1]

The trimmed aggregate ATT has three attributes that make it a good way to summarize results from a staggered adoption design. First, θ_κ^e is a coherent "causal aggregate" because it is a convex combination of underlying causal effects. Second, because the measure is computed from a trimmed set of identified group-time ATTs, changes in the aggregate parameter across event time periods reflect treatment effect dynamics, rather than compositional changes. Third, under the DID assumptions, pseudo-ATT effects computed in the pre-treatment periods should equal zero. Since the composition of the stacked data is stable over event time, the value of the aggregate in pre-treatment periods measures differential pre-trends, which would indicate non-common trends, anticipation, or both in one or more sub-experiments.[2]

We take θ_κ^e to be a target parameter of interest, and we examine the performance of *stacked DID regressions* as estimators of the target parameter using data from a staggered adoption design. We start by analyzing a basic stacked event study regression that is saturated in both event time and treatment status and is fit to stacked data created using clear inclusion criteria. We show that the coefficients from the basic stacked event study regression *do not* correspond to our target parameter or to any other convex combination of causal effects. Even when the DID assumptions hold within each sub-experiment, these stacked regressions are biased because treatment and control trends are implicitly weighted differently across sub-experiments. However, the bias is a function of known sub-experimental sample sizes. We derive sample weights to correct for the imbalance and show that a weighted stacked DID

[1] We use the ATT-weighting procedure as our primary example throughout the paper. But it is straightforward to use our methods to define different weighting procedures for any weights that are invariant across event time. For example, in some applications it may make sense to weight the group-time ATTs by the share of the overall analytic sample used to estimate the group-time ATT. These weights would depend on both the treatment group and control group share of the overall analytic sample rather than the treatment group share. Another option would be to weight by population size.

[2] θ_κ^e is similar to the balanced event time aggregate presented by Callaway and Sant'Anna (2021). The important difference is that our version enforces compositional balance in both a pre-treatment period and and post-treatment period.

estimator identifies the target aggregate. In practice, the effects can be estimated by fitting a saturated event study or a DID regression using weighted least squares. The coefficients from the weighted stacked regression correspond to the target aggregate, θ^e_κ.

The trimmed aggregate ATT (θ^e_κ) is a reasonable causal aggregate to focus on in applied work. However, there are other sensible ways to combine estimates from multiple sub-experiments. For example, two alternatives that seem logical are a population weighted aggregation and a sub-experiment sample size weighted aggregation. With some simple alterations to the corrective sample weights, we show that the weighted stacked event study can also be used to identify and estimate both of these parameters.

Because treatment varies at the group level in staggered adoption designs, it is customary to estimate standard errors using a cluster robust variance matrix and to treat observations as independent across groups but dependent within groups. The stacked data sets we consider in this paper will often include the same group in multiple sub-experiments and may contain duplicate observations if the same clean controls appear in multiple sub-experiments. Duplicate observations are an additional reason why the stacked data should be viewed as dependent across sub-experiments. We argue that it makes sense to cluster at the group level to allow for dependence across sub-experiments. However, some applications of stacked DID estimators report standard errors clustered at the $group \times sub-experiment$ level. We report results from a small Monte Carlo simulation study, examining the performance of both approaches. The results suggest both metrics have accurate coverage when the number of clusters is not too small.

Other versions of stacked DID estimators have been used in applied work by Cengiz et al. (2019), Deshpande and Li (2019), Callison and Kaestner (2014), and Butters et al. (2022). Our approach is closely related but not identical to these applications. Cengiz et al. (2019) and Deshpande and Li (2019) build the stacked data set using clean controls but do not enforce inclusion criteria that create compositional balance as we do. Butters et al. (2022) impose both clean controls and compositional balance. Callison and Kaestner (2014) use individual level survey data to build a stacked data set with two event times and a matched set of clean control states. To form the matched state control groups for each sub-experiment, they start with the set of clean controls and then discard candidate control states where the difference in mean baseline outcomes between the treated state and the control state was statistically significant. They estimate treatment effects using a stacked logistic regression model. Cengiz et al. (2019) report standard errors that cluster at the $group \times sub-experiment$ level, while Callison and Kaestner (2014), Deshpande and Li (2019) and Butters et al. (2022) cluster at the group level. None of these existing studies derives what parameter is identified by the stacked regression, or uses the weighting strategy we propose. Interestingly, none of these

papers uses a fully saturated stacked regression specification. Instead, all four applications use regression specifications that include $group \times sub-experiment$ and $time \times sub-experiment$ fixed effects to model the stacked data.

We examine the stacked fixed effect specification applied to a stacked data set created using our inclusion criteria. We show that the stacked fixed effect estimator does not identify our target aggregate parameter. It also does not – in general – identify any other convex combination of underlying causal effects. Estimating the stacked fixed effect model using weighted least squares and our proposed corrective weights resolves the problem and makes the fixed effects redundant. Thus, one contribution of our paper is to show that a simpler regression specification is all that is required to identify the trimmed aggregate ATT. However, we also show that the unweighted stacked fixed effect regression does recover the target aggregate in the special case where the treatment group sample share is fixed across sub-experiments.

Stacked estimation has several advantages for applied research. It is regression-based, making it easy to implement and explain to social scientists who are used to working with regressions. It focuses attention on the underlying research designs and it does not rely on ancillary assumptions about statistical modeling beyond the standard DID assumptions. The coefficients from the weighted stacked model correspond to a well-defined average of underlying group-time ATT parameters, which is a sensible rationale for using the method. In addition, the trimmed aggregate ATT parameter provides a summary that is not affected by compositional change over event time. This makes it suitable for assessing treatment effect dynamics in the post period, and for assessing the possibility of differential pre-trends in the pre-period. Finally, the stacked DID framework provides an intuitive platform for implementing more elaborate research designs. For instance, the common trend and no-anticipation assumptions may be more credible when applied to a set of clean controls that is matched on a vector of baseline covariates and outcomes. In principle, a matched comparison group can be formed for each sub-experiment, building on methods developed in Heckman et al. (1998); Callaway and Sant'Anna (2021); Callison and Kaestner (2014). The stacked framework makes it easy to compare estimates from each sub-experiment, which can help avoid treating the aggregate parameter as a black box.

2 Staggered Adoption Designs

Use $s = 1...S$ to index a collection of groups and $t = T_{min}...T_{max}$ to index calendar time periods. Treatment exposure occurs at the $group \times time$ level, and treatment remains in place until the end of the study period. Let A_s represent the calendar period when group

s is first exposed to treatment, and set $A_s = \infty$ for groups that never adopt treatment during the study period. $Y_{st}(0)$ represents the outcome group s would experience in calendar period t under a hypothetical scenario in which group s is never exposed to treatment. $Y_{st}(a)$ represents the outcome that group s would experience in calendar period t if the group was first exposed to treatment in calendar period a.

The causal effect of adopting treatment in period a compared to never adopting treatment is $\beta_{st}(a) = Y_{st}(a) - Y_{st}(0)$. The realized outcome is $Y_{st} = Y_{st}(0) + \sum_a \beta_{st}(a) \times 1(A_s = a)$. Most of the time, the object of interest is an average causal effect, such as the average treatment effect on the treated (ATT) evaluated at a particular calendar date. In our notation, this group-time ATT is written $ATT(a, a+e) = E[\beta_{s,a+e}(a)|A_s = a]$, where $e = t - a$ measures event time centered at the treatment adoption date. Thus $ATT(a, a+e)$ represents the average causal effect of adopting treatment in period a on outcomes experienced in period $t = a + e$ among groups that were first exposed to treatment in period $A_s = a$.

The staggered adoption design provides a way to identify these group-time ATT effects under two main assumptions:

Assumption 1. *No Anticipation: The average causal effect of adopting treatment in period a is equal to zero for all calendar periods prior to period a. This means that $ATT(a, a+e) = 0$ for all $e < 0$. Equivalently, no anticipation implies that for periods $t' < a$:*

$$E\left[Y_{s,t'}(a) - Y_{s,t'}(0)|A_s = a\right] = 0.$$

Assumption 2. *Common Trends: In the absence of treatment exposure, the average change across post-treatment time periods would be the same in the treatment group ($A_s = a$) and the comparison group ($A_s > a$). For post-treatment event periods $e >= 0$:*

$$E\left[Y_{s,a+e}(0) - Y_{s,a-1}(0)|A_s = a\right] = E\left[Y_{s,a+e}(0) - Y_{s,a-1}(0)|A_s > a + e\right]$$

Assumption 1 (no-anticipation) is a version of the *strict exogeneity* assumption, familiar from panel data models. The assumption could fail if treatment exposure occurs in response to volatility in the outcome variable, or if behavior changes due to expectations of future treatment. We state Assumption 2 (common trends) in terms of a comparison of a specified adoption group with all groups that have not yet adopted treatment, including both never treated groups and groups that adopt after the post-treatment period of interest, $A_s > a + e$.

In practice, researchers may choose to work with a specialized common trends assumption that only relies on a never treated comparison group.[3]

2.1 Identifying group-time ATTs

The staggered adoption design may identify multiple $ATT(a, a+e)$ parameters. The trick is to use the correct combination of periods and groups. To see the standard argument that a DID comparison of treated units to a clean comparison group identifies a given $ATT(a, a+e)$ parameter, write:

$$\begin{aligned}
DID_{a,e} &= E[Y_{s,a+e} - Y_{s,a-1}|A_s = a] - E[Y_{s,a+e} - Y_{s,a-1}|A_s > a+e] \\
&= E[Y_{s,a+e}(a) - Y_{s,a-1}(0)|A_s = a] - E[Y_{s,a+e}(0) - Y_{s,a-1}(0)|A_s > a+e] \\
&= E[Y_{s,a+e}(0) + \beta_{s,a+e}(a) - Y_{s,a-1}(0)|A_s = a] - E[Y_{s,a+e}(0) - Y_{s,a-1}(0)|A_s > a+e] \\
&= E[\beta_{s,a+e}(a)|A_s = a] + \left\{E[\Delta_s^{Y(0)}|A_s = a] - E[\Delta_s^{Y(0)}|A_s > a+e]\right\} \\
&= ATT(a, a+e)
\end{aligned}$$

The first equality gives the standard DID in observed outcomes, where the treatment group consists of all groups that adopt in period a, and the control group consists of groups that adopt after focal post-period $a+e$. The second line substitutes potential outcomes, and imposes Assumption 1 (no-anticipation), allowing $Y_{s,a-1}(a)$ to be replaced with $Y_{s,a-1}(0)$. The third line rewrites the expression to emphasize causal effects, and the fourth line collects terms to express treatment group specific time trends, using $\Delta_s^{Y(0)} = Y_{s,a+e}(a) - Y_{s,a-1}(0)$ to represent the change in untreated outcomes. The term in braces cancels under Assumption 2 (common trends). This shows that the simple DID identifies the group-time ATT.

In a staggered adoption design, one can swap in different treatment groups and control groups to identify group-time ATT parameters for each adoption group – simply change the value of a in the derivation above. Keeping the base period fixed at the last year before treatment exposure ($t = a-1$) and applying the method repeatedly for different choices of e traces out the treatment effect in event time for a given adoption group. Setting $e < 1$ and forming the DID leads to pseudo effects in the pre-period, which should equal 0 under the common trend and no-anticipation assumptions. Thus, the staggered adoption design will often identify a family of event time specific $ATT(a, a+e)$ parameters for each adoption

[3]We have used a set up with group and time period observations throughout the paper. But it is straightforward to extend the approach to cases with individual observations within each group by time cell – i.e. observations on individual i in state s in year t.

group. The main constraint on whether the effect is identified for a particular event time is how far the adoption event is from the earliest and latest calendar date in the available data.

2.2 Two group event studies

The two group event study is a useful special case that helps clarify more complicated situations. Suppose that Group $s = 1$ is never treated ($A_1 = \infty$) and group $s = 2$ is treated at time a ($A_2 = a$), with $T_{min} < a \leq T_{max}$. The pre-treatment period runs from $t = T_{min}...a - 1$ and the post-treatment period runs from $t = a...T_{max}$. Under the no anticipation and common trends assumptions, $ATT(a, a+e)$ is identified for each event time $e = t-a$ starting $e = T_{min}-a$ periods before treatment exposure and extending to to $e = T_{max} - a$ periods after exposure using the DID comparison $DID_{a,e} = E[Y_{2,a+e} - Y_{2,a-1}|A_s = a] - E[Y_{1,a+e} - Y_{1,a-1}|A_s = \infty]$. Note of course that $ATT(a, a + e)$ is normalized to zero for $e = -1$.

In practice, it is convenient to estimate the event study using a saturated linear regression that traces out the conditional expectation function of realized outcomes with respect to event time and treatment group membership:

$$Y_{st} = \alpha_0 + \alpha_1 1[A_s = a] + \sum_{\substack{h=T_{min}-a \\ h\neq -1}}^{T_{max}-a} \Big[\beta_h \big(1[A_s = a] \cdot 1[t - a = h]\big) + \delta_h 1[t - a = h]\Big] + U_{st} \quad (1)$$

The reference group in the specification is the control group in period $a - 1$, which is the period immediately before treatment. The model is parameterized so that the coefficients on the treatment $\times$ event time interaction terms are DID comparisons at different follow up times, each relative to the same reference period. This means that each of the β_h coefficients identifies a causal effect – $\beta_h = ATT(a, a + h)$ – just as if we had computed each DID comparison using the relevant group by period means rather than a regression. In applied work, it is common practice to plot the pre-period and post-period event study coefficients along with confidence intervals. Under the null hypothesis implied by the identifying assumptions, the collection of pre-period coefficients should be equal to zero, and the post-period coefficients will trace out the pattern of time varying treatment effects.

3 Trimming for Compositional Balance

The staggered adoption design is a collection of two group event studies. The main difference is that in the staggered adoption design the mapping between calendar time and event time is

not one-to-one across the different sub-experiments. For example, in a group that first adopts treatment in 1997, the year 2000 is three years post-treatment. In contrast, for a group that adopts treatment in 2005, the year 2000 is five years *pre-treatment*. Calendar time and event time are separate concepts from the perspective of the overall staggered adoption design.

Another issue is that causal effects may be identified for a larger number of event time periods for some adoption groups than others. For instance, suppose that group $A_s = a_1$ is first exposed only one year before the most recent year of data available so that $a_1 = T_{max} - 1$. The treatment effect for group a_1 measured three years after adoption, $ATT(a_1, a_1 + 3) = ATT(T_{max} - 1, T_{max} - 1 + 3)$, is not identified because there is no data available past T_{max} to measure the effect.[4] In contrast, suppose group a_2 first adopted the treatment ten years before the final year of data. Then $ATT(a_2, a_2 + e)$ is identified for every $e = 0 \ldots 10$ provided there are some never treated groups to serve as controls. The same problem arises in the pre-period: pre-treatment pseudo ATT effects may be identified for more periods in some sub-experiments than others.

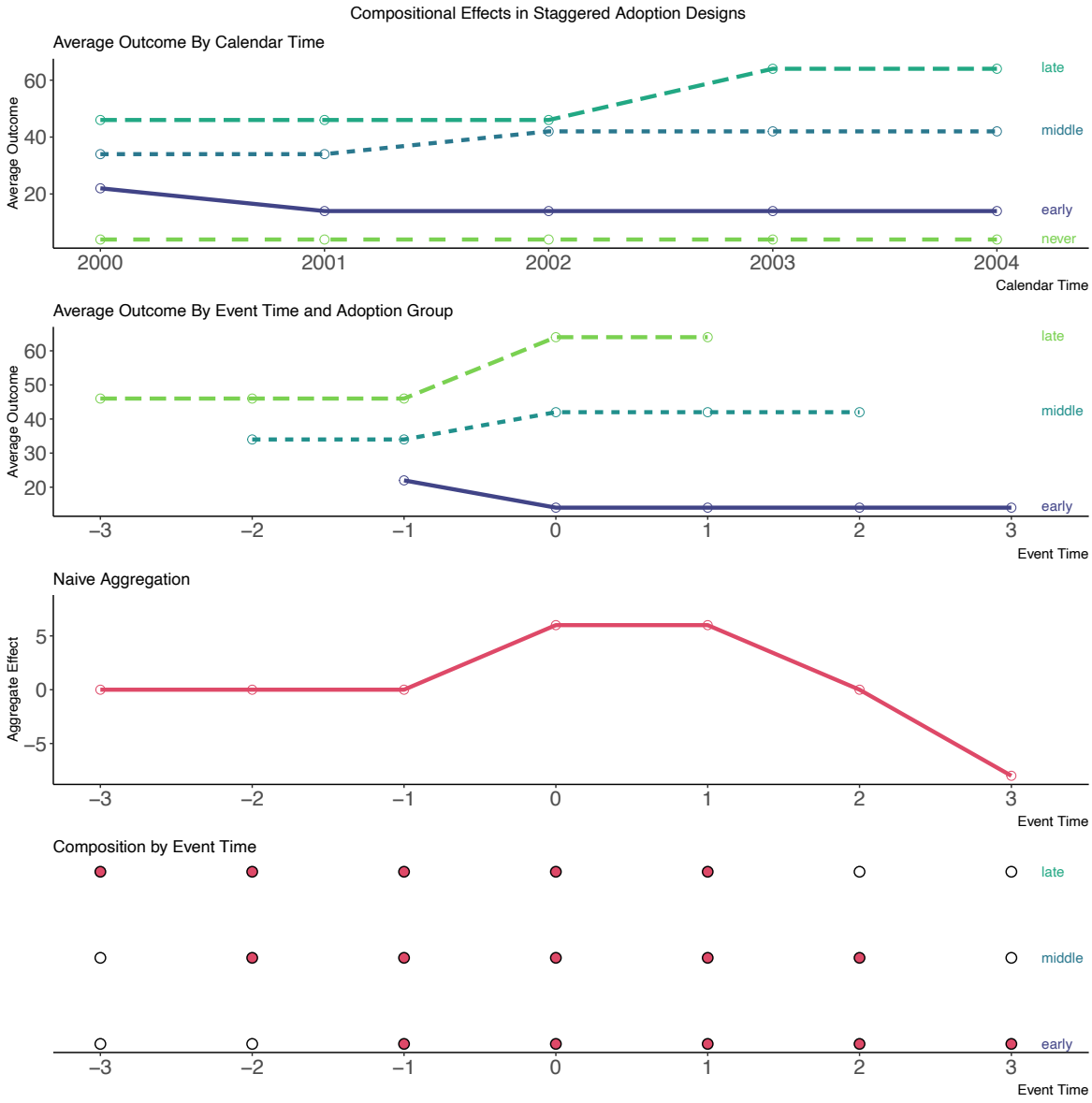
The collection of group-time ATT effects identified by a staggered adoption design can be unwieldy, especially when there is a large number of treatment adoption events. A natural impulse is to average the effects into some type of aggregate summary (Callaway and Sant'Anna, 2021). There are many ways to form an aggregate summary and no one measure will make sense for every occasion. In this paper, we focus on how an aggregate parameter might be useful for answering two types of empirical questions that commonly arise in applied work. First, how do treatment effects evolve over time since adoption? Second, what evidence exists for or against the validity of the no-anticipation and common trend assumptions? In the two group event study case, the event study regression coefficients – presented graphically, with confidence intervals – are a useful way for assessing evidence relevant to both questions.

One challenge with aggregating group-time ATTs is ensuring that the aggregation procedure preserves the ability to interpret event study results as evidence on differential pre-trends and dynamic causal effects in the post-period. The main threat is that the composition of the identified group-time ATTs being averaged together changes over event time periods. If the collection of group-time ATTs that are being combined changes over event time and treatment effects are heterogeneous, a naive event time plot can be misleading.

Figure 1 shows a stylized example of a staggered adoption design with three adoption groups (early, middle, late) and a never treated comparison group. The first panel shows the time series of observed outcomes in each group over a calendar time period running from

[4]This might happen in practice because the data is complete up to the present calendar date and so three years post treatment has yet to occur. In that case, it may eventually be possible to identify and estimate the effect as new data becomes available. In other circumstances, data are only available up to period T_{max} and no future data collection is possible. In that case, the effect will remain unidentified with available data.

Figure 1: The dangers of compositional changes in an aggregate event study in a staggered adoption design.



2000 to 2004. Each group has a different baseline level and the time trend is flat. The early adoption group receives a negative treatment effect when it adopts in 2001. The middle group adopts in 2002 and has a small positive treatment effect. And the late adoption group has a large treatment effect after it adopts in 2003. Thus, in this example, treatment effects are heterogeneous across groups but they do not change over time within groups. There are no "dynamic treatment effects".

The second panel shows average outcomes in the three adoption groups after centering each group around its adoption year to put the x-axis on event time rather than calendar

time. The event time graph makes it clear that treatment effects are identified for a different range of event times in each group. The early adoption group can be followed four years after adoption. In contrast, the middle group can be followed three years after adoption and the late group can only be followed for one year post adoption.

The third panel shows a naive aggregation of treatment effects in event time. These are simple averages of the identified causal effects in each event time period. But the set of effects included in the average changes over event time. The fourth panel highlights which of the three treated units is included in the average.In event times 0 and 1, the average is taken over all three groups for a net positive effect. In event time 2 only the early and middle groups are included in the average. And in event time 3 only the early group is included. Even though treatment effects are time invariant in this example, the naive aggregation strategy makes it appear as though the treatment effect is initially large and then fades out over time. In contrast, the compositional changes do not lead to misleading results in the pre-periods in this example because there is no heterogeneity across groups during the pre-period.

To build an aggregate that avoids the kind of problem shown in Figure 1, we develop inclusion criteria and use them to construct a trimmed set of sub-experiments that is balanced over a fixed event time window. The starting point is to define a uniform event window for the analysis. Use κ_{pre} and κ_{post} to represent the desired length of the pre- and post-treatment period. Then let Ω_A be the set of unique policy adoption dates contained in the staggered adoption design. Ω_κ is the trimmed subset of adoption events for which the ATT is identified for each event time in the κ window. Membership in Ω_κ is determined by two inclusion criteria.

IC 1. *Adoption Event Window:* The treatment adoption event $a \in \Omega_A$ must occur inside the κ event window. Let $IC1_a = 1[T_{min} + \kappa_{pre} + 1 \leq a \leq T_{max} - \kappa_{post} - 1]$ be an indicator variable set to 1 if adoption event a occurs inside the event window.

IC 2. *Existence of Clean Controls:* There must be one or more "clean control" units that can serve as a comparison group in the DID analysis for a treatment adoption event $a \in \Omega_A$. Specifically, let $IC2_a = 1[\sum_s 1(A_s > a + \kappa_{post}) \geq 1]$ be an indicator variable set to 1 if there are any clean controls available for adoption event a.

The trimmed set of adoption events is $\Omega_\kappa = \{a \in \Omega_A | IC1_a = IC2_a = 1\}$. Under the common trend and no-anticipation assumptions, the staggered adoption design identifies $ATT(a, a+e)$ for each $a \in \Omega_\kappa$ for each event time $e \in \{-\kappa_{pre}...\kappa_{post} | e \neq -1\}$, where the event time period immediately before treatment adoption serves as a fixed reference period in the DID comparisons.

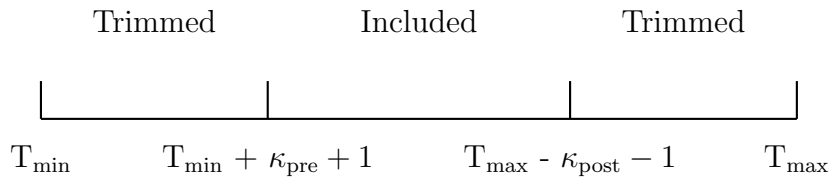


Figure 2: Admissible events with a κ event window

The κ parameters are a research design choice with practical implications because they control which adoption events are "trimmed" in order to ensure compositional balance. Figure 2 shows how the κ window determines which policy changes are admissible under the first inclusion criteria. The full set of treatment adoption events $a \in \Omega_A$ may occur anywhere between T_{min} and T_{max} in the diagram. However, any adoption event that occurs between before $T_{min} + \kappa_{pre} + 1$ occurs too close to the beginning of available data to identify an effect κ_{pre} periods in advance of treatment. Likewise, any event that occurs after $T_{max} - \kappa_{post} - 1$ occurs too late to be studied for a full κ_{post} periods after treatment. These events are trimmed. A shorter event time window may allow more policy events to be studied. A longer window allows dynamic treatment effects to be studied for a longer period, perhaps for a smaller subset of adoption events. Setting $\kappa_{pre} = \kappa_{post} = 0$ is the least restrictive because in that case any event that happens at least one period after T_{min} and one period before T_{max} is potentially valid.

The clean controls inclusion criteria we use here is based around the idea that a clean control is one that is not exposed to treatment at any time during the κ event window. But it is straightforward to implement alternative definitions of clean controls. In some cases, researchers may prefer to define clean controls as units that never adopt treatment at all, at least up to the time of publication. In other cases, it might be more credible to define clean controls as states that do adopt treatment at some future date but do not adopt during the κ window. This would exclude the never treated states and use only the not yet treated states as clean controls. Another possibility is to allow both never treated and not-yet treated states but require a more stringent criteria such that clean a controls must have $A_s > a + \kappa_{post} + \kappa_{pre}$. This would ensure that the clean controls are not even in their own pre-period if they are being used as clean controls for event a. There is no single correct set of inclusion criteria that is good for all settings. These are study design choices and the definition of what constitutes a clean control that is likely to meet the common trend and no-anticipation assumption may vary from one study to the next. However, these alternative inclusion criteria do not change the basic ideas of the stacked DID approach.

4 Aggregation and Target Parameters

With the trimmed set of adoption events in hand, we have a collection of group-time ATT parameters that are identified based on the observable data and the common trend and no-anticipation assumptions. Specifically, in each event time period from $e = -\kappa_{pre}...\kappa_{post}$ we can identify $ATT(a, a+e)$ for each $a \in \Omega_\kappa$. In practice, it will often be convenient to combine the collection of estimates into a summary average. One approach is to form a weighted average of group-time ATTs for a specific event time, weighting each group-time ATT by its share of the trimmed treatment group:

$$\theta^e_\kappa = \sum_{a \in \Omega_\kappa} ATT(a, a+e) \times \frac{N^D_a}{N^D_{\Omega_\kappa}} \tag{2}$$

We call θ^e_κ the trimmed aggregate ATT because it is an average of group-time ATTs across the sub-experiments contained in the trimmed set. N^D_a is the number of treated units in sub-experiment a, and $N^D_{\Omega_\kappa} = \sum_{a \in \Omega_\kappa} N^D_a$ is the total number of treated units in the trimmed set. The weights are non-negative and sum to one across the sub-experiments, which means that the aggregate parameter is a convex combination of underlying causal effects. The collection of groups included in the average and the weight assigned to each component ATT parameter is fixed across event time periods, ensuring that the composition of the aggregate is balanced over event time. This means that changes in the value of θ^e_κ across event time periods indexed by e reflect dynamic treatment effects in the post-period and evidence of differential pre-trends in the pre-periods. To form a summary average across the post-treatment event window, one could compute the simple average of the treatment effects across the post period event times: $\theta^{post}_\kappa = \frac{1}{\kappa_{post}} \sum_{h=1}^{\kappa_{post}} \theta^h_\kappa$.

Table 1 presents three other trimmed aggregate parameters that might be useful in some applications. The first row shows the trimmed aggregate ATT discussed above. The second row presents a population weighted ATT, which weights each of the group-time ATTs by it's share of the overall treated population. In a study organized around *state* $\times$ *year* policy changes, the population weighted ATT give more weight to sub-experiments that affect a larger group of people. The third row shows an aggregate in which each group-time ATT is weighted by the share of the overall analytic sample involved in its estimation. This would give more weight to group-time ATTs estimated using a larger sample size, which depends on the size of the treatment group and the control group in that sub-experiment. The first column of Table 1 simply shows the definition of each of these three aggregate parameters of interest. We discuss how to estimate them using stacked weighted least squares regressions

Table 1: Alternative weighting schemes

	Estimand	Treatment Weight	Control Weight	Notes
Trimmed Aggregate ATT	$\sum_{\Omega_\kappa} ATT(a,e)\frac{N_a^D}{N^D}$	1	$\frac{N_a^D/N^D}{N_a^C/N^C}$	N_a^D is the number of groups that first adopt treatment in a. N^D is the total number of groups that adopt treatment at any of the times included in the trimmed set. N_a^C and N^C give the analogous counts for the control groups. These ATT weights produce an aggregate parameter in which each group time $ATT(a,e)$ in the trimmed set is weighted by its share of the treated sample.
Population Weighted ATT	$\sum_{\Omega_\kappa} ATT(a,e)\frac{Pop_a^D}{Pop^D}$	$\frac{Pop_a^D/Pop^D}{N_a^D/N^D}$	$\frac{Pop_a^D/Pop^D}{N_a^C/N^C}$	Pop_a^D represents the total population of people in group a. Pop^D is the total population across all of the treated groups included in the trimmed set. These weights produce an aggregate parameter in which each group time $ATT(a,e)$ in the trimmed set is weighted by its share of the treated population.
Sample Share Weighted ATT	$\sum_{\Omega_\kappa} ATT(a,e)\frac{(N_a^D+N_a^C)}{N^D+N^C}$	$\frac{(N_a^D+N_a^C)/(N^D+N^C)}{N_a^D/N^D}$	$\frac{(N_a^D+N_a^C)/(N^D+N^C)}{N_a^C/N^C}$	These weights produce an aggregate parameter in which each group time $ATT(a,e)$ is weighted by its overall share of the stacked analytic sample. The sample share depends on the number of both treated and control units in each subexperiment.

later in the paper.

5 Stacked DID Estimation

All of the component elements of the trimmed aggregate ATT parameter – θ_κ^e – are identified by a staggered adoption design. In principle, one could use the strategy mapped out by Callaway and Sant'Anna (2021) to estimate each of the group-time ATTs and then form the aggregate manually. [5] However, in this section of the paper, we show that it is also possible

[5]To compute the trimmed aggregate ATT using the manual approach, use the event window and clean controls inclusion criteria to define the trimmed set of adoption groups. Compute the weights given in the estimand column of Table 1; note that the weights here come from the estimand itself and not from the corrective sample weights. Then separately estimate the DID contrasts for each adoption group in each event time in the $\kappa-$window. Finally, multiply each of the group-time ATT estimates by the weight and then sum up the weighted ATT estimates from each event time period. All of this is conceptually compatible with the Callaway and Sant'Anna (2021) method. However, there is no automatic way to accomplish these steps

to estimate the aggregate parameter in one step using a linear regression estimator applied to a stack of sub-experimental data sets.

We start by explaining how to construct the analytic sample for each sub-experiment and assemble them into a stacked data set. Then we analyze several regression specifications to clarify the causal estimand of the stacked regressions under the standard DID assumptions. We show that the most basic stacked regression specification is biased in the sense that its coefficients do not correspond to our target aggregate parameter (θ^e_κ) or to any other convex combination of underlying causal effect parameters. The bias in the basic stacked regression arises because the regression implicitly weights treatment trends and control trends differently. This matters when the treatment share differs across sub-experiments. Because the bias is a function of relative sample sizes in the treatment and control groups, it is possible to correct the bias using sample weights. We derive sample weights to correct the bias and propose a weighted least squares stacked regression that is straightforward to implement. We show that the coefficients from the weighted stacked regressions recover the target aggregate parameter, θ^e_κ. We study statistical inference for the weighted stacked event study estimator using a small monte carlo analysis.

Most of this section is focused on unweighted and weighted stacked event study specifications that are *fully saturated* models of the conditional expectation function linking outcomes across event time and treatment status. These models are different from some of the specifications used in applications by Cengiz et al. (2019), Butters et al. (2022), and Deshpande and Li (2019). Those studies use regressions that include unit by time and unit by sub-experiment fixed effects and that are not fully saturated specifications. The final part of this section considers these stacked fixed effect specifications and shows that they are also biased by differential weights.

5.1 Building the Stack

The first step in implementing the stacked DID estimator is to assemble a separate data set for each sub-experiment in the trimmed set of adoption events. Start with a long-form panel in which each row is a unit $\times$ calendar time observation. Let $D^a_s = 1(A_s = a)$ be an indicator variable set to one if group s first adopts treatment in period a. Next, let $C^a_s = 1(A_s > a + \kappa_{post})$ be a dummy variable set to one if group s is a valid clean control for adoption event a. Finally, $M^a_t = 1(a - \kappa_{pre} - 1 \leq t \leq a + \kappa_{post})$ indicates that calendar time t falls inside the $\kappa-$window for sub-experiment a.

using the packages and commands developed for the Callaway and Sant'Anna (2021) method in R and Stata. To use these packages for our purposes, you would need to apply the inclusion criteria to trim the set of identified parameters, manually compute the weights, and the form the aggregate.

With these definitions in hand, $I_{st}^a = M_t^a(D_s^a + C_s^a)$ is a dummy variable set to one if observation s from calendar period t belongs to the analytic sample for sub-experiment a. The sub-experimental data set for sub-experiment $a \in \Omega_\kappa$ consists of all observations with $I_{st}^a = 1$. Repeat this procedure for each $a \in \Omega_\kappa$ to construct each of the sub-experimental data sets.

In sub-experiment a, there will be $N_a^D = \sum_s D_s^a$ treatment group units and $N_a^C = \sum_s C_s^a$ control group units. In each sub-experiment, each unit will be observed for $\kappa_{pre} + \kappa_{post} + 1$ time periods. Ordering each sub-experiment by event time $e = t - a$ and *vertically concatenating* the sub-experimental data sets leads to a single "stacked" data set. In the stacked data set, each observation refers to a unit $\times$ *sub-experiment* $\times$ event time (s, a, e) observation. Notationally, this means we have changed the $a-$superscripts in the sub-experimental data sets into $a-$subscripts in the stacked data set. Thus, we use Y_{sae} to represent the observed outcome for unit s in sub-experiment a in event time e. And we let D_{sa} be a dummy variable set to one if unit s is a member of the treatment group in sub-experiment a and set to zero if unit s is a member of the control group in sub-experiment a. Within a sub-experiment, treatment status does not vary over event time.

Importantly, in the stacked data set the event time index runs from $-\kappa_{pre} ... \kappa_{post}$ in every sub-experiment in the stack. The event times will correspond to different calendar times since $e = t - a$, which means that the sub-experiments are not aligned in calendar time. There will be $N_{\Omega_\kappa}^D = \sum_{a \in \Omega_\kappa} N_a^D$ treatment group units in each event time period. Likewise, the stacked dataset will have $N_{\Omega_\kappa}^C = \sum_{a \in \Omega_\kappa} N_a^C$ control group units in each event time period. The total number of control observations in the stacked data set can be quite large because the same *group* $\times$ *time* observation may appear in multiple sub-experiments.

5.2 Simple Stacked Regressions

Looking at the stacked data set described above, it is straightforward to define conventional sub-experiment specific DID comparisons as $DID_{a=j,e} = E[Y_{sa,e} - Y_{sa,-1}|D_{sa} = 1, a = j] - E[Y_{sa,e} - Y_{sa,-1}|D_{sa} = 0, a = j]$. There is one of these DID comparisons for each sub-experiment, and they can be constructed for different event times to trace out an event study. But the point of the stacked data set is to examine DID comparisons that pool information across sub-experiments. From that point of view, the most basic stacked DID comparison is $DID_e^{stack} = E[Y_{sa,e} - Y_{sa,-1}|D_{sa} = 1] - E[Y_{sa,e} - Y_{sa,-1}|D_{sa} = 0]$. Under the common trends and no-anticipation assumption, we know that for each sub-experiment DID $DID_{a=j,e} = ATT(j, j + e)$. But what about the stacked DID comparison?

To work out the connection between the stacked DID and the underlying sub-experiment

specific DIDs, we use the law of iterated expectations:

$$
\begin{aligned}
DID_e^{stack} &= E\Big[E[DID_a^e|a]\Big] \\
&= E\Big[E[Y_{sa,e} - Y_{sa,-1}|D_{sa}=1,a]\Big] - E\Big[E[Y_{sa,e} - Y_{sa,-1}|D_{sa}=0,a]\Big] \\
&= \sum_{j\in\Omega_\kappa} E[Y_{sa,e}(j) - Y_{sa,-1}(0)|D_{sa}=1, a=j]\tfrac{N_j^D}{N^D} \\
&\quad - \sum_{j\in\Omega_\kappa} E[Y_{sa,e}(0) - Y_{sa,-1}(0)|D_{sa}=0, a=j]\tfrac{N_j^C}{N^C} \\
&= \sum_{j\in\Omega_\kappa} E[\beta_{s,j+e}(j)|D_{sa}=1, a=j]\tfrac{N_j^D}{N^D} \\
&\quad + \sum_{j\in\Omega_\kappa} E[\Delta_s^{Y(0)}|D_{sa}=1, a=j]\tfrac{N_j^D}{N^D} - \sum_{j\in\Omega_\kappa} E[\Delta_s^{Y(0)}|D_{sa}=0, a=j]\tfrac{N_j^C}{N^C} \\
&= \theta_\kappa^e + \sum_{j\in\Omega_\kappa} E[\Delta_s^{Y(0)}|D_{sa}=1, a=j]\tfrac{N_j^D}{N^D} - \sum_{j\in\Omega_\kappa} E[\Delta_s^{Y(0)}|D_{sa}=0, a=j]\tfrac{N_j^C}{N^C} \\
&\neq \theta_\kappa^e
\end{aligned}
$$

The first equality describes the stacked DID as an iterated expectation over sub-experiment-specific DIDs. The second equality decomposes the DID into the pre-post changes in the treatment and control sub-populations, maintaining the iterated expectation for each. In the third equality, we substitute potential outcomes for observed outcomes, impose the no-anticipation assumption, and replace the outer expectations with a sum over sub-experiments. In the treated arm of the study, the j^{th} sub-experiment is weighted by $\frac{N_j^D}{N^D}$, which is the share of all of the treated observations in the stack that belong to sub-experiment j. Similarly, in the control group arm, the j^{th} sub-experiment is weighted by $\frac{N_j^C}{N^C}$, which is the share of all controls in the stack that appear in sub-experiment j. The fourth equality rewrites the expression in terms of causal effects and time trends in untreated outcomes, where we re-write pre-post changes using $\Delta_s^{Y(0)} = Y_{sa,e} - Y_{sa,-1}$ as shorthand. The weighted sum in the first term is the trimmed aggregate ATT, θ_κ^e, a substitution we make in the fifth equality.

The inequality in the final line shows that the stacked DID does not identify the target aggregate causal parameter or any other sensible causal aggregate. The problem is that the time trends in untreated potential outcomes are averaged across sub-experiments using different weights for treated observations and control observations in the stacked data set.

These weights are implicit: they are simply the way that the stacked DID combines information across the sub-experiments. But because of the different weights, the trends in untreated outcomes do not necessarily cancel in the stacked DID even if the common trends assumption holds within each sub-experiment.

5.3 Weighted Stacked Regressions

The simple stacked regression does not identify a causal effect because it weights treatment and control group trends differently. To correct for the bias, we define the following sample weight:

$$Q_{sa} = \begin{cases} 1 & \text{if } D_{sa} = 1 \\ \frac{N_a^D / N_{\Omega_\kappa}^D}{N_a^C / N_{\Omega_\kappa}^C} & \text{if } D_{sa} = 0 \end{cases}$$

The weighted stacked DID using Q_{sa} as a sample weight is:

$$\begin{aligned} DID_{ws,e} &= E\Big[E[Y_{sa,e}Q_{sa} - Y_{sa,-1}Q_{sa}|D_{sa}=1,a]\Big] - E\Big[E[Y_{sa,e}Q_{sa} - Y_{sa,-1}Q_{sa}|D_{sa}=0,a]\Big] \\ &= \sum_{j\in\Omega_\kappa} E[\beta_{s,j+e}(j)|D_{sa}=1,a=j]\tfrac{N_j^D}{N_{\Omega_\kappa}^D} + \sum_{j\in\Omega_\kappa} E[Y_{sa,e}(0) - Y_{sa,-1}(0)|D_{sa}=1,a=j]\tfrac{N_j^D}{N_{\Omega_\kappa}^D} \\ &\quad - \sum_{j\in\Omega_\kappa} E[Y_{sa,e}(0) - Y_{sa,-1}(0)|D_{sa}=0,a=j] \times \frac{N_j^D/N_{\Omega_\kappa}^D}{N_j^C/N_{\Omega_\kappa}^C} \times \frac{N_j^C}{N_{\Omega_\kappa}^C} \\ &= \theta_\kappa^e + \sum_{j\in\Omega_\kappa} \tfrac{N_j^D}{N_{\Omega_\kappa}^D}\left(E[\Delta_{sa}^{Y(0)}|D_{sa}=1,a=j] - E[\Delta_{sa}^{Y(0)}|D_{sa}=0,a=j]\right) \\ &= \theta_\kappa^e \end{aligned}$$

The first line shows the weighted stacked DID as an iterated expectation over the sub-experiments, this time with the outcome variables multiplied by the Q-weights. Conditional on $D_{sa} = 1$, the Q-weight is equal to 1 and so the weight vanishes in the treatment arm of the DID. In contrast, conditional on $D_{sa} = 0$, the weight is equal to $\frac{N_a^D/N_{\Omega_\kappa}^D}{N_a^C/N_{\Omega_\kappa}^C}$ in the control arm. The second equality substitutes potential outcomes, imposes the no-anticipation assumption, substitutes the known values of the Q-weights, and replaces the outer expectations with a sum over the sub-experiments. The control group term simplifies because $\frac{N_a^D/N^D}{N_a^C/N^C} \times \frac{N_j^C}{N^C} = \frac{N_j^D}{N^D}$. Using $\Delta_{sa}^e(0) = Y_{sa,e}(0) - Y_{sa,-1}(0)$ to represent trends and re-arranging, the third equality shows that the weighted stacked DID is equal to the target ATT aggregate plus the difference

in treatment and control group trends. Because of the Q-weights, the treatment group trends and control group trends in each sub-experiment are now weighted equally. These trends cancel under the common trends assumption so that the stacked DID identifies the target parameter.

The identical point estimate could be obtained using weighted least squares regressions fitted to the stacked data set. To see this, start with a simple case in which $\kappa_{pre} = \kappa_{post} = 0$. That leads to a stacked data set in which each sub-experiment has two periods corresponding to event times $e = -1$ and $e = 0$. In this case, the stacked DID regression specification would would have the familiar form:

$$Y_{sae} = \beta_0 + \beta_1 D_{sa} + \beta_2 1(e = 0) + \beta_3 D_{sa} \times 1(e = 0) + U_{sae}$$

Using weighted least squares with the Q-weights described above, $\beta_3 = \theta^0_\kappa$. This is the average of the group-time ATTs measured at the initial adoption year across all sub-experiments where the initial adoption year is identified. In the more expansive event study case, there is a longer event window and the classic DID regression does not suffice. For a feasible choice of κ_{pre} and κ_{post} the weighted event study regression specification is:

$$Y_{sae} = \alpha_0 + \alpha_1 D_{sa} + \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\lambda_e 1[e = h] + \delta_e D_{sa} \times 1[e = h] \right] + U_{sae} \tag{3}$$

When the model is estimated using weighted least squares with the Q-weights, $\delta_e = \theta^e_\kappa$.[6] For $\kappa_{pre} > 0$, the specification includes estimated effects in the pre-treatment time periods. When the no anticipation and common trend assumptions hold in each sub-experiment, these pre-treatment pseudo ATT effects will equal zero. Similarly, for $\kappa_{post} > 0$ the event study traces out the aggregated ATTs over the post treatment event times for a balanced set of adoption groups. Changes in $\delta_e = \theta^e_\kappa$ over post-treatment periods $e = 0, 1, ..., \kappa_{post}$ measure time varying treatment effects without concerns about changes in composition.

The full event study regression is a convenient way to measure evidence of differential

[6]The coefficients in the expression are defined as the solution to

$$\underset{\alpha_0,\alpha_1,\lambda_e,\delta_e}{\arg\min} \sum_{sae} Q_{sa} \left(Y_{sae} - \alpha_0 - \alpha_1 D_{sa} - \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\lambda_e 1[e = h] + \delta_e D_{sa} \times 1[e = h] \right] \right)^2 .$$

trends in the pre period and to estimate time varying treatment effects over the post-period. However, in applied work, it may also be appealing to estimate a single summary measure the average effect over the post-treatment time periods. For example, it might be helpful to estimate a quantity like $\theta_{\kappa}^{post} = \frac{1}{\kappa_{post}} \sum_{h=1}^{\kappa_{post}} \theta_{\kappa}^{h}$. A point estimate of θ_{κ}^{post} is the simple average of the post-period event study coefficients: $\delta^{post} = \frac{1}{\kappa_{post}} \sum_{e=0}^{\kappa_{post}} \delta_e$. In practice, it is easy to estimate standard errors for the aggregate can be computed using standard methods for a linear combination of coefficients.[7] Alternatively, the average post-period effect can be estimated directly using a regression model with a different parameterization:

$$Y_{sae} = \alpha_0 + \alpha_1 D_{sa} + \alpha_3 1[e \geq 0] + \delta^{post} D_{sa} 1[e \geq 0]$$
$$+ \sum_{h=-\kappa_{pre}{}^{-2}} \left[\gamma_e 1[e = h] + \beta_e D_{sa} \times 1[e = h] \right] + U_{sae}$$

In this model, event time indicators and the interaction with treatment group membership are included for each pre-treatment time period. But the post-period event time indicators are replaced with a single indicator set to 1 for all post-treatment periods. The coefficient on the interaction between treatment status and the post variable is exactly equal to the simple average of the underlying event study coefficients from the event study specification. Estimating the model this way using the trimmed sample and the corrective weights provides a direct estimate of the post period average effect and its standard error. The linear combination approach and the re-parameterized regression approach will both produce the same point estimate and standard errors. But note that these methods are slightly different from estimating a summary parameter by using just $treat_{sa}$, $post_{ae}$, and $treat_{sa} \times post_{ae}$ as regressors. Grouping the multiple pre- and post-periods into blocks to form the average effect will produce a different point estimate of the post-period average effect because it uses the full pre-period average as the baseline outcome rather than only the period just prior to treatment.

The weighting procedure described in this section can be modified so that the stacked regression coefficients identify some alternative aggregate parameters of interest. Table 1 shows the definition of $Q-$ weights required to uncover the population weighted ATT and the sample share weighted ATT. All of the information required to implement the sample share weights is contained in the stacked data set itself. However, to implement the population weighted ATT, researchers will need to assemble information about the population size in

[7] For example, one can use the marginaleffects package in R, or the lincom command in stata.

each of the units in the study sample.

Estimating these weighted stacked event study regressions is as easy as fitting event study regressions in the two group event study case. The main tasks from a programming point of view involve building the sub-experimental data sets, stacking them up, and computing the Q-weights using information on sample sizes by sub-experiment and treatment status.

5.4 Statistical Inference

In conventional approaches to staggered adoption designs and other settings where treatment varies at the group level, the data are usually treated as independent across groups and dependent within groups. It is important to allow for this dependence – i.e. clustering at the group level – when estimating standard errors (Bertrand et al., 2004). In the stacked design, the same observations may appear in multiple sub-experimental data sets because clean controls may be used in several sub-experiments. For instance, there may be some never treated groups that qualify as clean controls for every sub-experiment. In addition, it is possible for some units to appear as a clean control in some sub-experiments and as a treated unit in other sub-experiments. This could happen, for example, if a group is a very late adopter and the κ window is relatively short. These issues do not pose an identification problem. However, duplicate clean control observations across sub-experiments, and repeated use of the same groups across sub-experiments may create additional dependence across sub-experiments.

One way to conduct inference is to make the standard clustering assumption: assume that all observations from the same unit may be dependent, even if they appear in different sub-experiments. An alternative approach is to allow for clustering at the $group \times sub-experiment$ level. This approach allows for dependence among observations within the same group and sub-experiment but treats the observations on the same group in different sub-experiments as independent. In applied work, Cengiz et al. (2019) estimate standard errors allowing for clustering at the $group \times sub-experiment$ level, and Deshpande and Li (2019) and Butters et al. (2022) cluster at the group level.

We conducted a small Monte Carlo study to examine the performance of these two cluster robust variance matrices for performing statistical inference in a stacked event study. Our simulation is built around data from the Medicare Geographic Variation (MGV) Public Use File, which provides information on Fee for Service Medicare Expenditures per capita at the county level. We limit the sample to a balanced panel of $2,724$ counties that are each observed for the 15 years from 2007 to 2021. We treat these data as a sampling frame for our simulations.

Table 2: Statistical inference in stacked DID – Monte Carlo simulation results

Event Time	Number of Clusters	True $\theta(e)$	Average $\hat{\theta}$	SD $\hat{\theta}$	Average $\widehat{se_{county}}$	Average $\widehat{se_{county-sub}}$	Rejection $\widehat{se_{county}}$	Rejection $\widehat{se_{county-sub}}$
0	50	112	116.14	148.62	138.28	137.77	0.07	0.07
0	100	112	115.81	104.94	101.58	101.47	0.05	0.05
0	500	112	111.65	47.75	47.22	47.26	0.05	0.06
0	1000	112	112.74	33.03	33.50	33.55	0.05	0.05
0	1500	112	112.49	26.97	27.41	27.45	0.05	0.04
0	2000	112	111.35	23.37	23.72	23.75	0.05	0.05
0	2500	112	112.25	20.53	21.24	21.27	0.04	0.04
1	50	132	137.24	178.66	164.33	165.86	0.08	0.07
1	100	132	134.36	122.32	120.28	121.55	0.06	0.06
1	500	132	133.03	55.41	55.74	56.37	0.05	0.04
1	1000	132	133.20	39.33	39.58	40.04	0.05	0.05
1	1500	132	132.20	32.34	32.42	32.79	0.05	0.05
1	2000	132	131.51	27.85	28.08	28.41	0.05	0.05
1	2500	132	132.14	25.06	25.14	25.43	0.05	0.05
2	50	152	156.67	205.69	194.28	191.79	0.06	0.07
2	100	152	153.50	139.68	140.59	139.25	0.06	0.06
2	500	152	152.98	63.12	65.10	64.68	0.04	0.04
2	1000	152	153.53	44.71	46.16	45.88	0.04	0.04
2	1500	152	151.62	37.18	37.75	37.52	0.05	0.05
2	2000	152	151.61	31.82	32.73	32.54	0.04	0.05
2	2500	152	152.15	28.52	29.28	29.11	0.04	0.04

In each run of the Monte Carlo simulation, we draw a random sample of G counties with replacement from the MGV file. Then we assign the selected county-year observations to treatment and control status using a simple staggered adoption design. In our design, 18% of the G counties are randomly assigned to be ever treated. Then these $.18 \times G = N_D$ counties are randomly apportioned to three timing groups: $5/9 \times N_D$ are exposed to treatment in calendar year 2011, $3/9 \times N_D$ are exposed in 2013, and $1/9 \times N_D$ are exposed in 2015. We set the untreated potential outcome in each county-year cell equal to the observed Medicare Expenditures per capita in that cell. The treated potential outcome is the sum of the untreated outcome and the time varying treatment effect for each treatment adoption group. In our design, the treatment effect function is \$107 + \$20 $\times$ Years Since Adoption for the 2011 adoption group, \$119 + \$20 $\times$ Years Since Adoption for the 2013 adoption group, and \$116 + \$20 $\times$ Years Since Adoption for the 2015 adoption group. We consider a design with $\kappa_{pre} = 2$ and $\kappa_{post} = 2$, so that we estimate the trimmed aggregate ATT at $e = 0, 1, 2$. Given the size of these three adoption group, this means that the target aggregate parameters of interest in the DGP are $\theta^0 = \$112$, $\theta^1 = \$132$, and $\theta^2 = \$152$.

In the simulation, a cluster is a 15 period vector of Medicare expenditure for a selected county so that the sampling process preserves the actual dependency between observations within a county. The level of the outcome and changes over time within a county are also realistic because they come from whatever happened to Medicare spending in the selected county. Because treatment timing is randomly assigned, the common trend assumption holds in the DGP.

With a random sample of counties in hand, we form the sub-experiments, stack the data, and fit the weighted stacked event study regression. We estimate two sets of cluster robust standard errors for the coefficients: standard errors that allow for dependence at the pseudo-county level, and standard errors that allow for dependence as the pseudo-county $\times$ sub-experiment level. Then we compute the cluster robust t-statistic for tests of the null hypothesis that each post period treatment effect is equal to its known true value. If the clustered standard errors are a good approximation to the sampling error, then a two-tailed test with $\alpha = .05$ should reject the null (wrongly) in 5% of simulations.

We conducted 5000 simulations with $G \in [50, 100, 500, 1000, 2000, 2500]$ clusters. Table 2 reports the results from the simulation. The first panel shows estimates of the immediate effect at period $e = 0$. The second and third panels show results for event times $e = 1$ and $e = 2$. Within a panel, each row in the table shows results from 5000 simulations with a fixed number of clusters.

The fourth column reports the average coefficient estimate across the 5000 simulations. As expected, the average simulation is a close match to the theoretical parameter, especially as the number of clusters gets large. The fifth column shows the empirical standard deviation of the point estimates across the 5000 simulations, and the sixth and seventh columns contains the average estimated standard error, clustering at the county level and the county x sub-experiment level. Both sets of standard errors correspond closely to the observed standard deviation, particularly when the number of clusters is 100 or more. The final columns show the rejection rates for the t-test against the true null. Both tests perform quite well with a large number of clusters: with 2500 clusters, the rejection rates are between .04 and .05 for all three treatment effect parameters. Performance is slightly worse with 50 clusters, where the rejection rates are between .06 and .08 across the three parameters. We expected worse performance when standard errors are clustered at the county $\times$ sub-experiment level because this approach does not account for repeated observations across sub-experiments. But in practice — at least in this example — the two methods produced nearly identical results.

Overall, the Monte Carlo simulations suggest that the weighted stacked event study does not pose any new challenges from a statistical inference point of view. Standard errors that allow for clustering at the level of the treatment work well when the number of clusters is not

too small. In applications, the standard alternatives (i.e. wild bootstrap, block bootstrap, and degree of freedom adjustments) could be appropriate for the usual reasons.

5.5 Alternative Stacked Regression Specifications – Fixed Effects

The stacked event study regressions we proposed above are different from the regression specifications used in applied work by Cengiz et al. (2019), Deshpande and Li (2019), and Butters et al. (2022). The regressions used in these papers include *state* × *sub-experiment* and *sub-experiment* × *event time* fixed effects. Using our notation, these stacked fixed effect regression specifications look like:

$$Y_{sae} = \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\delta_e^{fe} (D_{sa} \times 1[e=h]) \right] + m_{sa} + v_{ae} + U_{sae} \tag{4}$$

The stacked fixed effect specification looks more complicated than the event study specification in our main analysis because it incorporates potentially high dimensional fixed effects. But the complexity is not necessarily desirable. The event study model in equation 3 is a saturated specification for the conditional expectation function linking realized outcomes across cells defined by treatment status and event time, $E[Y_{sae}|D_{sa}, e]$. Because the specification contains a parameter for every value of the conditioning variables, the event study specification does not impose any parametric structure on the shape of the conditional expectation function in observed outcomes. The common trend and no anticipation assumptions are imposed on the potential outcomes, which gives a causal interpretation to the coefficients in the event study specification.

In contrast, the fixed effect specification is a model of the conditional expectation function linking observed outcomes across states, sub-experiments, and event time periods – $E[Y_{sae}|s, a, e]$. But the regression is not fully saturated in state, sub-experiments, and event times. That means that the specification imposes more parametric structure on the observed outcome function in addition to the common trends and no anticipation assumptions. Thus, estimates based on the fixed effect structure are more dependent on modeling assumptions than the saturated event study specification.

One way to build intuition for stacked fixed effects regressions is to consider the saturated event study regression fitted separately to each sub-experiment.

$$Y_{se}^a = \alpha_0^a + \alpha_1^a D_s^a + \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\lambda_e^a 1[e=h] + \delta_e^a (D_s^a \times 1[e=h]) \right] + \epsilon_{sae}$$

$$= \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\delta_e^a D_s^a \times 1[e=h] \right] + m_{sa} + v_{ae} + U_{sae}$$

The second line re-parameterizes the two group event study regression from equation 1, replacing the event time main effects with fixed effects. When fitted to data from a single sub-experiment, each $\delta_e^a = ATT(a, a+e)$ under the common trends and no-anticipation assumptions. These sub-experiment specific models could be fit in a single pass through the stacked data set by fully interacting all of the variables with sub-experiment identifiers:

$$Y_{sae} = \sum_{j \in \Omega_{\kappa_{pre},\kappa_{post}}} \sum_{\substack{h=-\kappa_{pre}...\kappa_{post} \\ h\neq -1}} \left[\delta_{ed} (D_{sa} \times 1[e=h] \times 1[a=j]) \right] + m_{sa} + v_{ae} + U_{sae}$$

The fully interacted model reproduces the parameters from the sub-experiment specific regressions, but it does not aggregate the effects into a summary measure. The stacked fixed effect specification in equation 4 produces an aggregate effect by suppressing the three way interaction terms. However, it is not clear how the coefficients from equation 4 relate to underlying causal effects, or to our target aggregate parameter, θ_κ^e.

To investigate the structure of the stacked fixed effect regression coefficient, focus on the simple case of a stacked data set with two periods corresponding to event times $e = -1$ and $e = 0$. In this case, there is no event study and only a single coefficient of interest:

$$Y_{sae} = \delta_0^{fe} D_{sae} + m_{sa} + v_{ae} + U_{sae}$$

Using the Frisch-Waugh-Lowell theorem, $\delta_0^{fe} = \frac{Cov(Y_{sae}, \widetilde{D_{sae}})}{Var(\widetilde{D_{sae}})}$, where $\widetilde{D_{sae}} = D_{sae} - \overline{D_{sa}} - \overline{D_{ea}} + \overline{D_a}$. Limiting the sample to a single sub-experiment and applying the same logic implies that the sub-experiment specific causal effect is $ATT(a,0) = \delta_e^a = \frac{Cov_a(Y_{sae}, \widetilde{D_{sae}})}{Var_a(\widetilde{D_{sae}})}$. Let $\frac{N_a}{N} = \frac{N_a^D + N_a^C}{N^D + N^C}$ represent the share of stacked observations that belong to sub-experiment a. Using the law of total covariance and variance, we can write the stacked fixed effect parameter as the sum

of two components.[8]

$$\delta_0^{fe} = \frac{\sum_a ATT(a,0) \times Var_a(\widetilde{D_{sae}})\frac{N_a}{N}}{\sum_a Var_a(\widetilde{D_{sae}})\frac{N_a}{N} + \sum_a (E[\widetilde{D_{sae}}|a] - E[\widetilde{D_{sae}}])^2 \frac{N_a}{N}}$$

$$+ \frac{\sum_a (E[Y_{sae}|a] - E[Y_{sae}])(E[\widetilde{D_{sae}}|a] - E[\widetilde{D_{sae}}])\frac{N_a}{N}}{\sum_a Var_a(\widetilde{D_{sae}})\frac{N_a}{N} + \sum_a (E[\widetilde{D_{sae}}|a] - E[\widetilde{D_{sae}}])^2 \frac{N_a}{N}}$$

The first component is a weighted sum of causal effects from each sub-experiment. The second component depends on the covariance between the outcomes and treatment exposure levels across sub-experiments. Although it is not particularly intuitive, this decomposition shows that, in general, the stacked fixed effect coefficient does not represent a coherent aggregation of underlying causal effects. The first term is a non-convex combination of effects, rather than a proper average of ATTs. And the second term potentially incorporates non-causal associations across sub-experiments.

The fixed effect expression simplifies in the special case where the share of treated units is constant across sub-experiments. Then $E[\widetilde{D_{sae}}|a] = E[\widetilde{D_{sae}}]$ for all sub-experiments, making second term drop out because the cross-experiment covariance equals zero. The first term simplifies as well because the second term in its denominator drops out, leaving a simpler form for the fixed effect coefficient:

[8]The law of total covariance implies that

$$\begin{aligned} Cov(Y_{sae}, \widetilde{D_{sae}}) &= E[Cov(Y_{sae}, \widetilde{D_{sae}}|a)] + Cov(E[Y_{sae}|a], E[\widetilde{D_{sae}}|a]) \\ &= \sum_a ATT(a,0) \times Var_a(\widetilde{D_{sae}})\frac{N_a}{N} \\ &+ \sum_a (E[Y_{sae}|a] - E[Y_{sae}])(E[\widetilde{D_{sae}}|a] - E[\widetilde{D_{sae}}])\frac{N_a}{N} \end{aligned}$$

Similarly, the law of total variance implies that:

$$\begin{aligned} Var(\widetilde{D_{sae}}) &= E[Var(\widetilde{D_{sae}}|a)] + Var(E[\widetilde{D_{sae}}|a]) \\ &= \sum_a Var_a(\widetilde{D_{sae}})\frac{N_a}{N} + \sum_a (E[\widetilde{D_{sae}}|a] - E[\widetilde{D_{sae}}])^2 \frac{N_a}{N} \end{aligned}$$

$$\delta_0^{fe} = \frac{\sum_a ATT(a,0) \times Var_a(\widetilde{D_{sae}}) Pr(a)}{\sum_a Var_a(\widetilde{D_{sae}}) Pr(a)}$$

This shows that in the constant treatment share case, the stacked fixed effect coefficient corresponds to a convex combination of sub-experiment specific group-time ATTs. Each group-time ATT receives a weight proportional to its sub-experiment's share of the total (within) variance in the treatment variable. However, since the within variance in treatment will be equal in each sub-experiment under the constant treatment shares case, the weights are really driven by the sample sizes in each sub-experiment. If sub-experiment total sample sizes also are equal then the stacked design gives equal weight to each sub-experiment.

6 Application

We illustrate the method by examining the effects of the ACA Medicaid expansion on health insurance coverage among adults ages 19 to 60. We use data from the 2011 to 2021 waves of the American Community Survey, obtained from IPUMS. The outcome variable is a dummy variable indicating that the person reports that they were not covered by any health insurance plan during the year. Using sampling weights, we collapse the microdata into state $\times$ calendar year uninsurance rates.

As of 2021, a total of 40 states had adopted the ACA Medicaid expansions. 28 states adopted in 2014, 3 states adopted in 2015, 2 states adopted in 2016, 2 states adopted in 2019, 3 states adopted in 2020, and 2 states adopted in 2021. These adoption events represent potential sub-experiments to be examined in the staggered adoption design.

We set $\kappa_{pre} = 3$ and $\kappa_{post} = 2$ to define a uniform event window for each sub-experiment. Table 3 has a row for each sub-experiment and shows the treated states, clean controls, and calendar years included in each sub-experiment. The 2020 and 2021 adoption events are trimmed from the analytic sample because they occur so recently that there is not enough follow up data to maintain the κ window. For each of the feasible adoption events, we form a sub-experimental data set consisting of data on the treated states and any states that are either never treated or are not treated for at least κ_{post} years after the focal adoption event. We limit these data to the relevant calendar years so that each of the 4 sub-experimental data sets consists of 6 event time observations for each treated and control state. Concatenating the feasible sub-experimental data sets yields a stacked dataset with 600 observations in total. Table 3 show the fraction of treated units and of all units in the stacked sample that fall into each sub-experiment.

Table 3: Sub-Experiments in the ACA Expansion Staggered Adoption Design

Sub-Experiment	Starting Year	Ending Year	Stack Share	Treated Share	Treated States	Control States
2014	2011	2016	0.46	0.80	AZ, AR, CA, CO, CT, DE, DC, HI, IL, IA, KY, MD, MA, MI, MN, NV, NH, NJ, NM, NY, ND, OH, OR, RI, VT, WA, WV, WI	AL, FL, GA, ID, KS, ME, MS, MO, NE, NC, OK, SC, SD, TN, TX, UT, VA, WY
2015	2012	2017	0.21	0.09	AK, IN, PA	AL, FL, GA, ID, KS, ME, MS, MO, NE, NC, OK, SC, SD, TN, TX, UT, VA, WY
2016	2013	2018	0.20	0.06	LA, MT	AL, FL, GA, ID, KS, ME, MS, MO, NE, NC, OK, SC, SD, TN, TX, UT, VA, WY
2019	2016	2021	0.13	0.06	ME, VA	AL, FL, GA, KS, MS, NC, SC, SD, TN, TX, WY
2020	Trimmed	Trimmed	–	–	ID, NE, UT	–
2021	Trimmed	Trimmed	–	–	MO, OK	–

Table 4 shows estimated coefficients from simple two-group event study regressions fitted to each sub-experiment separately. Standard errors that allow for clustering at the state level are shown in parentheses. The first column of Panel B shows results from the 2014 sub-experiment, in which 28 of the 35 treated states first adopted the ACA. The event study coefficients are small and not statistically significant during the pre-treatment period, but they are larger and statistically significantly negative during the post-treatment periods. The estimates imply that adopting the ACA reduced uninsurance rates by about 1.7 percentage points in the initial year and by about 2.4 percentage points 1 year and 2 years after adoption. Columns 2-4 show effects for the 2015, 2016, and 2019 sub-experiments. The results suggest the ACA reduced uninsurance rates by 2 to 5 percentage points after 2 years in each sub-experiment. The pre-treatment coefficients mostly support the common trend and no-anticipation assumptions, although there a few exceptions in 2016 and 2019. Panel A shows the average of the post-treatment event study coefficients for each group. The average post-period effect suggests that the over the first three years after adoption, the ACA Medicaid expansion reduced uninsurance by about 2.2 percentage points in the 2014 adoption group, 1.6 percentage points in the 2015 adoption group, 4.3 percentage points in the 2016 adoption group, and 1.5 percentage point in the 2019 adoption group.

The title row of the table reports the share of treated observations that belong to each of the four sub-experiments. For example, 28 of the 35 treated states in the trimmed sample

Table 4: Sub Experiment Specific Event Studies

	2014 28/35 Treated	2015 3/35 Treated	2016 2/35 Treated	2019 2/35 Treated	Weighted Stacked
	(1)	(2)	(3)	(4)	(5)
A. Post-Treatment Average Effect					
Treated (=1)×Post (=1)	−2.16***	−1.59***	−4.21***	−1.52***	−2.19***
	(0.645)	(0.512)	(0.407)	(0.383)	(0.561)
B. Event-studies					
Treated (=1)	−4.68***	−1.61	1.56	−5.66***	−4.12***
	(1.59)	(2.54)	(1.46)	(1.19)	(1.34)
Event-time, -3 (=1)	0.749***	3.30***	5.17***	−0.877***	1.13***
	(0.242)	(0.259)	(0.418)	(0.177)	(0.203)
Event-time, -2 (=1)	0.423	2.87***	2.30***	−0.240	0.703***
	(0.270)	(0.235)	(0.336)	(0.251)	(0.222)
Event-time, 0 (=1)	−2.87***	−2.30***	−1.04***	0.333	−2.54***
	(0.226)	(0.335)	(0.284)	(0.238)	(0.184)
Event-time, +1 (=1)	−5.17***	−3.34***	−0.576*	−0.150	−4.47***
	(0.400)	(0.347)	(0.292)	(0.274)	(0.336)
Event-time, +2 (=1)	−6.22***	−2.88***	−0.493**	−0.503*	−5.28***
	(0.465)	(0.311)	(0.215)	(0.254)	(0.384)
Treated (=1)×Event-time, -3 (=1)	−0.272	−0.366	1.55***	1.01***	−0.102
	(0.313)	(0.567)	(0.525)	(0.185)	(0.368)
Treated (=1)×Event-time, -2 (=1)	−0.390	−0.855**	0.868**	0.564**	−0.303
	(0.316)	(0.373)	(0.354)	(0.251)	(0.299)
Treated (=1)×Event-time, 0 (=1)	−1.68***	−1.20***	−2.53***	−0.657	−1.63***
	(0.450)	(0.366)	(0.648)	(0.609)	(0.393)
Treated (=1)×Event-time, +1 (=1)	−2.41***	−1.46*	−4.63***	−1.23	−2.39***
	(0.737)	(0.747)	(0.801)	(0.697)	(0.645)
Treated (=1)×Event-time, +2 (=1)	−2.38***	−2.09***	−5.48***	−2.68***	−2.55***
	(0.831)	(0.584)	(0.855)	(0.332)	(0.707)
Observations	276	126	120	78	600

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. The uninsurance rate is multipled by 100.

adopted the ACA in 2014, 3 of 35 adopted in 2015, and so on. To manually compute trimmed ATT aggregate, multiply each of the sub-experiment specific coefficients by the treatment share and then sum them up. The final column shows the point estimates from the weighted stacked event study regression, which pools observations from each of the four sub-experiments. These coefficients give the same answer as manually computing the weighted sum. Column 5 of Panel A shows that the average post-period trimmed aggregate ATT of the ACA expansion on uninsurance rates was about -2.2 percentage points.

Table 5 (Panel B) shows estimated coefficients from several different stacked regression specifications. Panel A shows the post-treatment average of the event study coefficients. In each regression, standard errors are shown in parenthesis and are estimated using a cluster robust variance matrix that allows for dependence at the state level. The first column shows point estimates from the stacked unweighted regression. The unweighted specification is quite misleading. It implies statistically significant negative pre-trends, which were not evident

Table 5: Stacked Event Study Regressions

	Stacked ES	Stacked ES Weights	Stacked FE	Stacked FE Weights
	(1)	(2)	(3)	(4)
A. Post-Treatment Average Effect				
Treated (=1)×Post (=1)	−3.97***	−2.19***	−2.22***	−2.19***
	(0.537)	(0.561)	(0.475)	(0.540)
B. Event-studies				
Treated (=1)	−1.95	−4.12***		
	(1.26)	(1.34)		
Event-time, -3 (=1)	2.40***	1.13***		
	(0.188)	(0.203)		
Event-time, -2 (=1)	1.51***	0.703***		
	(0.157)	(0.222)		
Event-time, 0 (=1)	−1.67***	−2.54***		
	(0.140)	(0.184)		
Event-time, +1 (=1)	−2.54***	−4.47***		
	(0.224)	(0.336)		
Event-time, +2 (=1)	−2.74***	−5.28***		
	(0.239)	(0.384)		
Treated (=1)×Event-time, -3 (=1)	−1.38***	−0.102	0.035	−0.102
	(0.365)	(0.368)	(0.291)	(0.282)
Treated (=1)×Event-time, -2 (=1)	−1.11***	−0.303	−0.232	−0.303
	(0.255)	(0.299)	(0.258)	(0.271)
Treated (=1)×Event-time, 0 (=1)	−2.50***	−1.63***	−1.59***	−1.63***
	(0.381)	(0.393)	(0.324)	(0.371)
Treated (=1)×Event-time, +1 (=1)	−4.31***	−2.39***	−2.38***	−2.39***
	(0.607)	(0.645)	(0.549)	(0.614)
Treated (=1)×Event-time, +2 (=1)	−5.09***	−2.55***	−2.69***	−2.55***
	(0.657)	(0.707)	(0.616)	(0.693)
Observations	600	600	600	600
Sub-experiment-State FEs	No	No	Yes	Yes
Sub-experiment-Event-time FEs	No	No	Yes	Yes
Stack weights	No	Yes	No	Yes

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. The uninsurance rate is multiplied by 100.

in the underlying sub-experiment specific event studies. This occurs because the regression implicitly weights treated and control group time trends differently when forming the average effects. The second column shows the same saturated event study specification but this time the model is estimated using weighted least squares and the corrective sample weights. The pre-trends are not apparent in the weighted specification and the event study coefficients are the correct weighted average across sub-experiments.

Columns 3 and 4 of Table 5 show estimates from unweighted and weighted stacked event study specifications that include state × sub-experiment and event time × sub-experiment fixed effects. The weighted stacked fixed effect specification produces the same treatment effect estimates as the simpler weighted event study without fixed effects. Interestingly, the coefficients from the unweighted stacked fixed effect model (Column 3) are only slightly different from the weighted stacked fixed effect specification (Column 4), implying that

– in this example – the stacked fixed effect specification does not suffer from substantial bias. Of course, the analysis in section 5.5 shows that the two estimators (fixed effects vs weighted event study) are not equivalent in general. See Appendix A for results from a simulated example in which the bias in the unweighted stacked fixed effect specification is more prominent.

7 Discussion

This paper shows how to use a stacked regression to analyze data from a staggered adoption DID design. Stacked estimators are appealing because they provide a one-step, regression-based method of pooling information from multiple (sub-experimental) difference-in-difference designs in a way that nevertheless results in a well-defined and logical average causal effect. We make three main contributions. First, we show how to apply a trimming rule to staggered adoption designs to ensure that the aggregate average does not suffer from compositional bias, which can be an important problem in event studies. Second, we show that the simple saturated stacked regression estimators do not identify a causal parameter and are biased because of differential (implicit) weighting. This negative identification result also applies to non-saturated stacked fixed effect regressions in the general case. Third, we derive sample weights that correct for the differential weighting bias and show that a simple weighted least squares estimator identifies a well-defined causal effect that we call the "trimmed aggregate ATT". We think that the trimmed aggregate ATT is a sensible target parameter in many applied settings. However, we also show how to create corrective sample weights to identify other sensible aggregations, such as the trimmed per capita (population) ATT and the trimmed sample share weighted ATT. The corrective weights are a function of sample sizes by sub-experiment and they are easy to compute using the data. The weighted stacked DID and Event Study estimators are intuitive and emphasize the key sources of variation in the study, making it clear how each treatment exposure event contributes to the analysis.

References

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics 119*(1), 249–275.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

Butters, R. A., D. W. Sacks, and B. Seo (2022). How do national firms respond to local cost shocks? *American Economic Review 112*(5), 1737–1772.

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

Callison, K. and R. Kaestner (2014). Do higher tobacco taxes reduce adult smoking? new evidence of the effect of recent cigarette tax increases on adult smoking. *Economic Inquiry 52*(1), 155–172.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

Deshpande, M. and Y. Li (2019). Who is screened out? application costs and the targeting of disability programs. *American Economic Journal: Economic Policy 11*(4), 213–248.

Dube, A., D. Girardi, Ò. Jordà, and A. M. Taylor (2023). A local projections approach to difference-in-differences event studies. Technical report, National Bureau of Economic Research.

Gardner, J. (2022). Two-stage differences in differences.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Heckman, J. J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The review of economic studies 65*(2), 261–294.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.

A Simulated Infant Mortality Rates–Staggered Adoption Design

We constructed another example of a staggered adoption design using CDC Wonder data on infant mortality rates in the 50 states from 1978 to 2020. We created 4 simulated adoption groups. The earliest group is Mississippi in 1985. Then Alabama, Alaska, and Arizona adopt in 1987. A collection of 14 states adopt in 2000: Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming. And finally, Massachusetts adopts in 2011. The remaining states are never treated.

In each adoption group, we generate treated outcomes by adding a constant and time invariant treatment effect to the observed infant mortality rate. Specifically, we set the treatment effect to be -12 deaths per 1000 births in the 1985 adoption group, $+12$ deaths per 1000 births in the 1987 group, -8 deaths per 1000 births in the 2000 group, and $+12$ deaths per 1000 births in the 2011 adoption group.

Setting the balanced event window to cover a period from 6 periods before treatment to 10 periods after treatment, we form a trimmed set of sub-experiments for 1985, 1987, and 2000. The 2011 adoption group is trimmed (excluded) because it occurs too recently to be studied for a full 10 year follow up. We formed the trimmed and balanced stacked data set and construct the corrective weights using the procedures described in the paper. With the stacked data in hand, we fit event study regression models with and without weights and with and without state $\times$ event time and state $\times$ sub-experiment fixed effects.

The true ATT within this balanced sample is a constant -4.9 in each event time period. However, because we are using a real data set rather than a full simulation, it is possible that the common trends assumption does not hold perfectly in this example: thus, the regression estimates will not necessarily perfectly recover the true treatment effect. The estimates are in Table A1. The weighted stacked event study and weighted stacked fixed effect estimates are identical – as expected – and they are in fact very close to the the true effect in most of the post-treatment time periods. The average effect across all post-treatment time periods is estimated to be about -4.7. However, unlike the ACA example presented in the paper, in this example the stacked event study coefficients are different in the weighted and unweighted fixed effect models. We emphasize this point to make it clear that the close correspondence between the fixed effect and weighted fixed effect estimates in the ACA example is simply something that happenend in that example and is not a generalizable result.

The clean controls inclusion criteria we focused on in the paper is based around the idea that a clean control is not exposed to treatment at any time during the κ event window. That is, units with treatment adoption dates $A_s > a + \kappa_{post}$ were eligible clean controls for sub-experiment a. In practice, this group may consist of a mixture of "never treated" and "not yet treated" units. But our overall approach to stacked DID analysis is compatible with some alternative clean control restrictions that may be useful in applied work. One alternative is to use a stricter rule that defines clean controls as units that are not exposed to treatment long enough after the κ window that they are not themselves in their own pre-treatment period. Formally, this strict rule includes as clean controls any units with $_s > a + \kappa_{post} + \kappa_{pre}$. A third approach is to use only units that are "never treated" as clean controls.

Table A2 illustrates how these different clean control definitions one matter when analyzing the CDC Wonder example. Column (1) shows results based on the standard sample inclusion criteria we have used throughout the paper. Column (2) presents results from the strict clean controls inclusion criteria ($_s > a + \kappa_{post} + \kappa_{pre}$), where we omit any not yet treated group who has pre-period data that falls within our κ_{pre} bound. Column (3) shows results where only never treated units are used as clean controls. Panel A of Table A2 shows estimates with our corrective weights, while Panel B shows estimates from analogous regressions without the corrective weights.The results make it clear that the corrective weights matter regardless of the specific inclusion criteria used.

Table A1: Stacked event study regressions, using infant mortality data and an imposed effect

	Stacked ES	Stacked ES Weights	Stacked FE	Stacked FE Weights
	(1)	(2)	(3)	(4)
Real effect	−4.9	−4.9	−4.9	−4.9
A. Post-Treatment Average Effect				
Treated (=1)×Post (=1)	−4.1**	−4.7***	−4.0*	−4.7***
	(1.7)	(1.7)	(2.0)	(1.6)
B. Event-studies				
Treated × Event-time, -6	−0.68**	−0.03	−0.007	−0.03
	(0.31)	(0.29)	(0.26)	(0.25)
Treated × Event-time, -5	−0.48**	0.20	0.17	0.20
	(0.21)	(0.21)	(0.20)	(0.19)
Treated × Event-time, -4	−0.57*	−0.03	−0.03	−0.03
	(0.30)	(0.31)	(0.28)	(0.30)
Treated × Event-time, -3	−0.14	0.30	0.28	0.30
	(0.19)	(0.23)	(0.21)	(0.23)
Treated × Event-time, -2	0.03	0.21	0.19	0.21
	(0.16)	(0.21)	(0.21)	(0.22)
Treated × Event-time, 0	−5.0***	−5.1***	−4.4**	−5.1***
	(1.8)	(1.8)	(2.1)	(1.6)
Treated × Event-time, 1	−4.8**	−4.7**	−4.0*	−4.7***
	(1.9)	(1.9)	(2.1)	(1.7)
Treated × Event-time, 2	−4.4**	−4.6**	−3.9*	−4.6***
	(1.8)	(1.8)	(2.0)	(1.6)
Treated × Event-time, 3	−4.5**	−4.7**	−4.0*	−4.7***
	(1.8)	(1.8)	(2.1)	(1.6)
Treated × Event-time, 4	−4.4**	−4.7**	−4.0*	−4.7***
	(1.7)	(1.7)	(2.0)	(1.6)
Treated × Event-time, 5	−3.9**	−4.6***	−3.9*	−4.6***
	(1.7)	(1.7)	(2.0)	(1.6)
Treated × Event-time, 6	−4.1**	−4.8***	−4.2**	−4.8***
	(1.7)	(1.7)	(2.0)	(1.6)
Treated × Event-time, 7	−3.8**	−4.8***	−4.1**	−4.8***
	(1.6)	(1.7)	(2.0)	(1.6)
Treated × Event-time, 8	−3.7**	−4.6***	−3.9*	−4.6***
	(1.6)	(1.7)	(2.0)	(1.6)
Treated × Event-time, 9	−3.6**	−4.6***	−3.8*	−4.6***
	(1.7)	(1.7)	(2.1)	(1.6)
Treated × Event-time, 10	−3.5**	−4.4**	−3.7*	−4.4***
	(1.7)	(1.7)	(2.0)	(1.6)
Observations	2465	2465	2465	2465
Treated in Sub-experiment FEs	Yes	Yes	No	No
Event-time FEs	Yes	Yes	No	No
Event-time by Sub-experiment FEs	No	No	Yes	Yes
State by Sub-experiment FEs	No	No	Yes	Yes

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A2: Stacked event study regressions, using infant mortality data and an imposed effect, changing inclusion criteria

	All not yet treated	Strict not yet treated	No not yet treated
	(1)	(2)	(3)
Real effect	−4.9	−4.9	−4.9
A. Post-Treatment Average Effect, with weights			
Treated (=1)×Post (=1)	−4.7***	−4.7***	−4.7***
	(1.6)	(1.6)	(1.5)
B. Post-Treatment Average Effect, without weights			
Treated (=1)×Post (=1)	−4.0*	−4.1**	−4.1**
	(2.0)	(2.0)	(2.0)
Observations	2465	1972	1938
Event-time-Sub-experiment FEs	Yes	Yes	Yes
State-Sub-experiment FEs	Yes	Yes	Yes

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.