# Two-Stage Differences in Differences

John Gardner  Neil Thakral  Linh T. Tô  Luther Yap

July 2023*

**Abstract**

This paper develops a framework for estimation of the effect of a policy or treatment in settings with variation in treatment timing. We propose a simple and intuitive two-stage estimator that is robust to treatment-effect heterogeneity. In the first stage, we identify group and period effects using the sample of untreated observations. In the second stage, we estimate average treatment effects by comparing treated and untreated outcomes after removing these group and period effects. The procedure can be extended to environments with individual fixed effects and inference can be conducted from within a conventional asymptotic framework. We establish the theoretical properties of the two-stage approach and highlight its advantages over existing proposals. Simulations of randomly generated placebo laws in state-level wage data demonstrate that our method outperforms alternative approaches for estimation and inference.

*Gardner: Department of Economics, University of Mississippi (email: jrgardne@olemiss.edu). Thakral: Department of Economics, Brown University (email: neil_thakral@brown.edu). Tô: Department of Economics, Boston University (email: linhto@bu.edu). Yap: Department of Economics, Princeton University (email: lyap@princeton.edu). We thank Kyle Butts, Tao Wang, Scott Cunningham, Brantly Callaway, Taylor Wright, David Drukker, Len Goff, Carolina Caetano, Gregorio Caetano and participants at the Western Economics Association International conference and the Southern Economic Association annual meeting for helpful comments and suggestions. This paper supersedes earlier versions that circulated under the same title by Gardner (2020).

# 1 Introduction

Difference-in-differences (DD) estimation has emerged as an indispensable tool for empirical researchers seeking to evaluate the impact of a given intervention or policy. Its appeal stems in part from the conceptual simplicity of comparing changes in outcomes for groups affected by an intervention to changes for unaffected groups. A potential reason for the widespread use of two-way fixed-effects (TWFE) in settings with multiple groups and time periods is a presumption that it should identify the average effect of the treatment on the treated. Although this intuition is accurate when the heterogeneous treatment effects are distributed identically across groups and periods (a condition that is automatically satisfied in the classic two-group, two-period setting), it does not hold in general. When these distributions are not identical, conditional mean outcomes are no longer linear in group, period, and treatment status, causing the TWFE regression model to be misspecified for conditional mean outcomes, and thus unable to identify the average treatment effect on the treated.

This paper develops a novel two-stage regression-based approach to identification that is robust to treatment-effect heterogeneity when adoption of the treatment is staggered over time. The first stage regresses outcomes on group and period fixed effects using the subsample of untreated observations. The second stage subtracts the estimated group and period effects from observed outcomes and regresses the resulting residualized outcomes on treatment status. Under the usual parallel trends assumption, this procedure identifies the overall average effect of the treatment on the treated (i.e., across groups and periods), even when average treatment effects are heterogeneous over groups and periods. This approach preserves the intuition behind identification in the two-group, two-period case: it recovers the average difference in outcomes between treated and untreated units, after removing group and period effects. Furthermore, we demonstrate how to extend this approach to recover a variety of treatment effect measures, including event-study analyses of pre-trends and duration-specific average treatment effects.

We derive the asymptotic distribution of the treatment effect estimates by interpreting the two-stage procedure as a joint GMM estimator. The two-stage estimator

can thus be implemented easily, along with valid asymptotic standard errors, using standard statistical software, and with little programming beyond that required to estimate a regression.[1]

To evaluate the performance of our proposed estimator, we conduct simulations of randomly generated placebo laws in state-level wage data. We build on the seminal work of Bertrand, Duflo and Mullainathan (2004) by analyzing the performance of various DD estimators for placebo laws in which states and their associated years of passage are randomly drawn. Using a 40-year panel, we compare the rejection rates at the 5 percent significance level from the estimators proposed in De Chaisemartin and d'Haultfoeuille (2020); Callaway and Sant'Anna (2021); Sun and Abraham (2021); Borusyak, Jaravel and Spiess (2023). The simulation exercises highlight the value of our approach to estimation and inference by demonstrating its finite-sample performance. Our estimator consistently offers the best performance in terms of rejection rates and efficiency, including in comparison to the imputation approach from Borusyak, Jaravel and Spiess (2023) that provides identical point estimates to ours with different variance estimators.[2] We document these advantages even in cases with homogeneous treatment effects or with independent and identically distributed data.

Our work adds to an emerging body of research highlighting limitations of the traditional TWFE approach for DD estimation in the presence of staggered treatment timing and the effects of a treatment vary across groups and time.[3] Several papers have provided alternative representations of the TWFE regression estimand. Borusyak, Jaravel and Spiess (2023) show that TWFE identifies a regression-weighted mean of

[1]Our approach to inference accommodates design-based sources of uncertainty. As Abadie et al. (2020) emphasize, the design-based perspective provides a coherent interpretation for standard errors, particularly for empirical settings where the source of randomness is known.

[2]The imputation estimator, which first appears in Borusyak, Jaravel and Spiess (2021), is numerically identical to the two-stage estimators initially proposed by Gardner (2020) and Thakral and Tô (2020). However, they develop a different asymptotic theory, resulting in an asymptotically conservative default variance estimator and a leave-one-out modification which they show results in improved finite-sample performance.

[3]See, for example, De Chaisemartin and d'Haultfoeuille (2020); Goodman-Bacon (2021); Imai and Kim (2021); Sun and Abraham (2021); Athey and Imbens (2022); Borusyak, Jaravel and Spiess (2023).

the average effect of the treatment in each post-treatment period, and De Chaisemartin and d'Haultfoeuille (2020) show that all two-way fixed-effects regression estimates (which include DD regressions as a special case) identify weighted averages of group- and period-specific average treatment effects. Since the weights in both of these representations can be negative, the TWFE estimand may be difficult to interpret. Goodman-Bacon (2021) further shows that the TWFE estimate represents a weighted average of all two-group, two-period differences in differences, which under parallel trends identifies a combination of weighted averages of group×period-specific average treatment effects and changes over time in those effects. We discuss how these decomposition results can be interpreted as describing how misspecified TWFE regression models project heterogeneous treatment effects onto treatment status, group effects, and period fixed effects.

When treatment adoption is staggered, several alternatives to the TWFE regression approach exhibit robustness to heterogeneity across groups and periods. One alternative is to estimate separate average treatment effects for each group and period, which can then be aggregated to form measures of the overall effect of the treatment.[4] Another approach is the "stacked" difference-in-differences (see, e.g., Gormley and Matsa, 2011; Deshpande and Li, 2019; Cengiz et al., 2019), which attempts to transform the staggered adoption setting to a two-group, two-period design (in which difference in differences identifies the overall average effect of the treatment on the treated) by stacking separate datasets containing observations on treated and control units for each treatment cohort. Our approach provides another procedure that exhibits such robustness and has several advantages, including simplicity in estimation and inference, interpretability of the estimand, and strong finite-sample performance.

The paper proceeds as follows. Section 2 presents the main idea in a simple setting

[4]Gibbons, Suárez Serrato and Urbancic (2018) suggest an approach like this for fixed effects models; Borusyak, Jaravel and Spiess (2023) suggest such a solution for difference-in-differences models in which the duration-specific effects of the treatment are identical across groups, as do Callaway and Sant'Anna, 2021 for the case when treatment effects vary by group and duration and Sun and Abraham, 2021 in the event-study context). The method developed by Callaway and Sant'Anna (2021, cf. Abadie, 2005) also accommodates covariates more flexibly than traditional regression-based methods.

with group and period fixed effects and without covariates. It provides intuition for why the TWFE approach for DD estimation may not identify the average effect of the treatment on the treated, and shows how our proposed two-stage regression-based approach is robust to treatment-effect heterogeneity in settings with variation in treatment timing. Section 3 provides theoretical results in a more general setting that can include covariates and individual fixed effects. Section 4 demonstrates the performance of the two-stage approach compared to alternative proposals, and Section 5 concludes.

# 2 Motivating the two-stage approach in a simplified setting

## 2.1 The problem with difference-in-differences regression

Difference-in-differences (DD) research designs attempt to identify the causal effects of treatments under the parallel or common trends assumption. This assumption asserts that, absent the treatment, treated units would experience the same change in outcomes as untreated units. Mathematically, this amounts to the assumption that average untreated potential outcomes decompose into additive group and period effects. Let $i$ index units (e.g., states or, with microdata, individuals) and $t$ index calendar time (often years). Further, partition units and time into treatment groups $g \in \{0, 1, \ldots, G\}$ and periods $p \in \{0, 1, \ldots, P\}$ defined by the adoption of the treatment among successive groups, so that members of group 0 are untreated in all periods, only members of group 1 are treated in period 1, members of groups 1 and 2 are treated in period 2, and so on. Let $Y_{gpit}$, $Y_{gpit}(1)$ and $Y_{gpit}(0)$ denote the observed, treated, and untreated potential outcomes for the $i$th member of group $g$ during time $t$ of period $p$, let $D_{gp}$ be an indicator for whether members of group $g$ are treated in period $p$, and let $\beta_{gp} = \mathbb{E}[Y_{gpit}(1) - Y_{gpit}(0) \,|\, g, p]$ denote the average causal effect of the treatment for members of $g$ in $p$.[5] Assume for simplicity that the

[5] Causal effects for the never-treated group may be normalized to zero.

treatment is both irreversible and unanticipated (both of these assumptions can be at least partially relaxed).[6] Under parallel trends, mean outcomes satisfy

$$\mathbb{E}[Y_{gpit} \mid g, p, D_{gp}] = \lambda_i + \alpha_p + \beta_{gp} D_{gp}. \tag{1}$$

The idea behind differences in differences is to eliminate the permanent group effects $\lambda_g$ and secular period effects $\alpha_p$ in order to identify the average effect of the treatment. In the classic setup, there are only two periods (pre and post) and two groups (treatment and control). In this setting, within-group differences over time eliminate the group effects and within-period differences between groups eliminate the period effects. Hence the between-group difference in post-pre differences (i.e., the difference in differences) identifies the average effect of the treatment for members of the treatment group during the post-treatment period.

The two-period, two-group difference-in-differences estimate can be obtained using a regression of outcomes on group and period fixed effects and a treatment-status indicator:

$$Y_{gpit} = \lambda_g + \alpha_p + \beta D_{gp} + \varepsilon_{gpit}. \tag{2}$$

It follows from (1) that the coefficient on $D_{gp}$ in (2) identifies the average effect of the treatment on the treated, $\mathbb{E}[Y_{gpit}(1) - Y_{gpit}(0) \mid D_{gp} = 1]$.[7]

The regression approach suggests a natural way to extend the DD idea to settings with multiple groups and time periods. Unfortunately, as several authors have noted

[6] Briefly, if the treatment is anticipated for $r$ periods before adoption, the procedure developed below can be applied after redefining treated to mean having adopted the treatment for at least $r$ periods. If the treatment is reversible, our results still apply under the (strong) assumption that there are no within-unit spillovers of the treatment to future periods.

[7] There are several equivalent variations on this regression. Specification (2) is identical to a regression of outcomes on an indicator $Post_{it}$ for whether $t$ occurs in the post-treatment period, an indicator $Treat_{it}$ for whether $i$ belongs to the treatment group, and an interaction between the two. Often, the group and period effects $\lambda_g$ and $\alpha_p$ in (2) are replaced with individual and time effects $\lambda_i$ and $\gamma_t$. By the Frisch-Waugh-Lovell theorem, the coefficient on $D_{gp}$ in (2) can be obtained by regressing $Y_{gpit}$ on the residuals from a regression of treatment status on group and period effects. Since treatment status only varies by group and period, these residuals are the same as those from a regression of treatment status on individual and time effects, so the coefficients on treatment status from both specifications are identical (despite the fact that the latter model is misspecified for $\mathbb{E}[Y_{it} \mid i, t, D_{it}]$).

(De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Imai and Kim, 2021; Athey and Imbens, 2022; Borusyak, Jaravel and Spiess, 2023), when the average effect of the treatment varies across groups and over periods, the coefficient on $D_{gp}$ in specification (2) does not always identify an easily interpretable measure of the "typical" effect of the treatment. Although this result is now well established, because it is also somewhat counterintuitive, it bears further clarification.

While there are multiple ways to think about the typical effect of the treatment when that effect varies across groups and over time (see Section 2.4 below), an obvious candidate is the average $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1] = \mathbb{E}[Y_{gpit}(1) - Y_{gpit}(0) \mid D_{gp} = 1]$ of group- and period-specific average treatment effects, taken over all units that receive the treatment and all times during which they receive it (i.e., the expectation of $\beta_{gp}$ over the joint distribution of $g$ and $p$, conditional on being treated). This is analogous to the average $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1]$ identified by difference in differences in the two-period, two-group case. Hence, parallel trends can be expressed as

$$\mathbb{E}[Y_{gpit} \mid g, p, D_{gp}] = \lambda_g + \alpha_p + \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]D_{gp} + [\beta_{gp} - \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]]D_{gp}.$$

The difficulty with the regression approach is that, except in special cases, the "error term" $[\beta_{gp} - \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]]D_{gp}$ in this expression varies at the group×period level, and is not mean zero *conditional on group membership, period, and treatment status*. Consequently, the regression is misspecified in the sense that the conditional expectation $\mathbb{E}[Y_{gpit} \mid g, p, D_{gp}]$ is not a linear function of those variables (at least, not one in which the coefficient on $D_{gp}$ is $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1]$) in contrast to the two-group, two-period case, the coefficient on $D_{gp}$ from the regression DD specification (2) does not identify $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1]$ unless those average effects are independent of group and period (in which case $\beta_{gp} = \mathbb{E}[\beta_{gp} \mid D_{gp} = 1] = \beta$). Outside of this special case, when average treatment effects vary across groups and periods, and the adoption of the treatment by different groups is staggered over time, difference-in-differences regression does not recover a simple group×period average treatment effect (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Borusyak, Jaravel and Spiess, 2023).

## 2.2 The difference-in-differences regression estimand

In light of the preceding argument, we discuss what the DD regression identifies. To provide additional insight into the difference-in-differences estimand, it can be shown that, under parallel trends, the coefficient on $D_{gp}$ from the difference-in-differences regression specification (2) identifies the following weighted average of $\beta_{gp}$:

$$\beta^* = \sum_{g=1}^{G} \sum_{p=g}^{P} \omega_{gp} \beta_{gp},$$

with weights that take the form

$$\omega_{gp} = \frac{\tilde{\omega}_{gp}}{\sum_{g'=1}^{G} \sum_{p'=g'}^{P} \tilde{\omega}_{g'p'}}, \tag{3}$$

where

$$\tilde{\omega}_{gp} = [(1 - \Pr(D_{gp} = 1 \,|\, g)) - (\Pr(D_{gp} = 1 \,|\, p) - \Pr(D_{gp} = 1))] \Pr(g, p),$$

$\Pr(D_{gp} = 1 \,|\, p)$ is the fraction of units that are treated in period $p$, $\Pr(D_{gp} = 1 \,|\, g)$ is the fraction of periods in which members of group $g$ are treated, $\Pr(D_{gp} = 1)$ is the fraction of unit×times that are treated, and $\Pr(p, g)$ is the population share of observations that correspond to group $g$ and period $p$. This representation can be obtained from Theorem 1 of De Chaisemartin and d'Haultfoeuille (2020), who note that the weights $\omega_{gp}$ may also be negative. Our Appendix B presents an alternative derivation based on population regression algebra.[8]

Appearances notwithstanding, this weighting scheme is deeply intuitive. Specification (2) assumes a conditional expectation function that is linear in group, period, and treatment status. When misspecified, it will attribute some of the heterogeneous impacts of the treatment to group and period fixed effects.[9] As a group's observed

[8] One immediate implication of Equation (3) is that the weights must sum to one. Another is that $\omega_{11} = 1$ when there is only one treatment group, so the regression DD specification (2) identifies the average effect of the treatment on the treated, as noted above.

[9] This is consistent with the intuition that Equation (2) uses already-treated units as controls for newly treated ones (De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Borusyak,

treatment duration increases (i.e., the greater $\Pr(D_{gp} = 1 \mid g)$ is), more of that group's treatment effects will be absorbed by group fixed effects. Similarly, as the probability of being treated in a particular period (i.e., $\Pr(D_{gp} = 1 \mid p)$) increases, more of that period's treatment effects will be absorbed by period effects.

## 2.3 A two-stage approach

The observation that the problem arises from misspecification of (2) suggests a simple two-stage average treatment effect estimator for the multiple group and period case. As long there are untreated and treated observations for each group and period, $\lambda_g$ and $\alpha_p$ are identified from the subpopulation of untreated groups and periods. The overall group×period average effect of the treatment on the treated is then identified from a comparison of mean outcomes between treated and untreated groups, after removing the group and period effects.

This logic suggests the following regression-based two-stage estimation procedure:

1. Estimate the model
$$Y_{gpit} = \lambda_g + \alpha_p + u_{gpit}$$
on the sample of observations for which $D_{gp} = 0$, retaining the estimated group and time effects $\hat{\lambda}_g$ and $\hat{\alpha}_p$.

2. Regress adjusted outcomes $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p$ on $D_{gp}$.

Since parallel trends implies that

$$\mathbb{E}[Y_{gpit} \mid g, p, D_{gp}] - \lambda_g - \alpha_p = \beta_{gp} D_{gp} = \mathbb{E}[\beta_{gp} \mid D_{gp} = 1] D_{gp} + [\beta_{gp} - \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]] D_{gp},$$

where $\mathbb{E}[[\beta_{gp} - \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]] D_{gp} \mid D_{gp}] = 0$, this procedure identifies $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1]$, even when the adoption and average effects of the treatment are heterogeneous with respect to groups and periods.

Unbiasedness of the first-stage (and hence second-stage) estimates follows from standard arguments. If $P$ is fixed as the sample size increases, so does the consistency

Jaravel and Spiess, 2023).

of the first stage for the group and period effects. The consistency of the second-stage for $\mathbb{E}[\beta_{gp} \,|\, D_{gp} = 1]$ follows from the consistency of the first stage for the group and period effects and the continuous mapping theorem.[10]

In the DD analyses based on two-way fixed-effects regression, it is common to control for observable time-varying covariates by simply including them in the regression. The two-stage approach can readily be adapted to allow for such covariates: simply include them in the first-stage regression and amend the second-stage to

2' Regress $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p - X'_{gpit}\hat{\gamma}$ on $D_{gp}$ (where $X_{gpit}$ is the vector of covariates and $\hat{\gamma}$ is the estimated vector of coefficients on $X_{gpit}$ from the first-stage regression).

While this approach allows the effect of the treatment to depend arbitrarily on observable covariates, it does implicitly rule out the possibility of feedback from the treatment to the covariates and, as Sant'Anna and Zhao (2020) note, covariate-specific trends.[11] We revisit this point in Section 3.

## 2.4 2SDD estimands

Implemented as described, the two-stage difference-in-differences (2SDD) estimator identifies $\mathbb{E}[\beta_{gp} \,|\, D_{gp} = 1]$, where the expectation is implicitly taken with respect to all observed units and periods. This expectation can be expressed as

$$\mathbb{E}[\beta_{gp} \,|\, D_{gp} = 1] = \sum_{g=1}^{G} \sum_{p=g}^{P} \beta_{gp} \Pr(g, p \,|\, D_{gp} = 1), \tag{4}$$

where $\Pr(g, p \,|\, D_{gp} = 1)$ is the population share of the treated unit-times that correspond to group $g$ in period $p$. While this is a natural summary measure of

[10] Also note that restricting the sample to untreated observations does not introduce sample-selection bias because the selection is with respect to treatment status, which is a deterministic function of the group and period variables included in the model.

[11] In principle, the two-stage approach can be modified to accommodate the more stringent notion of conditional parallel trends introduced by Callaway and Sant'Anna (2021) by interacting the covariates with time indicators, although this remains more parametric than their inverse-probability weighting estimators. Caetano et al. (2022) discuss how a two-stage approach (in addition to methods based on inverse-probability weighting) can be used when the treatment also affects the covariates.

group×period-specific average treatment effects, and can be interpreted as an average effect on the treated (ATT), it may not be especially informative for program evaluation and policy analysis. For example, even if the effects of the treatment are identical across groups, this measure will put more weight on groups that are in early stages of the treatment.[12] Callaway and Sant'Anna (2021) provide a much richer discussion of how heterogeneous average treatment effects can be summarized.

If there is some treatment duration $\bar{P}$ such that a subset of groups has been treated for $\bar{P}$ periods, then an alternative summary measure is the $\bar{P}$-period average

$$\sum_{g=1}^{G} \sum_{p=g}^{g+\bar{P}-1} \beta_{gp} \Pr(g \mid D_g = 1)/\bar{P}, \tag{5}$$

where $\Pr(g \mid D_g = 1)$ is the fraction of treated units that belong to group $g$. Because this measure averages the group-specific average effects of the treatment for a common set of completed durations, it may provide a more balanced picture of the typical effect of the treatment, although it ignores the effects of the treatment for durations longer than $\bar{P}$ periods. The two-stage procedure can be modified to identify this measure by restricting the sample used in the second step to untreated observations and treated observations with durations no greater than $\bar{P}$.

It is worth noting that the two-stage procedure is equivalent to estimating the two-way fixed effects model

$$Y_{gpit} = \lambda_g + \alpha_p + B'_{gpit}\theta + e_{it},$$

where $B_{gpit}$ is a saturated set of interactions between group and period indicators for all treated observations, then aggregating the group×period-specific treatment effects estimates in $\theta$ as the sample analog of $\mathbb{E}[\beta_{gp} \mid D_{gp} = 1]$. One way to see this is to note that, by the Frisch-Waugh-Lovell (FWL) theorem, estimates of the $\lambda_g$ and $\alpha_p$ can be obtained by regressing $Y_{gpit}$ on the residuals from auxiliary regressions of group

[12] When treatment effects vary by group, it is unclear whether any summary measure will be informative about how the treatment might affect future groups. External validity with this type of heterogeneity is inherently challenging.

and period indicators on $B_{gpit}$. Since $B_{gpit}$ perfectly predicts group and time for all treated observations, the residuals from these auxiliary regressions will be zero for all treated units. Consequently, $\lambda_g$ and $\alpha_p$ are identified from variation in untreated outcomes, as they are in the two-stage procedure. In either case, the overall ATT is identified as $\mathbb{E}[\beta_{gp} \,|\, D_{gp} = 1] = \mathbb{E}[Y_{gpit} - \lambda_g - \alpha_p \,|\, D_{gp} = 1]$.[13]

## 2.5 Event studies

DD analyses are often accompanied by event-study regressions of the form

$$Y_{gpit} = \lambda_g + \alpha_p + \sum_{r=-\underline{R}}^{\overline{R}} \eta_r W_{rgp} + u_{gpit}, \tag{6}$$

where for $r \leq 0$ the $W_{rgp} \in \{W_{-\underline{R}gp}, \ldots, W_{0gp}\}$ are $(r+1)$-period leads of treatment adoption, and for $r > 0$ the $W_{rgp} \in \{W_{1gp}, \ldots, W_{\overline{R}gp}\}$ are $r$-period lags of adoption (i.e., indicators for being $r$ periods since treatment).[14] In principle, such regressions serve a dual purpose. First, they can be used to show how the effect of the treatment evolves over the course of the treatment. Second, the coefficients on the treatment adoption leads can be used as placebo tests for the plausibility of parallel trends.

Sun and Abraham (2021) show that, when duration-specific average treatment effects vary across groups, event-study regressions suffer from the same problem as DD regressions. This can be seen using an argument similar to the one presented for DD regressions in Section 2.1. Let $Y_{rgpit}$ denote potential outcomes after $r$ periods of treatment, and $\eta_{rgp} = \mathbb{E}[Y_{rgpit} - Y_{gpit}(0) \,|\, g, p, W_{rgp} = 1]$ be the average effect of being treated for $r$ periods for members of group $g$ in time period $p$.[15] Under parallel

[13] The same equivalence applies when covariates are included in the first stage of the two-stage procedure, with the caveat that, in this case, the two-way fixed-effects regression should include unit and time (rather than only group and period) indicators, and $B_{gpit}$ should contain a saturated set of unit and time indicators for treated observations.

[14] In event-study regressions, it is common practice to use calendar times $t$ in place of more coarse treatment periods $p$. When researchers do not wish to include leads, $\underline{R}$ can be set to zero.

[15] There is a one-to-one correspondence between duration- and period-specific treatment effects. In terms of the group×period average treatment effects $\beta_{gp}$, the duration-specific effects satisfy $\eta_{rgp} = \beta_{g,p-g+1}$. While in principle the duration-specific average treatment effects for each group might vary over time, in practice we only ever observe each treatment duration at most once for

trends,

$$\mathbb{E}\left[Y_{gpit} \,\middle|\, g, p, \{W_{rgp}\}_{r=-\underline{R}}^{\overline{R}}\right] =$$
$$\lambda_g + \alpha_p + \sum_{r=1}^{\overline{R}} \mathbb{E}[\eta_{rgp} \,|\, W_{rgp} = 1] D_{rpg} + \sum_{r=1}^{\overline{R}} [\eta_{rgp} - \mathbb{E}[\eta_{rgp} \,|\, W_{rgp} = 1]] W_{rgp},$$

where, in general, $\mathbb{E}\left[\sum_{r=1}^{P^*} [\eta_{rgp} - \mathbb{E}[\eta_{rgp} \,|\, W_{rgp} = 1]] W_{rgp} \,\middle|\, g, p, (W_{rgp})\right] \neq 0$. Hence, mean outcomes are not necessarily linear in group, period, and treatment-duration indicators, so the coefficients on the $W_{rgp}$ from (6) do not identify the average effects of being treated for $r$ periods. Sun and Abraham (2021) further show that the coefficients on the adoption leads and duration indicators identify weighted averages of all of the group×period-specific average treatment effects. An important consequence of this expression is that the coefficients on the treatment-adoption leads $W_{rgp}$, $r \leq 0$, may be nonzero even if trends are, in fact, parallel.

The two-stage procedure developed above can be extended to the event-study setting by amending the second stage of the procedure to:

2'. Regress $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p$ on $W_{-\underline{R}gp}, \ldots, W_{0gt}, \ldots, W_{\overline{R}gp}$.

Following the logic of the previous section, because $\mathbb{E}[Y_{gpit} \,|\, g, p, (W_{rgp})] - \lambda_g - \alpha_p$ is linear in the $W_{rgp}$, the coefficients on the $W_{rgp}$, $r > 0$, identify the average effects $\mathbb{E}[\eta_{rgp} \,|\, W_{rgp} = 1]$.[16] For $r \leq 0$, the coefficients on the $W_{rgp}$ can be used to test the hypothesis that $\mathbb{E}[Y_{gpit} \,|\, g, p, W_{rgp} = 1] = \lambda_g + \alpha_p$ (i.e., that the average first-stage residual is zero for all units who are $r + 1$ periods away from adopting the treatment), as implied by parallel trends.[17]

---

each group.

[16] This expectation is taken over all groups with treatment durations of at least $r$. Since under staggered adoption the completed treatment duration varies by group, the groups over which these duration-specific effects are averaged will vary across durations. These averages are also what the interaction-weighted estimator proposed by Sun and Abraham (2021) identifies. If all groups are treated for at least $\bar{P}$ periods, an alternative is to exclude observations corresponding to treatment durations longer than $\bar{P}$ periods from the second-stage sample, in which case the two-stage approach identifies duration-specific treatment effects, averaged over all groups.

[17] There are alternative approaches to testing the validity of parallel trends within the two-stage framework. Liu, Wang and Xu (2022) discuss how the method discussed above can be adapted to

## 2.6 Inference

The standard errors for the two-stage estimators need to be adjusted to account for the fact that the dependent variable $Y_{gpit} - \hat{\lambda}_g - \hat{\alpha}_p$ in the second-stage is generated using estimates obtained from the first stage of the procedure (Dumont et al., 2005). Perhaps the simplest way to obtain valid standard errors is using a bootstrap procedure in which both stages of the estimator are estimated in each bootstrap replication (this is the approach used in Liu, Wang and Xu, 2022). The asymptotic distribution of the second-stage estimates can also be obtained by interpreting the two-stage procedure as a joint GMM estimator (Hansen, 1982).[18]

Let $Z_{gpit} = [Y_{gpit}, (1(g)_{gpit}), (1(p)_{gpit}), D_{gp}]$ denote the data for observation $(g, p, i, t)$, consisting of the outcome $Y_{gpit}$, the $G$-vector of group-membership indicators $(1(g)_{gpit})$, a $P$-vector $(1(p)_{gpit})$ of period indicators for periods $p \in \{1, \ldots, P\}$, and the treatment-status indicator $D_{gp}$. Let $\lambda$ be the $G$-vector of group fixed effects, and $\alpha$ the $P$-vector of period fixed effects. The two-stage difference-in-differences estimator solves the sample analog of the moment condition

$$\begin{aligned}\mathbb{E}[f(\lambda, \alpha, \beta; W_{gpit})] &= \mathbb{E}\begin{bmatrix}[Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\alpha][(1(g)_{gpit}), (1(p)_{gpit})]'(1 - D_{gp}) \\ [Y_{gpit} - (1(g)_{gpit})'\lambda - (1(p)_{gpit})'\alpha - \beta D_{gp}]D_{gp}\end{bmatrix} \\ &= 0.\end{aligned}$$

By Theorem 6.1 of Newey and McFadden (1994, cf. Newey, 1984), and under standard regularity conditions, $\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, v)$, where $v$ is the last element of

$$\mathbb{E}\left[\frac{\partial f(\lambda, \alpha, \beta; Z_{gpit})}{\partial(\xi, \alpha, \beta)}\right]^{-1} \mathbb{E}[f(\lambda, \alpha, \beta; Z_{gpit}) f(\xi, \alpha, \beta; Z_{gpit})'] \mathbb{E}\left[\frac{\partial f(\xi, \alpha, \beta; Z_{gpit})}{\partial(\xi, \alpha, \beta)}\right]^{-1'}.$$

implement formal tests for parallel trends. Another approach is to replace treatment status $D_{gp}$ with the $r+1$-period lag $W_{rgp}$, $r \leq 0$, then use the two-stage procedure to estimate an $r+1$-period placebo ATT (as Liu, Wang and Xu, 2022, also note). Alternatively, Borusyak, Jaravel and Spiess (2023) recommend testing for parallel trends by including leads of treatment status in the first stage of the estimator, noting that their approach can sometimes circumvent concerns regarding conditioning difference-in-differences estimates on passing tests for parallel trends.

[18] Borusyak, Jaravel and Spiess (2023) provide an alternative derivation of the asymptotic distribution of the two-stage difference-in-differences and related imputation estimators.

The preceding expression can be used to manually correct the estimated second-stage variances for the use of a generated dependent variable. With modern statistical software, a simpler approach is to estimate both stages of the procedure simultaneously using a GMM routine.[19]

# 3 General theory for 2SDD

This section provides the theoretical results behind the main ideas presented in Section 2, and considers a more general setting with covariates and individual fixed effects, nesting Section 2 as a special case. We observe $\{(Y_{it}, X_{it}, D_{it})\}$, where $i \in \{1, 2, \cdots, N\}$ indexes individuals and $t \in \{1, 2, \cdots, T\}$ indexes time, so there are $NT$ observations in a balanced panel. Since indices $i$ and $t$ are sufficient to determine the group $g$ and period $p$, the indices $g, p$ are dropped to avoid notational clutter. The parallel trends assumption now takes the form

$$Y_{it} = \lambda_i + \alpha_t + D_{it}\beta_{it} + X_{it}'\gamma + \varepsilon_{it}, \tag{7}$$

where

$$\mathbb{E}\Big[\varepsilon_{it} \,\Big|\, \{D_{it}, X_{it}\}_{t=1}^{T}\Big] = 0. \tag{8}$$

Here, $D_{it}$ is an indicator for treatment status, $\beta_{it}$ is the heterogenous treatment effect, and $X_{it} \in \mathbb{R}^K$ is a vector of covariates. Since we consider setting where $N$ is large and $T$ is fixed, in a slight abuse of notation we now redefine $X_{it}$ to include time indicators, so that the vector $\gamma$ of coefficients on $X_{it}$ also includes time fixed effects. Note that, compared to the simplified setting in Section 2, (7) now includes individual fixed effects, the coefficient on $D_{it}$ is now $\beta_{it}$ and, accordingly, the error term is denoted by $\varepsilon_{it}$.

We also make the following substantive assumptions.

**Assumption 1.** Assume that:

[19] It is also possible to use the result of Theorem 6.1 of Newey and McFadden (1994, cf. Newey, 1984) to isolate the component of the variance matrix corresponding to the treatment effect estimate(s) (this approach is detailed in the Online Appendix).

1. (Parallel trends) Outcomes satisfy Equations (7) and (8).

2. For all $i$, there exists some $t$ where $D_{it} = 0$, and $\mathbb{E}\left[\sum_{t=1}^{T} D_{it}\right] > 0$.

3. Observations $\{(Y_{it}, D_{it}, X_{it})\}_{t=1}^{T}$ are independent and identically distributed over individuals $i$.

4. $T$ is fixed as $N \to \infty$.

Assumption 1.1 tells us that our model is correctly specified when heterogeneous treatment effects are accounted for. In the special case without covariates, and heterogeneity of $\beta_{gp}$ only at the group and period level, Assumption 1.1 reduces exactly to the parallel trends assumption stated in Section 2.1.[20] Assumption 1.2 requires everyone to be untreated for at least one period. As is standard in this environment, Assumption 1.3 does not require independence across time for a given individual, but requires independence over individuals. The assumption of i.i.d. data is unnecessary, but it is assumed for exposition and to avoid notational clutter.[21]

## 3.1 2SDD

Our general proposed procedure is:

1. Regress $Y_{it}$ on $X_{it}$ and individual fixed effects to obtain $\hat{\gamma}$ and $\hat{\lambda}_i$.

2. Regress adjusted outcomes $Y_{it} - \hat{\lambda}_i - X_{it}'\hat{\gamma}$ on $D_{it}$.

[20] To see this, using the notation in Section 2.1, the assumption becomes: $\mathbb{E}[Y_{it} - D_{gp}\beta_{gp} - \xi_g - \alpha_p \mid D_{gp}, g, p] = 0$, which implies:

$$\begin{aligned}\mathbb{E}[Y_{gpit} \mid g, p, D_{gp}] &= \lambda_g + \alpha_p + \beta_{gp} D_{gp} \\ &= \lambda_g + \alpha_p + \mathbb{E}[\beta_{gp} \mid D_{gp} = 1] D_{gp} + [\beta_{gp} - \mathbb{E}[\beta_{gp} \mid D_{gp} = 1]] D_{gp}.\end{aligned}$$

[21] Identical distribution of the triple does not contradict heterogeneous treatment effects, because the identical distribution of $Y_{it}$ can be attributed to the identical distribution of $\varepsilon_{it}$ instead of $\beta_{it}$. Heterogeneous treatment effects can also arise from variation in treatment timing even if the time- and duration-specific effects of the treatment are identically distributed.

Due to the FWL theorem, the first step of this procedure is equivalent to running the regression using data that have been transformed into deviations from individual means in the untreated sample. Consequently, we can recover the same $\hat{\gamma}$ and consequently $\hat{\beta}$, even though $\lambda_i$ are not consistently estimated.

To be precise, let $T_i^0 := \sum_{t=1}^{T} (1 - D_{it})$ denote the number of periods that individual $i$ is untreated, and define $\tilde{Y}_{it}$ as

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T_i^0} \sum_{t=1}^{T_i^0} Y_{it} (1 - D_{it}) .$$

Define $\tilde{\varepsilon}_{it}$ similarly, and let $\tilde{X}_{it}$ denote the matrix of deviations of the elements of $X_{it}$ from their individual untreated means. The equivalent procedure is:

1. Regress $\tilde{Y}_{0i}$ on $\tilde{X}_{0i}$ to obtain $\hat{\gamma}$.
2. Regress adjusted outcomes $\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\gamma}$ on $D_{it}$ to obtain $\hat{\beta}$.

Using $X_{kit}$ to denote regressor $k$ for individual $i$ at time $t$, $\tilde{X}_{0i}$ is a $T \times K$ matrix of the form

$$\tilde{X}_{0i} = \begin{bmatrix} \left(X_{1i1} - \frac{1}{T_i^0} \sum_{t=1}^{T} X_{1it} (1 - D_{it})\right) (1 - D_{i1}) & \cdots & \left(X_{Ki1} - \frac{1}{T_i^0} \sum_{t=1}^{T} X_{Kit} (1 - D_{it})\right) (1 - D_{i1}) \\ \vdots & & \\ \left(X_{1iT} - \frac{1}{T_i^0} \sum_{t=1}^{T} X_{1it} (1 - D_{it})\right) (1 - D_{iT}) & \cdots & \left(X_{KiT} - \frac{1}{T_i^0} \sum_{t=1}^{T} X_{Kit} (1 - D_{it})\right) (1 - D_{iT}) \end{bmatrix} .$$

Similarly, $\tilde{Y}_{0i}$ is a $T \times 1$ vector of the form

$$\tilde{Y}_{0i} = \begin{bmatrix} \tilde{Y}_{i1} (1 - D_{i1}) & \cdots & \tilde{Y}_{iT} (1 - D_{iT}) \end{bmatrix}' .$$

The coefficient estimator from the first stage regression is then

$$\hat{\gamma} = \left( \sum_{i=1}^{N} \tilde{X}_{0i}' \tilde{X}_{0i} \right)^{-1} \left( \sum_{i=1}^{N} \tilde{X}_{0i}' \tilde{Y}_{0i} \right) .$$

Observe that in both sums, we have a sum over independent individuals $i$ (the sum over time is already implicit in the matrix multiplication). The second stage-regression

is done without a constant, because the data are already demeaned. Hence, the two-stage difference in difference estimator is

$$\hat{\beta} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \left( \tilde{Y}_{it} - \tilde{X}_{it} \hat{\gamma} \right) \right).$$

We define $\beta$ as the limiting average treatment effect on the treated (ATT):

$$\beta := \mathbb{E}[\beta_{it} \mid D_{it} = 1] = \operatorname{plim} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \beta_{it} \right).$$

The estimators can be written as the solution to the following GMM problem:

$$\mathbb{E} \left[ \begin{array}{c} \tilde{X}_{0i}' \left( \tilde{Y}_{0i} - \tilde{X}_{0i} \gamma \right) \\ \sum_{t=1}^{T} D_{it} \left( \tilde{Y}_{it} - \tilde{X}_{it}' \gamma - \beta D_{it} \right) \end{array} \right] = 0$$

so we have $K+1$ moment conditions, with $K$ in the first stage and one in the second stage. The moment condition reduces to that of Section 2.6 as a special case.

**Assumption 2.** Assume that:

1. $\mathbb{E}\left[ \left\| \tilde{X}_{0i}' \tilde{\varepsilon}_{0i} \right\|^2 \right] < \infty$, $\mathbb{E}[\tilde{\varepsilon}_{it}^2] < \infty$, and $\mathbb{E}\left[ (\beta_{it} - \beta)^2 \right] < \infty$.
2. $\mathbb{E}\left[ \tilde{X}_{0i}' \tilde{X}_{0i} \right]$ is invertible and $\mathbb{E}\left[ \left\| \tilde{X}_{0i}' \tilde{X}_{0i} \right\|^2 \right] < \infty$.

**Theorem 1.** *If Assumptions 1 and 2 hold, then $\hat{\gamma}$ and $\hat{\beta}$ are asymptotically normal, $\hat{\gamma} \xrightarrow{p} \gamma$, $\hat{\beta} \xrightarrow{p} \beta$, and $\sqrt{NT} \left( \hat{\beta} - \beta \right) \xrightarrow{d} N(0, V)$, where $V = G_\beta^{-1} \mathbb{E}\left[ (g + G_\gamma \psi)(g + G_\gamma \psi)' \right] G_\beta^{-1\prime}$, with:*

$$\begin{aligned} G_\beta &= -\mathbb{E}\left[ \sum_{t=1}^{T} D_{it} \right] \\ G_\gamma &= -\mathbb{E}\left[ \sum_{t=1}^{T} D_{it} \tilde{X}_{it} \right] \\ \psi &= \mathbb{E}\left[ \tilde{X}_{0i}' \tilde{X}_{0i} \right]^{-1} \left( \tilde{X}_{0i}' \left( \tilde{Y}_{0i} - \tilde{X}_{0i} \gamma \right) \right) \end{aligned}$$

Theorem 1 tells us that the 2SDD estimator is consistent for $\beta$, and is asymptotically normal. Hence, using a consistent variance estimator provides valid inference asymptotically. The proof proceeds by the arguments in Section 2 and verifying the conditions in Newey and McFadden (1994).

## 3.2 Event Studies

As we note in Section 2.5, there are multiple ways to implement event-studies using the two stage approach. All of these variations can be understood from within the following framework. Let $W_{rit} = 1[t - t^*(i) = r]$ denote whether individual $i$ is $r$ periods away from treatment at time $t$. With slight abuse of notation, our model is:

$$\begin{aligned} Y_{it} &= \lambda_i + \sum_{r=-\underline{R}}^{\overline{R}} \eta_{rit} W_{rit} + X_{it}'\gamma + \varepsilon_{it} \\ &= \lambda_i + W_{it}'\eta_{it} + X_{it}'\gamma + \varepsilon_{it}. \end{aligned}$$

The second equality comes from stacking the $\underline{R} + \overline{R} + 1$ $W_{rit}$ and $\eta_{rit}$ objects. Notice now that $\eta_{it}$ is a vector with $\eta_{rit}$ as its components. In the first stage, we have $Y_{it} = \lambda_i + X_{it}'\gamma + \varepsilon_{it}$. This regression uses all observations with $t - t^*(i) < -\underline{R}^*$, where $\underline{R}^*$ may be zero. Let $Q_{it} := 1[t - t^*(i) < -\underline{R}]$. Then, define the analogous objects to 2SDD as $T_i^Q := \sum_{t=1}^T Q_{it}$. Now,

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T_i^Q} \sum_{t=1}^{T_i^Q} Y_{it} Q_{it},$$

and a similar definition applies to $\tilde{X}_{it}$ and $\tilde{\varepsilon}_{it}$. Analogously,

$$\tilde{X}_{Qi} = \begin{bmatrix} \left(X_{1i1} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{1it}Q_{it}\right) Q_{i1} & \cdots & \left(X_{Ki1} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{Kit}Q_{it}\right) Q_{i1} \\ \vdots & & \\ \left(X_{1iT} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{1it}Q_{it}\right) Q_{iT} & \cdots & \left(X_{KiT} - \frac{1}{T_i^Q}\sum_{t=1}^T X_{Kit}Q_{it}\right) Q_{iT} \end{bmatrix},$$

and

$$\tilde{Y}_{Qi} = \left[\tilde{Y}_{i1}Q_{i1} \quad \cdots \quad \tilde{Y}_{iT}Q_{iT}\right].$$

In this environment, our analogous two-stage procedure is:

1. Regress $\tilde{Y}_{Qi}$ on $\tilde{X}_{Qi}$ to obtain $\hat{\gamma}$.
2. Regress adjusted outcomes $\tilde{Y}_{it} - \tilde{X}'_{it}\hat{\gamma}$ on $W_{it}$ to obtain $\hat{\eta}$.

Hence, the estimators are:

$$\hat{\gamma} = \left(\frac{1}{N}\sum_i \tilde{X}'_{Qi}\tilde{X}_{Qi}\right)^{-1}\left(\frac{1}{N}\sum_i \tilde{X}'_{Qi}\tilde{Y}_{Qi}\right),$$

and

$$\hat{\eta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}\right)\right).$$

The object of interest is now:

$$\eta = \text{plim}\begin{bmatrix} \frac{1}{N_{-\underline{R}}}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t - t^*(i) = -\underline{R}\right]\eta_{-\underline{R}it} \\ \vdots \\ \frac{1}{N_{\overline{R}}}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t - t^*(i) = \overline{R}\right]\eta_{\overline{R}it} \end{bmatrix},$$

where $N_r := \sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t - t^*(i) = r\right]$ is the number of individuals whom we can get an observation when she was $r$ periods away from treatment. Then, every element $\eta_r = \text{plim}\,\frac{1}{N_r}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t - t^*(i) = r\right]\eta_{rit}$ is the average coefficient across individuals who are observed $r$ periods away from their treatment. For $r > 0$, $\eta_r$ can be interpreted as the treatment effect $r$ periods after treatment. If there are no pre-trends, $\eta_r = 0$ for all $r \leq 0$.

As before, the estimators can be written as the solution to a GMM problem:

$$\mathbb{E}\begin{bmatrix} \tilde{X}'_{0i}\left(\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma\right) \\ \sum_{t=1}^{T} W_{it}\left(\tilde{Y}_{it} - \tilde{X}'_{it}\gamma - W'_{it}\beta_{it}\right) \end{bmatrix} = 0$$

For the asymptotics to go through as before, we want a condition analogous to Assumption 2 that is suited for event studies.

**Assumption 3.** Assume that:

1. $\mathbb{E}\left[\left\|\tilde{X}'_{Qi}\tilde{\varepsilon}_{0i}\right\|^2\right] < \infty$, $\mathbb{E}[\tilde{\varepsilon}^2_{it}] < \infty$, and $\mathbb{E}\left[(\eta_{rit} - \eta_r)^2\right] < \infty$.
2. $\mathbb{E}\left[\tilde{X}'_{Qi}\tilde{X}_{Qi}\right]$ is invertible and $\mathbb{E}\left[\left\|\tilde{X}'_{Qi}\tilde{X}_{Qi}\right\|^2\right] < \infty$.
3. For all $i$, there exists some $t$ where $Q_{it} = 0$, and $\mathbb{E}[N_{ir}] > 0$ for all $r \in \{-\underline{R}, \cdots, \overline{R}\}$, where $N_{ir} := \sum_{t=1}^{T} 1\left[t - t^*(i) = r\right]$.

Note that in event studies, we are regressing on $1\left[t - t^*(i) = r\right]$ in the second stage. Assumption 3.2 is required for invertibility in the second stage, playing the role of Assumption 1.3.

**Theorem 2.** *If Assumptions 1 and 3 hold, then $\hat{\gamma}$ and $\hat{\eta}$ are asymptotically normal, $\hat{\gamma} \xrightarrow{p} \gamma$ and $\hat{\eta} \xrightarrow{p} \eta$, and $\sqrt{NT}\left(\hat{\eta} - \eta\right) \xrightarrow{d} N(0, V)$, where $V = G_\eta^{-1}\mathbb{E}\left[\left(g + G_\gamma\psi\right)\left(g + G_\gamma\psi\right)'\right]G_\eta^{-1\prime}$, with:*

$$
\begin{aligned}
G_\eta &= -\mathbb{E}\left[\sum_{t=1}^{T} W_{it}W'_{it}\right] \\
G_\gamma &= -\mathbb{E}\left[\sum_{t=1}^{T} W_{it}\tilde{X}'_{it}\right] \\
\psi &= \mathbb{E}\left[\tilde{X}'_{Qi}\tilde{X}_{Qi}\right]^{-1}\left(\tilde{X}'_{Qi}\left(\tilde{Y}_{Qi} - \tilde{X}_{Qi}\gamma\right)\right)
\end{aligned}
$$

.

# 4 Rejection rates for randomly generated interventions

This section conducts Monte Carlo simulation exercises inspired by Bertrand, Duflo and Mullainathan (2004) to evaluate the two-stage approach and provide insight

into how various difference-in-differences (DD) methods perform under realistic conditions. First, we aim to assess finite-sample performance in environments that resemble common empirical applications. Second, acknowledging that theoretical frameworks often rely on the assumption of i.i.d. data, we simulate scenarios that incorporate autocorrelation and reflect real-world datasets more accurately. Third, the proliferation of recently proposed alternatives for DD estimation necessitates a comparative analysis to discern their relative strengths and weaknesses. Lastly, since the Borusyak, Jaravel and Spiess (2023) method shares point estimates with ours, it becomes essential to assess the distinct approaches to inference.

## 4.1 Data and methodology

Our primary dataset consists of wage data for women between the ages of 25 and 50 from the Current Population Survey (CPS). We define wage as the natural logarithm of weekly earnings, which are recorded in the fourth interview month in the Merged Outgoing Rotation Group of the CPS.[22] The data span a 42-year period from 1979 to 2020 and contain over one million women reporting strictly positive weekly earnings. We obtain three datasets based on this for our Monte Carlo exercises. We construct a state-by-year panel dataset comprising average wages in 2,100 state-year cells. In addition, we generate an i.i.d. dataset with similar features.

Our simulation study adopts a "random design" strategy. This approach introduces stochasticity by randomly drawing treated states, treatment effects, and treatment timing in each iteration. By doing so, we create a more realistic representation of real-world scenarios where the assignment of treatments may not follow a fixed pattern (Athey and Imbens, 2022). Importantly, we also document the inherent limitations of considering treatment and treatment timing as non-stochastic as in the "fixed design" approach of Borusyak, Jaravel and Spiess (2023).

To simulate a staggered treatment setting, we randomly assign states to the

[22]Using the logarithmic transformation excludes women with zero weekly earnings. While many recent papers use quasi-logarithmic transformations to incorporate zero-valued observations, Thakral and Tô (2023) document substantial biases arising from the use of such transformations, and thus we focus on women with strictly positive earnings following Bertrand, Duflo and Mullainathan (2004).

treatment group and generate treatments that occur randomly over a specified period. This contrasts with the original exercise by Bertrand, Duflo and Mullainathan (2004), in which the placebo treatment timing is homogeneous across treated states and drawn uniformly at random. In all cases, we restrict the earliest treatment year to 1982 and the latest treatment year to 2014, ensuring that each state is treated for at least 5 periods.

We estimate the effects of the randomly generated interventions using the two-stage approach (with our analytical standard error) as well as a number of alternative methods for comparison. In particular, we consider the imputation approach from Borusyak, Jaravel and Spiess (2023), using both their "default" asymptotically conservative standard errors and "leave-out" version with improved finite-sample performance, as well as various estimators that use different forms of bootstrap to compute standard errors (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).[23] Standard errors are adjusted for clustering at the state level, following Bertrand, Duflo and Mullainathan (2004).

## 4.2 Simulation results

We conduct an event-study analysis to estimate the effect of the randomly generated interventions in each of the five years starting from the time of treatment. The primary measure we use to evaluate the performance of each method is the relative frequency of rejecting the null hypothesis of the true effect size at the 5 percent significance level over 500 simulations. We also report the mean bias, root-mean-square error (RMSE), and average per-simulation computational speed.

The baseline environment consists of states being treated over a 20-year period, which corresponds to an empirical example highlighted in the recent Miller (2023) guide to event study models (the impact of state-level school finance reforms in 26 states from 1990–2011 from Lafortune, Rothstein and Schanzenbach, 2018). However, we consider 40 treated states in our baseline environment and ensure at least 2 treated

[23] We conduct these analyses in Stata using the packages `did2s`, `did_imputation`, `did_multiplegt`, `csdid`, and `eventstudyinteract`.

states per year, with the goal of providing the Borusyak, Jaravel and Spiess (2023) approach to inference with a more balanced assessment. In particular, computing their leave-out variance estimator requires that no treatment cohort consists only of a single state. Treatment effects are heterogeneous and drawn from a normal distribution, with an average value randomly drawn between 10 percent and 50 percent of the average wage and a standard deviation equal to 50 percent of the average wage.

Table 1 reports results from the baseline environment, in which the average true effect is approximately 2. Our proposed two-stage method with the GMM approach to inference leads to rejection rates near 5 percent, with standard errors around 0.61. Despite having the same point estimates, the default Borusyak, Jaravel and Spiess (2023) variance estimator leads to the most substantial levels of over-rejection, ranging from 8.6 percent to 22.9 percent, with standard errors around 0.45. Their leave-out variance estimator, on the other hand, leads to overly conservative estimates, with rejection rates around 1 percent and standard errors around 0.84. Compared to the leave-out variance estimator, the Sun and Abraham (2021) method leads to similar rejection rates with a larger standard error (around 0.98) and the Callaway and Sant'Anna (2021) method results in similar standard errors (around 0.84) but achieves a rejection rate closer to 5 percent. The De Chaisemartin and d'Haultfoeuille (2020) estimator also achieves rejection rates near 5 percent, with standard errors around 0.70.

The two-stage approach and the imputation approach share a speed advantage, outperforming the alternatives by a factor of 100 or more. This highlights the simplicity of the two-stage estimator, which can be computed straightforwardly using OLS regressions, and the value of having analytical standard errors based on the familiar GMM approach to inference.

To further evaluate these methods, we proceed to vary the minimum number of treated states in each year, the number of years during which the treatment can occur, and the total number of treated states. We then extend our analysis to environments with homogeneous treatment effects and i.i.d. data.

### 4.2.1 Size of treatment cohorts

Many datasets, such as the example from Lafortune, Rothstein and Schanzenbach (2018), have the feature that treatment cohorts may consist of only a single treated unit. The results in Table 2 remove the restriction that at least two states must be treated in each period. In this case, the leave-out variance estimator from Borusyak, Jaravel and Spiess (2023) can no longer be computed. Aside from that, removing the restriction leads to similar results for all methods, though with slightly higher standard errors (Table 2). With the (overly) conservative leave-out option no longer available, over-rejection becomes a significant concern with the imputation approach.

### 4.2.2 Number of treatment cohorts

Table 3 shows how the results change after increasing the number of treatment cohorts to 30 from the baseline of 20. This change has little effect on two-stage approach and the Callaway and Sant'Anna (2021) estimator, with both leading to similar rejection rates (near 5 percent) and standard errors (around 0.61 for 2SDD and around 0.84 for CS) as before. The Sun and Abraham (2021) standard error also changes little (a slight increase to around 1.01) and leads to similar rates of under-rejection as before. In contrast, the default Borusyak, Jaravel and Spiess (2023) variance estimator leads to even more severe over-rejection rates than before, ranging from 20 percent to 30 percent, with much smaller standard errors of around 0.34. In this case, the leave-out variance estimator cannot be computed.

Decreasing the number of treatment cohorts to 15 similarly has little effect on the performance of the two-stage approach, the Callaway and Sant'Anna (2021) estimator, and the Sun and Abraham (2021) estimator, as Tables 4 and 5 show. The default Borusyak, Jaravel and Spiess (2023) variance estimator continues to lead to over-rejection, though with a rejection rate of only around 12 percent, and the under-rejection rates of the leave-out variance estimator improve slightly as the standard errors decrease to around 0.78.

Overall, these results highlight the anti-conservativeness of the default imputation

approach to inference. This can be attributed to over-fitting in finite samples.[24] This observation also explains why the imputation default performs poorly when the number of groups increases relative to $N$.[25] Due to over-fitting when the groups size is small, the extent of over-rejection using their approach becomes more severe if treatment timing is staggered over a longer period. In practice, we find evidence of over-rejection using their default variance estimator even if the treatment is staggered over only 5 periods (see Table 6).

### 4.2.3 Number of treatment units

The baseline environment consists of 40 treated states. However, as noted earlier, many empirical examples such as the Lafortune, Rothstein and Schanzenbach (2018) setting consist of fewer treated units (26 states in that case). Thus we examine the consequences of decreasing the number of treated units from 40 to 30 or 20. Before proceeding, we note that Borusyak, Jaravel and Spiess (2023) suggest a minimum effective number of treated observations of 30 because, as their documentation states, "inference on coefficients which are based on a small number of observations is unreliable" (see the Herfindahl condition in their paper). Nevertheless, given the prevalence of empirical examples with smaller numbers of treated units, we evaluate the performance of the various methods in such settings to shed light on their relative strengths and weaknesses.

To hold fixed the number of treatment cohorts while ensuring that the Borusyak, Jaravel and Spiess (2023) leave-out variance estimator can be computed even when the number of treated states is only 20, we consider settings with 10 treatment cohorts.

[24] The standard errors for the imputation estimator developed in Borusyak, Jaravel and Spiess (2023) are constructed based on the residuals $\tilde{\varepsilon}_{it} = \hat{\tau}_{it} - \hat{\bar{\tau}}_{it}$, where $\hat{\tau}_{it}$ is the estimated treatment effect for unit $i$ at time $t$ and $\hat{\bar{\tau}}_{it}$ is some average of these estimated individual treatment effects. Their "default" is to use cohort-period averages for $\hat{\bar{\tau}}_{it}$, i.e., $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}}_{gp}$. However, by using cohort-period averages, they are partway to the degenerate limit of zero variance. Hence, the variance estimator in Borusyak, Jaravel and Spiess (2023) is anti-conservative when the groups are small: in the extreme case, $\hat{\bar{\tau}}_{it} = \hat{\tau}_{it}$, so $\tilde{\varepsilon}_{it} = 0$.

[25] Since their default is to use $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}}_{gp}$, as $G$ increases, the groups become finer, so $\tilde{\varepsilon}_{it} \to 0$, which underestimates the variance. This problem is avoided if the imputation method were to use the largest group available, where $\hat{\bar{\tau}}_{it} = \hat{\bar{\tau}} = \hat{\beta}$, as GMM does.

When decreasing the number of treated states from 40 (Table 7) to 30 (Table 8) and further to 20 (Table 9), the standard error appreciably increases, and the resulting rejection rates remain stable, in all cases except for the default Borusyak, Jaravel and Spiess (2023) variance estimator. The simulation results suggest that most difference-in-difference methods may still apply reliably in empirical settings with smaller numbers of treated units and furthermore highlights an important advantage of the GMM approach to inference.

### 4.2.4 Homogeneous treatment effects

While the possibility of misspecification under the TWFE regression model in situations with heterogeneous treatment effects motivates the development of alternative methods for DD estimation (see Section 2.2), the case of homogeneous treatment effects provides a useful benchmark for comparing different methods. The various alternative approaches eliminate bias that arise when estimating average treatment effects in the presence of treatment effect heterogeneity with staggered treatment timing. A natural question, however, is whether the reduction in bias comes at the cost of substantially increasing the variance even when the TWFE model is correctly specified.

We therefore conduct a set of simulations in which treatment effects are homogeneous across units and time periods. In these simulations, the normal distribution from which treatment effects are drawn has an average value equal to 30 percent of the average wage, the midpoint of the range from before.

When treatment effects are homogeneous, we find that the two-stage approach performs almost as well as TWFE, as Table 10 shows. Both methods achieve rejection rates around 5 percent, though TWFE gives slightly smaller standard errors (about 0.60 instead of 0.62). Since homogeneous treatment effects is a special case of our setup, the 2SDD estimand converges to the true treatment effect $\beta$, which is the same limit as TWFE. Due to the FWL theorem, the estimator $\hat{\beta}_{\text{TWFE}}$ is numerically identical to what we would get if we regressed $Y_{it}$ on $1(g)'_{it}, 1(p)'_{it}$ and $D_{it}$ on $1(g)'_{it}, 1(p)'_{it}$ for all observations, then regressing the residual of $Y_{it}$ on the

residual of $D_{it}$. In the first stage, $\hat{\lambda}_{\text{TWFE}} = \lambda(1 + o_P(1)), \hat{\alpha}_{\text{TWFE}} = \alpha(1 + o_P(1))$, when there are homogeneous treatment effects. The 2SDD approach is similar, except that the first stage regression uses only the untreated observations, so $\hat{\lambda}_{\text{2SDD}} = \xi(1 + o_P(1)), \hat{\alpha}_{\text{2SDD}} = \alpha(1 + o_P(1))$. Then, asymptotically, the residual generated in both procedures will be $\tilde{Y}_{it} = Y_{it} - \hat{\xi}'1(g)_{it} - \hat{\alpha}'1(p)_{it} = Y_{it} - \xi'1(g)_{it} - \alpha'1(p)_{it} + o_P(1)$. 2SDD and TWFE hence only differ in the second stage: TWFE regresses $\tilde{Y}_{it}$ on the residual of $D_{it}$ while 2SDD regresses $\tilde{Y}_{it}$ on $D_{it}$. Since both estimators converge to the same limit, the only difference in inference is the variance.

The other methods, however, are substantially outperformed by TWFE. The Borusyak, Jaravel and Spiess (2023) default variance estimator gives much smaller standard errors of around 0.45, about 25 percent smaller than under TWFE, leading to rejections of the null hypothesis of the true effect about three times as often as under TWFE. The Borusyak, Jaravel and Spiess (2023) leave-out variance estimator (standard error 0.84) and the Sun and Abraham (2021) approach (standard error 0.99) reject only 20–40 percent as often as TWFE. The Callaway, Goodman-Bacon and Sant'Anna (2021) estimator yields similar rejection rates as our approach and TWFE, but with a relatively large standard error of around 0.85. The two-stage approach, in comparison, provides the most natural way to extend DD estimation to achieve robustness to treatment effect heterogeneity without much efficiency loss.

#### 4.2.5 I.i.d. data

The data that we use for our primary simulation exercises exhibit realistic features such as higher-order serial correlation. However, we note that the advantages of the two-stage approach do not rely on this particular feature of the data. We show this by conducting the much simpler exercise of generating i.i.d. data and comparing the performance of the different estimators.

All of our conclusions persist in the i.i.d. environment. The baseline environment (Table 11) continues to show rejection rates close to 5 percent for the two-stage approach and the Callaway, Goodman-Bacon and Sant'Anna (2021) estimator, with the standard error for the latter being nearly 40 percent larger. Also as before, the

default Borusyak, Jaravel and Spiess (2023) variance estimator leads to over-rejection (rejection rates ranging from 13.8 percent to 18.6 percent), while their leave-out variance estimator is overly conservative (rejection rates ranging from 1 percent to 2 percent), as is the Sun and Abraham (2021) estimator. The same holds in the simple case of homogeneous treatment effects (Table 12). The comparison between Tables 13 and 14 shows, as before, that increasing the number of treatment cohorts leads to smaller standard errors for all methods but keeps rejection rates constant for all except the default Borusyak, Jaravel and Spiess (2023) variance estimator, for which rejection rates become about 30 percent. Analogously, the comparison between Tables 15 to 17 shows, as before, that decreasing the number of treated states leads to notably larger standard errors and hence similar rejection rates for all methods except the default Borusyak, Jaravel and Spiess (2023) variance estimator, for which rejection rates increase from 8–10 percent to 18–22 percent as the number of treated states decreases from 40 to 20.

# 5 Conclusion

When adoption of a treatment is staggered across time, and the average effects of the treatment vary by group and period, the usual difference-in-differences regression specification does not identify an easily interpretable measure of the typical effect of the treatment. When the duration-specific effects are also heterogeneous, neither do the coefficients from the usual event-study specification. The ultimate source of these identification failures is that outcomes are not necessarily linear in group, period and treatment status, as difference-in-differences and event-study regression specifications assume.

The two-stage approach developed in this paper is motivated by the observation that, under parallel trends, untreated outcomes are linear in group and period effects. Those effects are therefore identified from a first-stage regression estimated using the sample of untreated observations. The average effect of the treatment on the treated is then identified from a regression of outcomes on treatment status, after removing group and period effects. This procedure is robust to the presence of heterogeneous

treatment effects when treatment adoption is staggered. Estimation and inference are simple and intuitive, and can be extended to identify a variety of different treatment effect measures. Monte Carlo simulations show that the two-stage estimators correctly identify informative average treatment effect measures, outperforming alternative estimators that are also more difficult to implement.

# A Extensions

## A.1 Additional estimands

Our procedure can be extended to any linear combination of treatment effects when some assumptions hold. Recall that we have the model $Y_{it} = D_{it}\beta_{it} + X'_{it}\gamma + \varepsilon_{it}$ with $E\left[\varepsilon_{it} | \{D_{it}, X_{it}\}_{t=1}^{T}\right] = 0$. This model implies that $E\left[Y_{it} - D_{it}\beta_{it} - X'_{it}\gamma\right] = 0$. We are interested in $\tau := w_{it}\beta_{it}$, where $w_{it}$ is a nonstochastic weight. Due to the moment condition, and $w_{it}$ being nonstochastic,

$$E\left[w_{it}Y_{it} - w_{it}X'_{it}\gamma\right] - E\left[D_{it}\right]w_{it}\beta_{it} = 0$$

Assume that heterogeneity in $E\left[D_{it}\right]$ occurs at some level $h$, and $\zeta$ is the vector of values it can take, so that $E\left[D_{it}\right] = 1(h)'_{it}\zeta$. Assume that $\zeta$ is either known or can be consistently estimated, and all elements of $\zeta$ are nonzero. Then, by summing $w_{it}\beta_{it} = E\left[w_{it}Y_{it} - w_{it}X'_{it}\gamma\right]/E[D_{it}]$ over $i, t$:

$$\tau = \sum_{i,t} w_{it}\beta_{it} = \sum_{i,t} w_{it} E\left[\frac{Y_{it} - X'_{it}\gamma}{1(h)'_{it}\zeta}\right]$$

Hence, writing everything as a system of moment conditions,

$$E\left[\begin{array}{c} 1(h)_{it}\left(D_{it} - 1(h)'_{it}\zeta\right) \\ X_{it}\left(1 - D_{it}\right)\left(Y_{it} - X'_{it}\gamma\right) \\ \tau - w_{it}\left(\frac{Y_{it} - X'_{it}\gamma}{1(h)'_{it}\zeta}\right) \end{array}\right] = 0$$

The just-identified system of equations allows GMM to be applied as before.

## A.2 Design-based analysis

The 2SDD estimand is also interpretable in a design-based world. Let $A_{1i}$ denote the number of periods that individual $i$ has been treated, with $A_{1i} = 1$ on the period that

$i$ was first treated. Further, assume that $\beta_{it} = \beta_i$ for all $t$. Define

$$\beta := \text{plim} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it} \beta_{it} \right)$$

Then,

$$\beta = \text{plim} \left( \sum_{i=1}^{N} A_i \right)^{-1} \left( \sum_{i=1}^{N} A_i \beta_i \right)$$

Under the Athey and Imbens (2022) setup where the adoption time of treatment is as good as random, $A_i$ is randomly assigned across individuals in our setting. Then, $\frac{1}{N} \sum_{i=1}^{N} A_i \xrightarrow{p} a$, and $\mathbb{E}[A_i] = a$ for all $i$. Further,

$$\frac{1}{aN} \mathbb{E} \left[ \sum_{i=1}^{N} A_i \beta_i \right] = \frac{1}{aN} \sum_{i=1}^{N} \mathbb{E}[A_i] \beta_i = \frac{1}{N} \sum_{i=1}^{N} \beta_i$$

which is exactly the average treatment effect (ATE).

# B Proofs

*Derivation of Equation* (3). From (1), we can write

$$Y_{gpit} = \xi_g + \alpha_p + \sum_{h=1}^{G} \sum_{q=h}^{P} \beta_{hq} 1(h,q)_{gpit} + e_{gpit}, \tag{9}$$

where $1(h,q)_{gpit}$ is an indicator for whether observation $(g,p,i,t)$ corresponds to group $h$ and period $q$, and $\mathbb{E}[e_{gpit} \mid g, p, (1(h,q)_{gpit})] = 0$.

Let $\tilde{D}_{gp}$ denote the residual from a population regression of $D_{gp}$ on group and period fixed effects. By the Frisch-Waugh-Lovell theorem, the coefficient on $D_{gp}$ from

a population regression of $Y_{gpit}$ on $D_{gp}$ and group and period effects is

$$\begin{aligned}
\beta^* &= \frac{\mathbb{E}\left[\tilde{D}_{gp} Y_{gpit})\right]}{\mathbb{E}\left[\tilde{D}_{gp}^2\right]} \\
&= \frac{\mathbb{E}\left[\tilde{D}_{gp} \sum_{h=1}^{G} \sum_{q=h}^{P} \beta_{hq} 1(h,q)_{gpit}\right]}{\mathbb{E}\left[\tilde{D}_{gp}^2\right]} \\
&= \sum_{h=1}^{G} \sum_{q=h}^{P} \frac{\mathbb{E}\left[\tilde{D}_{gp} 1(h,q)_{gpit}\right]}{\mathbb{E}\left[\tilde{D}_{gp}^2\right]} \beta_{hq} \\
&= \sum_{g=1}^{G} \sum_{p=g}^{P} \omega_{gp} \beta_{gp}.
\end{aligned}$$

where $\omega_{gp}$ is the coefficient from a regression of $1(h,q)_{gpit}$ on $D_{gp}$ and group and period fixed effects. The second equality uses the facts that $e_{gpit}$ is mean-independent of the regressors and that $\tilde{D}_{gp}$ is uncorrelated with group and period effects by construction.[26]

The weight $\omega_{gp}$ that difference in differences places on $\beta_{gp}$ is the coefficient on $D_{gp}$ from a regression of $1(g,p)_{gpit}$ on $D_{gp}$ and group and period fixed effects. By the Frisch-Waugh-Lovell theorem, this is equivalent to the slope coefficient from a population regression of $1(g,p)_{gpit}$ on the residual from an auxiliary regression of $D_{gp}$ on group and period effects. Using the two-way within or double-demeaned transformation, this residual can be expressed as

$$\tilde{D}_{gp} = [D_{gp} - \Pr(D_{gp} = 1 \,|\, g)] - [\Pr(D_{gp} = 1 \,|\, p) - \Pr(D_{gp} = 1)]. \tag{10}$$

[26]This, and the related result in Sun and Abraham (2021), can also be established by thinking of the term $\sum_{h=1}^{G} \sum_{q=h}^{P} \beta_{hq} 1(h,q)_{gpit}$ in (9) as an omitted variable, and taking its projection onto the included regressors.

Since $\mathbb{E}\left[\tilde{D}_{gp}^2\right] = \mathbb{E}\left[\tilde{D}_{gp} D_{gp}\right]$, $\omega_{gp}$ can also be expressed as

$$\begin{aligned}
\omega_{gp} &= \frac{\mathbb{E}\left[1(g,p)_{gpit}\tilde{D}_{gp}\right]}{\operatorname{Var}\left[\tilde{D}_{gp}\right]} \\
&= \frac{\mathbb{E}\left[\tilde{D}_{gp} \,\middle|\, 1(g,p)_{gpit} = 1\right] \Pr(1(g,p)_{gpit} = 1)}{\mathbb{E}\left[\tilde{D}_{gp} \,\middle|\, D_{gp} = 1\right] \Pr(D_{gp} = 1)} \\
&= \frac{[1 - \Pr(D_{gp} = 1 \,|\, g) - (\Pr(D_{gp} = 1 \,|\, p) - \Pr(D_{gp} = 1))] \Pr(g,p)}{\sum_{g'=1}^{G} \sum_{p'=g'}^{P} [1 - \Pr(D_{g'p'} = 1 \,|\, g') - (\Pr(D_{g'p'} = 1 \,|\, p') - \Pr(D_{g'p'} = 1))] \Pr(g',p')},
\end{aligned}$$

where the final equality uses (10). □

**Lemma 1.** *Under Assumption 1 and Assumption 2, $\hat{\gamma} \xrightarrow{p} \gamma$ and $\hat{\beta} \xrightarrow{p} \beta$.*

*Proof of Lemma 1.* We have

$$\tilde{Y}_{0i} - \tilde{X}_{0i}\gamma = \begin{bmatrix} (\tilde{\varepsilon}_{i1})\,(1 - D_{i1}) \\ \vdots \\ (\tilde{\varepsilon}_{iT})\,(1 - D_{iT}) \end{bmatrix} =: \tilde{\varepsilon}_{0i}$$

Hence,

$$\hat{\gamma} = \gamma + \left(\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{X}_{0i}\right)^{-1} \left(\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right)$$

Due to Assumption 1.3, and the existence of second moments in Assumption 2.2, by the weak law of large numbers (WLLN), $\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{X}_{0i} \xrightarrow{p} \mathbb{E}\left[\tilde{X}_{0i}'\tilde{X}_{0i}\right]$. With Assumption 1.1 on correct specification and the WLLN, $\mathbb{E}\left[\tilde{X}_{0i}'\tilde{\varepsilon}_{0i}\right] = 0$. Hence, $\frac{1}{N}\sum_i \tilde{X}_{0i}'\tilde{\varepsilon}_{0i} \xrightarrow{p} 0$. Then, by the continuous mapping theorem, $\hat{\gamma} \xrightarrow{p} \gamma$.

The OLS estimator is:

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}\hat{\gamma}\right)\right) \\
&= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\beta_{it}D_{it} + \tilde{\varepsilon}_{it} - \tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)\right)\right) \\
&= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\beta_{it}\right) + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{\varepsilon}_{it} - \tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)\right)\right)
\end{aligned}$$

Due to Assumption 2.1 and Assumption 2.2, $\left|\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\tilde{X}_{it}\right)\right|$ is bounded in probability, so, using the definition of $\beta$, and the first-stage consistency result that $\hat{\gamma} \xrightarrow{p} \gamma$,

$$\begin{aligned}
\hat{\beta} &= \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{\varepsilon}_{it} - \tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)\right)\right) + o_P(1) \\
&= \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\tilde{\varepsilon}_{it}\right) + o_P(1).
\end{aligned}$$

Due to Assumption 1.2 and Assumption 1.3, $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it} \xrightarrow{p} \Sigma_D > 0$. $D_{it}\tilde{\varepsilon}_{it}$ are also independent over individuals. Using a similar argument as before, $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\tilde{\varepsilon}_{it} \xrightarrow{p} 0$. Then, $\hat{\beta} = \beta + o_P(1)$ as required. □

*Proof of Theorem 1.* If the conditions of Theorem 6.1 of Newey and McFadden (1994) are satisfied, the result automatically follows. Hence, the proof verifies its conditions. Due to the consistency, we already have $\hat{\gamma} \xrightarrow{p} \gamma$ and $\hat{\beta} \xrightarrow{p} \beta$, fulfilling the probability limit requirement. Next, we want to show the following:

1. $\beta$ is in the interior of the parameter space.
2. $g(Z;\gamma,\beta)$ is continuously differentiable around $\beta$.
3. $\mathbb{E}[g(Z;\gamma,\beta)] = 0$ and $\mathbb{E}\left[\|g(Z;\gamma,\beta)\|^2\right]$ is finite.

4. $\mathbb{E}\Big[sup_{(\gamma,\beta)} \|\nabla g(Z;\gamma,\beta)\|\Big] < \infty$, where $\nabla g(Z;\gamma,\beta)$ is the derivative of $g$ with respect to $(\gamma',\beta)$.

5. $\mathbb{E}[\nabla g(Z;\gamma,\beta)]'\mathbb{E}[\nabla g(Z;\gamma,\beta)]$ is nonsingular.

6. $\frac{1}{N}\sum_{i=1}^{N} g\left(Z_i;\hat{\gamma},\beta\right) \xrightarrow{p} 0$ and $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{0i}'\left(\tilde{Y}_{0i}-\tilde{X}_{0i}\gamma\right)\right) \xrightarrow{p} 0$.

Condition 1 is straightforward as long as no further constraints are imposed on $\beta$, which is true in the setting. For condition 2, observe that $\nabla_\beta g(Z;\gamma,\beta) = -\sum_t D_t$, which is continuously differentiable. In condition 3, $\mathbb{E}[g(Z;\gamma,\beta)] = 0$ is immediate by assumption.

$$\begin{aligned}\mathbb{E}\Big[\|g(W;\gamma,\beta)\|^2\Big] &= \mathbb{E}\left[\left(\sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it}-\tilde{X}_{it}'\gamma - D_{it}\beta\right)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{t=1}^{T}\left[\tilde{\varepsilon}_{it}+\left(\beta_{it}-\beta\right)D_{it}\right]D_{it}\right)^2\right] < \infty\end{aligned}$$

due to Assumption 2.1 giving those objects finite moments and $T$ being finite due to Assumption 1.4. Condition 4 is immediate from finite moments and condition 5 is immediate from Assumption 1.2. For Condition 6,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} g(Z_i;\hat{\gamma},\beta) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\tilde{Y}_{it}-\tilde{X}_{it}\hat{\gamma}-D_{it}\beta\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\beta_{it}D_{it}+\tilde{\varepsilon}_{it}-\tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)-D_{it}\beta\right]D_{it} \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\tilde{\varepsilon}_{it}-\tilde{X}_{it}'\left(\hat{\gamma}-\gamma\right)+\left(\beta_{it}-\beta\right)D_{it}\right]D_{it} = o_P(1)\end{aligned}$$

due to previous arguments. Finally, the second part of condition 6 is immediate from the WLLN. □

**Lemma 2.**

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W_{it}'\beta_{it}\right) = \begin{bmatrix} \frac{1}{N_{-\underline{R}}}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=-\underline{R}\right]\beta_{-\underline{R}it} \\ \vdots \\ \frac{1}{N_{\overline{R}}}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=\overline{R}\right]\beta_{\overline{R}it} \end{bmatrix}$$

*Proof of Lemma 2.*

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W_{it}' &= \sum_{i=1}^{N}\sum_{t=1}^{T} \begin{bmatrix} 1\left[t-t^*(i)=-\underline{R}\right] \\ \vdots \\ 1\left[t-t^*(i)=\overline{R}\right] \end{bmatrix}\begin{bmatrix} 1\left[t-t^*(i)=-\underline{R}\right] \\ \vdots \\ 1\left[t-t^*(i)=\overline{R}\right] \end{bmatrix}' \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \operatorname{diag}\left(1\left[t-t^*(i)=-\underline{R}\right],\cdots,1\left[t-t^*(i)=\overline{R}\right]\right) \\
&= \operatorname{diag}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=-\underline{R}\right],\cdots,\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=\overline{R}\right]\right) \\
&= \operatorname{diag}\left(N_{\underline{R}},\cdots,N_{\overline{R}}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W_{it}'\beta_{it} &= \sum_{i=1}^{N}\sum_{t=1}^{T} \operatorname{diag}\left(1\left[t-t^*(i)=-\underline{R}\right],\cdots,1\left[t-t^*(i)=\overline{R}\right]\right)\beta_{it} \\
&= \begin{bmatrix} \sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=-\underline{R}\right]\beta_{-\underline{R}it} \\ \vdots \\ \sum_{i=1}^{N}\sum_{t=1}^{T} 1\left[t-t^*(i)=-\underline{R}\right]\beta_{\overline{R}it} \end{bmatrix}
\end{aligned}$$

□

*Proof of Theorem 2.* The proof is analogous to that of 2SDD. The first-stage regression then yields:

$$\hat{\gamma} = \gamma + \left(\frac{1}{N}\sum_{i}\tilde{X}_{Qi}'\tilde{X}_{Qi}\right)^{-1}\left(\frac{1}{N}\sum_{i}\tilde{X}_{Qi}'\tilde{\varepsilon}_{Qi}\right)$$

Due to Assumption 1.3, and the existence of second moments in Assumption 3.2, by the weak law of large numbers (WLLN), $\frac{1}{N}\sum_i \tilde{X}'_{Qi}\tilde{X}_{Qi} \xrightarrow{p} \mathbb{E}\left[\tilde{X}'_{Qi}\tilde{X}_{Qi}\right]$. Similarly, $\frac{1}{N}\sum_i \tilde{X}'_{Qi}\tilde{\varepsilon}_{Qi} \xrightarrow{p} 0$ and $\hat{\gamma} \xrightarrow{p} \gamma$.

$$\hat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\beta_{it}\right) + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\left(\tilde{\varepsilon}_{it} + \tilde{X}_{it}\left(\hat{\gamma}-\gamma\right)\right)\right)$$

Due to Assumption 3.2, $\left|\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\tilde{X}_{it}\right)\right|$ is bounded in probability, so

$$\begin{aligned}\hat{\beta} &= \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\left(\tilde{\varepsilon}_{it} - \tilde{X}'_{it}\left(\hat{\gamma}-\gamma\right)\right)\right) + o_P(1)\\ &= \beta + \left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\tilde{\varepsilon}_{it}\right) + o_P(1)\end{aligned}$$

Due to Assumption 1.3, $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}W'_{it} \to \mathbb{E}\left[\text{diag}\left(N_{i,-\underline{R}},\cdots,N_{i,\overline{R}}\right)\right] =: \Sigma_W$.

Due to Assumption 3.3, $\Sigma_W$ is invertible and finite. $\sum_{t=1}^{T} W_{it}\tilde{\varepsilon}_{it}$ are also independent over individuals. Due to finite moments, we can apply the law of large numbers to obtain $\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\tilde{\varepsilon}_{it} \xrightarrow{p} 0$. Then, $\hat{\beta} = \beta + o_P(1)$ as required.

If the conditions of Theorem 6.1 of Newey and McFadden (1994) are satisfied, the result automatically follows. Verifying the conditions is analogous to the proof of Theorem 1.

□

# References


**Abadie, Alberto.** 2005. "Semiparametric difference-in-differences estimators." *The review of economic studies*, 72(1): 1–19. 3

**Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge.** 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1): 265–296. 2

**Athey, Susan, and Guido W Imbens.** 2022. "Design-based analysis in difference-in-differences settings with staggered adoption." *Journal of Econometrics*, 226(1): 62–79. 2, 6, 21, 31

**Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan.** 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics*, 119(1): 249–275. 2, 20, 21, 22

**Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2021. "Revisiting event study designs: Robust and efficient estimation." *arXiv preprint arXiv:2108.12419*. 2

**Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2023. "Revisiting Event Study Designs: Robust and Efficient Estimation." *Available at SSRN 2826228*. 2, 3, 6, 7, 13, 21, 22, 23, 24, 25, 26, 27, 28, 41

**Caetano, Carolina, Brantly Callaway, Stroud Payne, and Hugo Sant'Anna Rodrigues.** 2022. "Difference in differences with time-varying covariates." *arXiv preprint arXiv:2202.02903*. 9

**Callaway, Brantly, and Pedro HC Sant'Anna.** 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230. 2, 3, 9, 10, 22, 23, 24

**Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna.** 2021. "Difference-in-differences with a continuous treatment." *arXiv preprint arXiv:2107.02637*. 27, 41

**Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer.** 2019. "The effect of minimum wages on low-wage jobs." *The Quarterly Journal of Economics*, 134(3): 1405–1454. 3

**De Chaisemartin, Clément, and Xavier d'Haultfoeuille.** 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996. 2, 3, 6, 7, 22, 23, 41

**Deshpande, Manasi, and Yue Li.** 2019. "Who is screened out? Application costs and the targeting of disability programs." *American Economic Journal: Economic Policy*, 11(4): 213–248. 3

**Dumont, Michel, Glenn Rayp, Olivier Thas, and Peter Willeme.** 2005. "Correcting standard errors in two-stage estimation procedures with generated regressands." *Oxford Bulletin of Economics and Statistics*, 67(3): 421–433. 13

**Gardner, John.** 2020. "Two-stage differences in differences." *Mimeo.* , 2

**Gibbons, Charles E, Juan Carlos Suárez Serrato, and Michael B Urbancic.** 2018. "Broken or fixed effects?" *Journal of Econometric Methods*, 8(1): 20170002. 3

**Goodman-Bacon, Andrew.** 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277. 2, 3, 6, 7

**Gormley, Todd A, and David A Matsa.** 2011. "Growing out of trouble? Corporate responses to liability risk." *The Review of Financial Studies*, 24(8): 2781–2821. 3

**Hansen, Lars Peter.** 1982. "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the econometric society*, 1029–1054. 13

**Imai, Kosuke, and In Song Kim.** 2021. "On the use of two-way fixed effects regression models for causal inference with panel data." *Political Analysis*, 29(3): 405–415. 2, 6

**Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach.** 2018. "School finance reform and the distribution of student achievement." *American Economic Journal: Applied Economics*, 10(2): 1–26. 22, 24, 25

**Liu, Licheng, Ye Wang, and Yiqing Xu.** 2022. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*. 12, 13

**Miller, Douglas L.** 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives*, 37(2): 203–230. 22

**Newey, Whitney K.** 1984. "A method of moments interpretation of sequential estimators." *Economics Letters*, 14(2-3): 201–206. 13, 14

**Newey, Whitney K, and Daniel McFadden.** 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245. 13, 14, 18

**Sant'Anna, Pedro HC, and Jun Zhao.** 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics*, 219(1): 101–122. 9

**Sun, Liyang, and Sarah Abraham.** 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, 225(2): 175–199. 2, 3, 11, 12, 22, 23, 24, 27, 28, 32, 41

**Thakral, Neil, and Linh Tô.** 2020. "Anticipation and consumption." *Available at SSRN 3756188*. 2

**Thakral, Neil, and Linh T Tô.** 2023. "When Are Estimates Independent of Measurement Units?" *Mimeo*. 21

Table 1: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 20 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.0 | 0.6114 | 0.0310 | 0.5931 | 0.148 |
| | 1 | 7.2 | 0.6108 | -0.0188 | 0.6340 | |
| | 2 | 5.6 | 0.6177 | 0.0501 | 0.6370 | |
| | 3 | 4.6 | 0.6206 | -0.0062 | 0.6036 | |
| | 4 | 5.0 | 0.6275 | 0.0160 | 0.6182 | |
| BJS | 0 | 17.6 | 0.4461 | 0.0310 | 0.6269 | 0.3364 |
| | 1 | 19.0 | 0.4422 | -0.0188 | 0.6425 | |
| | 2 | 16.8 | 0.4516 | 0.0501 | 0.5873 | |
| | 3 | 12.6 | 0.4586 | -0.0062 | 0.5329 | |
| | 4 | 13.8 | 0.4578 | 0.0160 | 0.6027 | |
| BJS (leave out) | 0 | 1.0 | 0.8436 | 0.0310 | 0.6269 | 0.3545 |
| | 1 | 1.4 | 0.8284 | -0.0188 | 0.6425 | |
| | 2 | 2.0 | 0.8453 | 0.0501 | 0.5873 | |
| | 3 | 1.0 | 0.8528 | -0.0062 | 0.5329 | |
| | 4 | 1.6 | 0.8481 | 0.0160 | 0.6027 | |
| dCDH | 0 | 4.8 | 0.6980 | 0.0435 | 0.6741 | 208.938 |
| | 1 | 6.4 | 0.6992 | 0.0004 | 0.6981 | |
| | 2 | 5.8 | 0.7094 | 0.0576 | 0.7063 | |
| | 3 | 5.2 | 0.7156 | 0.0096 | 0.6848 | |
| | 4 | 5.4 | 0.7157 | 0.0300 | 0.6925 | |
| CS | 0 | 5.0 | 0.8457 | 0.0621 | 0.8044 | 30.558 |
| | 1 | 6.0 | 0.8389 | 0.0047 | 0.8869 | |
| | 2 | 5.8 | 0.8401 | 0.0719 | 0.8178 | |
| | 3 | 4.2 | 0.8548 | 0.0227 | 0.7865 | |
| | 4 | 5.0 | 0.8605 | 0.0324 | 0.8514 | |
| SA | 0 | 1.8 | 0.9858 | 0.0626 | 0.8022 | 43.392 |
| | 1 | 1.6 | 0.9770 | 0.0051 | 0.8808 | |
| | 2 | 2.4 | 0.9834 | 0.0726 | 0.8165 | |
| | 3 | 1.6 | 0.9965 | 0.0233 | 0.7657 | |
| | 4 | 2.0 | 1.0000 | 0.0331 | 0.8381 | |

Note: The table reports results from 500 simulations of 40 treated states over 20 years, with two treated states in each of those years. The data consist of log wages for women between the ages of 25 and 50 from the Current Population Survey (CPS). Treatment effects are heterogeneous and drawn from a normal distribution, with an average value randomly drawn between 10 percent and 50 percent of the average wage and a standard deviation equal to 50 percent of the average wage. Rejection rate denotes the percentage of simulations in which the specified parameter estimate significantly differs from the true value at the 5 percent significance level. S.E. denotes the standard error averaged across all simulations. Bias denotes the average difference between the point estimate and the true value. RMSE denotes the root-mean-square error. 2SDD refers to the method proposed in the current paper. BJS and BJS (leave out) refer to the default asymptotic standard errors and leave-out versions from Borusyak, Jaravel and Spiess (2023). dCDH, CS, and SA refer to the methods proposed by De Chaisemartin and d'Haultfoeuille (2020), Callaway, Goodman-Bacon and Sant'Anna (2021), and Sun and Abraham (2021), respectively. TWFE denotes the two-way fixed effects estimator. Average speed per simulation using the corresponding Stata package for each method is reported in seconds.

Table 2: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 20 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 6.0 | 0.6308 | 0.0478 | 0.6006 | 0.16 |
| | 1 | 5.8 | 0.6295 | 0.0234 | 0.6348 | |
| | 2 | 5.2 | 0.6227 | -0.0083 | 0.5825 | |
| | 3 | 5.6 | 0.6299 | 0.0061 | 0.6113 | |
| | 4 | 4.6 | 0.6318 | 0.0204 | 0.6171 | |
| BJS | 0 | 15.2 | 0.4633 | 0.0478 | 0.6006 | 0.432 |
| | 1 | 16.8 | 0.4624 | 0.0234 | 0.6348 | |
| | 2 | 14.8 | 0.4545 | -0.0083 | 0.5825 | |
| | 3 | 15.0 | 0.4702 | 0.0061 | 0.6113 | |
| | 4 | 15.0 | 0.4673 | 0.0204 | 0.6171 | |
| CS | 0 | 6.2 | 0.8803 | 0.0469 | 0.8220 | 33.086 |
| | 1 | 4.2 | 0.8742 | 0.0342 | 0.8061 | |
| | 2 | 4.6 | 0.8740 | -0.0051 | 0.8344 | |
| | 3 | 6.0 | 0.8783 | 0.0174 | 0.8028 | |
| | 4 | 6.2 | 0.8726 | 0.0343 | 0.8514 | |
| SA | 0 | 2.4 | 1.0339 | 0.0468 | 0.8189 | 46.87 |
| | 1 | 2.2 | 1.0274 | 0.0346 | 0.8013 | |
| | 2 | 2.2 | 1.0213 | -0.0050 | 0.8244 | |
| | 3 | 2.4 | 1.0327 | 0.0177 | 0.7915 | |
| | 4 | 2.8 | 1.0239 | 0.0351 | 0.8411 | |

Note: The table reports results from 500 simulations of 40 treated states over 20 years, with at least one treated state in each of those years. See the note accompanying Table 1 for further information.

Table 3: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 30 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.2 | 0.6169 | -0.0033 | 0.5360 | 0.125 |
| | 1 | 4.4 | 0.6172 | 0.0758 | 0.5756 | |
| | 2 | 3.6 | 0.6070 | 0.0075 | 0.6527 | |
| | 3 | 8.2 | 0.6141 | 0.0340 | 0.6785 | |
| | 4 | 6.2 | 0.6182 | 0.0190 | 0.5984 | |
| BJS | 0 | 31.2 | 0.3213 | -0.0033 | 0.5360 | 0.825 |
| | 1 | 23.8 | 0.3325 | 0.0758 | 0.5756 | |
| | 2 | 29.2 | 0.3380 | 0.0075 | 0.6527 | |
| | 3 | 26.2 | 0.3487 | 0.0340 | 0.6785 | |
| | 4 | 29.6 | 0.3463 | 0.0190 | 0.5984 | |
| CS | 0 | 3.8 | 0.8509 | -0.0298 | 0.8199 | 58.5 |
| | 1 | 4.8 | 0.8385 | 0.0462 | 0.8377 | |
| | 2 | 6.2 | 0.8430 | -0.0235 | 0.9195 | |
| | 3 | 5.6 | 0.8521 | 0.0062 | 0.9548 | |
| | 4 | 4.4 | 0.8472 | -0.0048 | 0.7631 | |
| SA | 0 | 2.6 | 1.0145 | -0.0276 | 0.8125 | 182.125 |
| | 1 | 2.2 | 0.9985 | 0.0487 | 0.8221 | |
| | 2 | 2.2 | 1.0156 | -0.0205 | 0.9032 | |
| | 3 | 3.4 | 1.0259 | 0.0088 | 0.9495 | |
| | 4 | 0.4 | 1.0153 | -0.0026 | 0.7515 | |

Note: The table reports results from 500 simulations of 40 treated states over 30 years, with at least one treated state in each of those years. See the note accompanying Table 1 for further information.

Table 4: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 15 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.4 | 0.6252 | 0.0112 | 0.5946 | 0.128 |
| | 1 | 7.2 | 0.6256 | -0.0260 | 0.6694 | |
| | 2 | 5.6 | 0.6302 | -0.0379 | 0.6759 | |
| | 3 | 4.4 | 0.6337 | -0.0301 | 0.6406 | |
| | 4 | 6.0 | 0.6309 | 0.0008 | 0.6324 | |
| BJS | 0 | 10.4 | 0.5041 | 0.0112 | 0.5946 | 0.398 |
| | 1 | 12.6 | 0.5090 | -0.0260 | 0.6694 | |
| | 2 | 10.0 | 0.5174 | -0.0379 | 0.6759 | |
| | 3 | 12.6 | 0.5136 | -0.0301 | 0.6406 | |
| | 4 | 11.8 | 0.5145 | 0.0008 | 0.6324 | |
| BJS (leave out) | 0 | 2.0 | 0.7717 | 0.0112 | 0.5946 | 0.386 |
| | 1 | 3.2 | 0.7763 | -0.0260 | 0.6694 | |
| | 2 | 1.4 | 0.7893 | -0.0379 | 0.6759 | |
| | 3 | 2.8 | 0.7859 | -0.0301 | 0.6406 | |
| | 4 | 2.0 | 0.7761 | 0.0008 | 0.6324 | |
| CS | 0 | 4.6 | 0.8864 | -0.0124 | 0.8785 | 22.218 |
| | 1 | 4.2 | 0.8869 | -0.0518 | 0.8949 | |
| | 2 | 6.4 | 0.8857 | -0.0750 | 0.9284 | |
| | 3 | 4.8 | 0.8882 | -0.0580 | 0.8772 | |
| | 4 | 5.6 | 0.8887 | -0.0346 | 0.9072 | |
| SA | 0 | 2.0 | 1.0063 | -0.0092 | 0.8809 | 24.09 |
| | 1 | 2.0 | 1.0095 | -0.0480 | 0.8873 | |
| | 2 | 3.4 | 1.0107 | -0.0719 | 0.9218 | |
| | 3 | 1.8 | 1.0098 | -0.0544 | 0.8607 | |
| | 4 | 3.0 | 1.0112 | -0.0308 | 0.8936 | |

Note: The table reports results from 500 simulations of 40 treated states over 15 years, with at least two treated states in each of those years. See the note accompanying Table 1 for further information.

Table 5: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 15 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.2 | 0.6278 | 0.0316 | 0.6675 | 0.1409 |
| | 1 | 4.6 | 0.6393 | -0.0167 | 0.6299 | |
| | 2 | 5.6 | 0.6320 | 0.0307 | 0.6738 | |
| | 3 | 4.2 | 0.6368 | 0.0462 | 0.6221 | |
| | 4 | 6.4 | 0.6395 | -0.0026 | 0.6184 | |
| BJS | 0 | 12.7 | 0.5176 | 0.0316 | 0.6675 | 0.9318 |
| | 1 | 10.9 | 0.5267 | -0.0167 | 0.6299 | |
| | 2 | 15.0 | 0.5188 | 0.0307 | 0.6738 | |
| | 3 | 10.9 | 0.5175 | 0.0462 | 0.6221 | |
| | 4 | 10.5 | 0.5329 | -0.0026 | 0.6184 | |
| CS | 0 | 5.9 | 0.8952 | 0.0486 | 0.8910 | 26.2 |
| | 1 | 6.4 | 0.9086 | 0.0076 | 0.9216 | |
| | 2 | 6.4 | 0.8989 | 0.0675 | 0.9158 | |
| | 3 | 3.6 | 0.8970 | 0.0789 | 0.8529 | |
| | 4 | 4.5 | 0.9006 | 0.0437 | 0.8693 | |
| SA | 0 | 1.8 | 1.0175 | 0.0475 | 0.8802 | 28.4 |
| | 1 | 2.3 | 1.0339 | 0.0071 | 0.9069 | |
| | 2 | 3.2 | 1.0223 | 0.0664 | 0.9011 | |
| | 3 | 2.3 | 1.0204 | 0.0781 | 0.8466 | |
| | 4 | 2.7 | 1.0269 | 0.0419 | 0.8433 | |

Note: The table reports results from 500 simulations of 40 treated states over 15 years, with at least one treated state in each of those years. See the note accompanying Table 1 for further information.

Table 6: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 5 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 6.4 | 0.6913 | -0.0482 | 0.7036 | 0.1380 |
| | 1 | 4.6 | 0.6970 | 0.0020 | 0.6953 | |
| | 2 | 5.4 | 0.7076 | 0.0606 | 0.7157 | |
| | 3 | 7.2 | 0.7134 | -0.0164 | 0.7328 | |
| | 4 | 5.8 | 0.7176 | 0.0080 | 0.7339 | |
| BJS | 0 | 6.8 | 0.6575 | -0.0482 | 0.7036 | 0.3580 |
| | 1 | 6.4 | 0.6622 | 0.0020 | 0.6953 | |
| | 2 | 6.0 | 0.6734 | 0.0606 | 0.7157 | |
| | 3 | 8.8 | 0.6794 | -0.0164 | 0.7328 | |
| | 4 | 6.6 | 0.6835 | 0.0080 | 0.7339 | |
| CS | 0 | 6.2 | 1.0497 | -0.0807 | 1.0157 | 5.3620 |
| | 1 | 6.6 | 1.0455 | -0.0286 | 1.0907 | |
| | 2 | 4.8 | 1.0429 | 0.0132 | 1.0852 | |
| | 3 | 5.6 | 1.0444 | -0.0575 | 1.0801 | |
| | 4 | 6.2 | 1.0503 | -0.0470 | 1.0750 | |
| SA | 0 | 2.4 | 1.0899 | -0.0825 | 0.9947 | 1.4340 |
| | 1 | 4.6 | 1.0882 | -0.0320 | 1.0715 | |
| | 2 | 4.0 | 1.0846 | 0.0120 | 1.0594 | |
| | 3 | 4.4 | 1.0878 | -0.0574 | 1.0646 | |
| | 4 | 3.0 | 1.0935 | -0.0472 | 1.0470 | |

Note: The table reports results from 500 simulations of 40 treated states over 5 years, with at least one treated state in each of those years. See the note accompanying Table 1 for further information.

Table 7: Simulations (CPS wage data, heterogeneous treatment effects): 40 states treated over 10 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.4 | 0.6431 | 0.0141 | 0.6108 | 0.166 |
| | 1 | 5.0 | 0.6470 | -0.0361 | 0.6692 | |
| | 2 | 5.2 | 0.6522 | 0.0090 | 0.6250 | |
| | 3 | 6.2 | 0.6522 | -0.0239 | 0.6907 | |
| | 4 | 5.8 | 0.6568 | -0.0149 | 0.6630 | |
| BJS | 0 | 8.4 | 0.5694 | 0.0141 | 0.6108 | 0.392 |
| | 1 | 9.8 | 0.5746 | -0.0361 | 0.6692 | |
| | 2 | 7.2 | 0.5772 | 0.0090 | 0.6250 | |
| | 3 | 9.8 | 0.5793 | -0.0239 | 0.6907 | |
| | 4 | 10.2 | 0.5817 | -0.0149 | 0.6630 | |
| BJS (leave out) | 0 | 3.0 | 0.7307 | 0.0141 | 0.6108 | 0.336 |
| | 1 | 3.2 | 0.7365 | -0.0361 | 0.6692 | |
| | 2 | 2.6 | 0.7359 | 0.0090 | 0.6250 | |
| | 3 | 4.8 | 0.7364 | -0.0239 | 0.6907 | |
| | 4 | 4.0 | 0.7384 | -0.0149 | 0.6630 | |
| CS | 0 | 7.4 | 0.9391 | -0.0005 | 0.9660 | 12.858 |
| | 1 | 6.4 | 0.9255 | -0.0384 | 0.9420 | |
| | 2 | 6.2 | 0.9293 | 0.0084 | 0.9225 | |
| | 3 | 6.2 | 0.9251 | -0.0109 | 0.9729 | |
| | 4 | 4.2 | 0.9261 | -0.0047 | 0.9174 | |
| SA | 0 | 4.6 | 1.0302 | 0.0030 | 0.9593 | 6.924 |
| | 1 | 4.0 | 1.0178 | -0.0338 | 0.9298 | |
| | 2 | 3.8 | 1.0199 | 0.0124 | 0.9081 | |
| | 3 | 3.6 | 1.0155 | -0.0059 | 0.9567 | |
| | 4 | 2.8 | 1.0154 | -0.0003 | 0.9058 | |

Note: The table reports results from 500 simulations of 40 treated states over 10 years, with at least two treated states in each of those years. See the note accompanying Table 1 for further information.

Table 8: Simulations (CPS wage data, heterogeneous treatment effects): 30 states treated over 10 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.8 | 0.7160 | 0.0402 | 0.7323 | 0.152 |
| | 1 | 5.6 | 0.7112 | 0.0400 | 0.7226 | |
| | 2 | 6.0 | 0.7093 | 0.0178 | 0.7134 | |
| | 3 | 8.0 | 0.7100 | 0.0585 | 0.7370 | |
| | 4 | 4.4 | 0.7124 | -0.0059 | 0.6885 | |
| BJS | 0 | 13.2 | 0.6007 | 0.0402 | 0.7323 | 0.43 |
| | 1 | 11.2 | 0.5950 | 0.0400 | 0.7226 | |
| | 2 | 11.0 | 0.5928 | 0.0178 | 0.7134 | |
| | 3 | 12.2 | 0.5958 | 0.0585 | 0.7370 | |
| | 4 | 8.8 | 0.5939 | -0.0059 | 0.6885 | |
| BJS (leave out) | 0 | 2.6 | 0.8803 | 0.0402 | 0.7323 | 0.376 |
| | 1 | 2.8 | 0.8673 | 0.0400 | 0.7226 | |
| | 2 | 3.0 | 0.8639 | 0.0178 | 0.7134 | |
| | 3 | 3.6 | 0.8619 | 0.0585 | 0.7370 | |
| | 4 | 1.6 | 0.8579 | -0.0059 | 0.6885 | |
| CS | 0 | 5.6 | 0.9843 | -0.0055 | 0.9208 | 14.686 |
| | 1 | 6.0 | 0.9799 | -0.0164 | 0.9647 | |
| | 2 | 5.4 | 0.9749 | -0.0321 | 0.9590 | |
| | 3 | 6.2 | 0.9828 | 0.0139 | 0.9864 | |
| | 4 | 5.6 | 0.9742 | -0.0631 | 0.9641 | |
| SA | 0 | 2.4 | 1.0689 | -0.0025 | 0.9155 | 7.276 |
| | 1 | 3.4 | 1.0630 | -0.0137 | 0.9528 | |
| | 2 | 3.8 | 1.0584 | -0.0284 | 0.9474 | |
| | 3 | 4.2 | 1.0657 | 0.0173 | 0.9698 | |
| | 4 | 2.8 | 1.0577 | -0.0605 | 0.9556 | |

Note: The table reports results from 500 simulations of 30 treated states over 10 years, with at least two treated states in each of those years. See the note accompanying Table 1 for further information.

Table 9: Simulations (CPS wage data, heterogeneous treatment effects): 20 states treated over 10 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.0 | 0.8386 | 0.0645 | 0.8260 | 0.14 |
| | 1 | 5.6 | 0.8371 | -0.0755 | 0.8592 | |
| | 2 | 5.4 | 0.8316 | -0.0218 | 0.8632 | |
| | 3 | 8.0 | 0.8502 | 0.0127 | 0.8903 | |
| | 4 | 7.6 | 0.8435 | 0.0352 | 0.8568 | |
| BJS | 0 | 18.8 | 0.6120 | 0.0645 | 0.8260 | 0.29 |
| | 1 | 18.8 | 0.6105 | -0.0755 | 0.8592 | |
| | 2 | 19.2 | 0.6083 | -0.0218 | 0.8632 | |
| | 3 | 21.4 | 0.6239 | 0.0127 | 0.8903 | |
| | 4 | 17.2 | 0.6199 | 0.0352 | 0.8568 | |
| BJS (leave out) | 0 | 0.6 | 1.1776 | 0.0645 | 0.8260 | 0.27 |
| | 1 | 1.6 | 1.1718 | -0.0755 | 0.8592 | |
| | 2 | 1.6 | 1.1659 | -0.0218 | 0.8632 | |
| | 3 | 2.0 | 1.1944 | 0.0127 | 0.8903 | |
| | 4 | 3.0 | 1.1858 | 0.0352 | 0.8568 | |
| CS | 0 | 5.4 | 1.1523 | 0.0441 | 1.1802 | 10.716 |
| | 1 | 6.6 | 1.1601 | -0.0939 | 1.2108 | |
| | 2 | 7.0 | 1.1500 | -0.0353 | 1.2051 | |
| | 3 | 7.2 | 1.1492 | -0.0078 | 1.2191 | |
| | 4 | 7.4 | 1.1481 | 0.0141 | 1.2007 | |
| SA | 0 | 3.2 | 1.2223 | 0.0422 | 1.1795 | 4.354 |
| | 1 | 4.8 | 1.2324 | -0.0954 | 1.1853 | |
| | 2 | 4.8 | 1.2212 | -0.0368 | 1.1766 | |
| | 3 | 3.6 | 1.2242 | -0.0094 | 1.1952 | |
| | 4 | 5.2 | 1.2224 | 0.0126 | 1.1831 | |

Note: The table reports results from 500 simulations of 20 treated states over 10 years, with at least two treated states in each of those years. See the note accompanying Table 1 for further information.

Table 10: Simulations (CPS wage data, homogeneous treatment effects): 40 states treated over 20 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.0 | 0.6100 | 0.0310 | 0.5931 | 0.134 |
| | 1 | 7.2 | 0.6120 | -0.0188 | 0.6340 | |
| | 2 | 5.6 | 0.6184 | 0.0501 | 0.6370 | |
| | 3 | 4.6 | 0.6228 | -0.0062 | 0.6036 | |
| | 4 | 5.0 | 0.6267 | 0.0160 | 0.6182 | |
| BJS | 0 | 17.6 | 0.4413 | 0.0310 | 0.5931 | 0.384 |
| | 1 | 19.0 | 0.4418 | -0.0188 | 0.6340 | |
| | 2 | 16.8 | 0.4510 | 0.0501 | 0.6370 | |
| | 3 | 12.6 | 0.4591 | -0.0062 | 0.6036 | |
| | 4 | 13.8 | 0.4570 | 0.0160 | 0.6182 | |
| BJS (leave out) | 0 | 1.0 | 0.8361 | 0.0310 | 0.5931 | 0.3 |
| | 1 | 1.4 | 0.8297 | -0.0188 | 0.6340 | |
| | 2 | 2.0 | 0.8444 | 0.0501 | 0.6370 | |
| | 3 | 1.0 | 0.8541 | -0.0062 | 0.6036 | |
| | 4 | 1.6 | 0.8462 | 0.0160 | 0.6182 | |
| CS | 0 | 5.0 | 0.8502 | 0.0621 | 0.8038 | 29.814 |
| | 1 | 6.0 | 0.8471 | 0.0047 | 0.8405 | |
| | 2 | 5.8 | 0.8508 | 0.0719 | 0.8710 | |
| | 3 | 4.2 | 0.8586 | 0.0227 | 0.8219 | |
| | 4 | 5.0 | 0.8588 | 0.0324 | 0.8191 | |
| SA | 0 | 1.8 | 0.9879 | 0.0626 | 0.8010 | 43.332 |
| | 1 | 1.6 | 0.9846 | 0.0051 | 0.8397 | |
| | 2 | 2.4 | 0.9914 | 0.0726 | 0.8671 | |
| | 3 | 1.6 | 1.0006 | 0.0233 | 0.8116 | |
| | 4 | 2.0 | 1.0005 | 0.0331 | 0.8114 | |
| TWFE | 0 | 3.2 | 0.5992 | 0.0442 | 0.5780 | 0.062 |
| | 1 | 6.8 | 0.6002 | -0.0130 | 0.6209 | |
| | 2 | 5.8 | 0.6080 | 0.0479 | 0.6255 | |
| | 3 | 4.4 | 0.6096 | -0.0067 | 0.5927 | |
| | 4 | 4.0 | 0.6127 | 0.0125 | 0.6058 | |

Note: The table reports results from 500 simulations of 40 treated states over 20 years, with two treated states in each of those years. Treatment effects are homogeneous and drawn from a normal distribution, with an average value set to 30 percent of the average wage and a standard deviation equal to 50 percent of the average wage. See the note accompanying Table 1 for further information.

Table 11: Simulations (i.i.d. data, heterogeneous treatment effects): 40 states treated over 20 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.0 | 0.6159 | 0.0299 | 0.5976 | 0.146 |
| | 1 | 7.6 | 0.6166 | -0.0120 | 0.6432 | |
| | 2 | 5.6 | 0.6234 | 0.0567 | 0.6406 | |
| | 3 | 5.4 | 0.6281 | -0.0018 | 0.6119 | |
| | 4 | 5.2 | 0.6316 | 0.0173 | 0.6206 | |
| BJS | 0 | 15.6 | 0.4457 | 0.0299 | 0.5976 | 0.352 |
| | 1 | 18.6 | 0.4456 | -0.0120 | 0.6432 | |
| | 2 | 16.8 | 0.4549 | 0.0567 | 0.6406 | |
| | 3 | 13.8 | 0.4624 | -0.0018 | 0.6119 | |
| | 4 | 13.8 | 0.4605 | 0.0173 | 0.6206 | |
| BJS (leave out) | 0 | 1.0 | 0.8442 | 0.0299 | 0.5976 | 0.3 |
| | 1 | 1.8 | 0.8370 | -0.0120 | 0.6432 | |
| | 2 | 2.0 | 0.8515 | 0.0567 | 0.6406 | |
| | 3 | 1.0 | 0.8603 | -0.0018 | 0.6119 | |
| | 4 | 1.4 | 0.8525 | 0.0173 | 0.6206 | |
| CS | 0 | 5.4 | 0.8573 | 0.0539 | 0.8180 | 26.164 |
| | 1 | 6.2 | 0.8551 | 0.0043 | 0.8550 | |
| | 2 | 6.6 | 0.8583 | 0.0702 | 0.8802 | |
| | 3 | 5.2 | 0.8655 | 0.0192 | 0.8354 | |
| | 4 | 4.8 | 0.8658 | 0.0266 | 0.8338 | |
| SA | 0 | 1.8 | 0.9964 | 0.0544 | 0.8157 | 35.854 |
| | 1 | 2.0 | 0.9938 | 0.0047 | 0.8532 | |
| | 2 | 2.4 | 1.0000 | 0.0711 | 0.8768 | |
| | 3 | 1.6 | 1.0084 | 0.0199 | 0.8258 | |
| | 4 | 2.8 | 1.0082 | 0.0275 | 0.8265 | |

Note: The table reports results from 500 simulations of 40 treated states over 20 years, with two treated states in each of those years. The outcome data are drawn i.i.d. from a normal distribution with the same mean and variance as that of the wage data used in Table 1. See the note accompanying Table 1 for further information.

Table 12: Simulations (i.i.d. data, homogeneous treatment effects): 40 states treated over 20 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.0 | 0.6160 | 0.0416 | 0.6109 | 0.2 |
| | 1 | 7.8 | 0.6186 | 0.0027 | 0.6502 | |
| | 2 | 5.2 | 0.6237 | 0.0551 | 0.6360 | |
| | 3 | 5.8 | 0.6270 | -0.0066 | 0.6185 | |
| | 4 | 4.4 | 0.6323 | 0.0233 | 0.6159 | |
| BJS | 0 | 16.4 | 0.4467 | 0.0416 | 0.6109 | 0.48 |
| | 1 | 18.9 | 0.4486 | 0.0027 | 0.6502 | |
| | 2 | 16.1 | 0.4545 | 0.0551 | 0.6360 | |
| | 3 | 14.3 | 0.4614 | -0.0066 | 0.6185 | |
| | 4 | 13.0 | 0.4613 | 0.0233 | 0.6159 | |
| BJS (leave out) | 0 | 1.1 | 0.8455 | 0.0416 | 0.6109 | 0.53 |
| | 1 | 1.8 | 0.8430 | 0.0027 | 0.6502 | |
| | 2 | 2.3 | 0.8510 | 0.0551 | 0.6360 | |
| | 3 | 1.1 | 0.8586 | -0.0066 | 0.6185 | |
| | 4 | 1.4 | 0.8540 | 0.0233 | 0.6159 | |
| CS | 0 | 6.1 | 0.8577 | 0.0521 | 0.8312 | 44.965 |
| | 1 | 6.4 | 0.8554 | 0.0062 | 0.8576 | |
| | 2 | 5.7 | 0.8601 | 0.0539 | 0.8660 | |
| | 3 | 4.8 | 0.8648 | 0.0032 | 0.8383 | |
| | 4 | 4.3 | 0.8662 | 0.0192 | 0.8167 | |
| SA | 0 | 2.0 | 0.9966 | 0.0550 | 0.8308 | 82.439 |
| | 1 | 2.0 | 0.9946 | 0.0089 | 0.8572 | |
| | 2 | 1.8 | 1.0015 | 0.0572 | 0.8651 | |
| | 3 | 1.6 | 1.0073 | 0.0062 | 0.8301 | |
| | 4 | 2.7 | 1.0083 | 0.0226 | 0.8112 | |
| TWFE | 0 | 3.4 | 0.6048 | 0.0556 | 0.5942 | 0.095 |
| | 1 | 7.0 | 0.6067 | 0.0058 | 0.6375 | |
| | 2 | 6.1 | 0.6134 | 0.0513 | 0.6224 | |
| | 3 | 4.3 | 0.6133 | -0.0055 | 0.6018 | |
| | 4 | 3.4 | 0.6178 | 0.0161 | 0.6035 | |

Note: The table reports results from 500 simulations of 40 treated states over 20 years, with two treated states in each of those years. Treatment effects are homogeneous and drawn from a normal distribution, with an average value set to 30 percent of the average wage and a standard deviation equal to 50 percent of the average wage. See the note accompanying Table 11 for further information.

Table 13: Simulations (i.i.d. data, heterogeneous treatment effects): 40 states treated over 30 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.2 | 0.6175 | -0.0226 | 0.6011 | 0.263 |
| | 1 | 4.6 | 0.6209 | 0.0890 | 0.5839 | |
| | 2 | 4.2 | 0.6210 | -0.0046 | 0.6145 | |
| | 3 | 7.0 | 0.6161 | 0.0219 | 0.6492 | |
| | 4 | 4.4 | 0.6237 | 0.0240 | 0.6212 | |
| BJS | 0 | 28.6 | 0.3225 | -0.0226 | 0.6011 | 0.545 |
| | 1 | 24.4 | 0.3373 | 0.0890 | 0.5839 | |
| | 2 | 32.4 | 0.3382 | -0.0046 | 0.6145 | |
| | 3 | 27.8 | 0.3398 | 0.0219 | 0.6492 | |
| | 4 | 30.8 | 0.3457 | 0.0240 | 0.6212 | |
| CS | 0 | 3.6 | 0.8593 | -0.0364 | 0.8515 | 62.127 |
| | 1 | 5.5 | 0.8551 | 0.0720 | 0.8072 | |
| | 2 | 5.5 | 0.8603 | -0.0216 | 0.8613 | |
| | 3 | 5.5 | 0.8444 | 0.0067 | 0.8609 | |
| | 4 | 3.6 | 0.8586 | 0.0104 | 0.7894 | |
| SA | 0 | 2.7 | 1.0085 | -0.0350 | 0.8452 | 188.263 |
| | 1 | 1.8 | 1.0155 | 0.0736 | 0.8040 | |
| | 2 | 2.3 | 1.0168 | -0.0197 | 0.8597 | |
| | 3 | 3.2 | 1.0071 | 0.0085 | 0.8548 | |
| | 4 | 0.5 | 1.0167 | 0.0117 | 0.7874 | |

Note: The table reports results from 500 simulations of 40 treated states over 30 years, with at least one treated state in each of those years. See the note accompanying Table 11 for further information.

Table 14: Simulations (i.i.d. data, heterogeneous treatment effects): 40 states treated over 5 years (at least 1 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.6 | 0.7080 | 0.0485 | 0.6927 | 0.13 |
| | 1 | 5.2 | 0.7125 | 0.0111 | 0.6976 | |
| | 2 | 4.6 | 0.7109 | 0.0128 | 0.6812 | |
| | 3 | 7.2 | 0.7159 | 0.0084 | 0.7502 | |
| | 4 | 5.8 | 0.7230 | 0.0062 | 0.7284 | |
| BJS | 0 | 6.8 | 0.6741 | 0.0485 | 0.6927 | 0.334 |
| | 1 | 7.0 | 0.6791 | 0.0111 | 0.6976 | |
| | 2 | 5.8 | 0.6747 | 0.0128 | 0.6812 | |
| | 3 | 8.6 | 0.6800 | 0.0084 | 0.7502 | |
| | 4 | 6.6 | 0.6886 | 0.0062 | 0.7284 | |
| CS | 0 | 5.4 | 1.0624 | 0.0004 | 1.0546 | 4.904 |
| | 1 | 6.6 | 1.0645 | -0.0317 | 1.0510 | |
| | 2 | 5.0 | 1.0530 | -0.0067 | 1.0271 | |
| | 3 | 4.4 | 1.0598 | -0.0592 | 1.0484 | |
| | 4 | 6.0 | 1.0613 | -0.0296 | 1.0849 | |
| SA | 0 | 4.0 | 1.1057 | 0.0093 | 1.0127 | 1.392 |
| | 1 | 4.4 | 1.1070 | -0.0243 | 1.0308 | |
| | 2 | 3.0 | 1.0965 | -0.0012 | 1.0116 | |
| | 3 | 3.8 | 1.1034 | -0.0485 | 1.0204 | |
| | 4 | 4.4 | 1.1040 | -0.0250 | 1.0485 | |

Note: The table reports results from 500 simulations of 40 treated states over 5 years, with at least one treated state in each of those years. See the note accompanying Table 11 for further information.

Table 15: Simulations (i.i.d. data, heterogeneous treatment effects): 40 states treated over 10 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 5.4 | 0.6485 | 0.0229 | 0.6152 | 0.19 |
| | 1 | 5.2 | 0.6514 | -0.0384 | 0.6709 | |
| | 2 | 4.2 | 0.6569 | 0.0097 | 0.6320 | |
| | 3 | 6.2 | 0.6576 | -0.0186 | 0.6939 | |
| | 4 | 5.6 | 0.6625 | -0.0096 | 0.6671 | |
| BJS | 0 | 8.0 | 0.5741 | 0.0229 | 0.6152 | 0.426 |
| | 1 | 9.2 | 0.5786 | -0.0384 | 0.6709 | |
| | 2 | 7.2 | 0.5814 | 0.0097 | 0.6320 | |
| | 3 | 10.4 | 0.5842 | -0.0186 | 0.6939 | |
| | 4 | 9.6 | 0.5866 | -0.0096 | 0.6671 | |
| BJS (leave out) | 0 | 2.6 | 0.7369 | 0.0229 | 0.6152 | 0.45 |
| | 1 | 3.2 | 0.7417 | -0.0384 | 0.6709 | |
| | 2 | 2.0 | 0.7411 | 0.0097 | 0.6320 | |
| | 3 | 4.8 | 0.7426 | -0.0186 | 0.6939 | |
| | 4 | 3.8 | 0.7447 | -0.0096 | 0.6671 | |
| CS | 0 | 7.2 | 0.9460 | 0.0106 | 0.9790 | 14.604 |
| | 1 | 6.0 | 0.9313 | -0.0385 | 0.9526 | |
| | 2 | 6.6 | 0.9366 | 0.0133 | 0.9383 | |
| | 3 | 6.2 | 0.9323 | -0.0028 | 0.9759 | |
| | 4 | 4.0 | 0.9336 | 0.0029 | 0.9273 | |
| SA | 0 | 4.8 | 1.0379 | 0.0135 | 0.9739 | 7.926 |
| | 1 | 4.0 | 1.0241 | -0.0343 | 0.9405 | |
| | 2 | 3.2 | 1.0284 | 0.0171 | 0.9239 | |
| | 3 | 3.6 | 1.0233 | 0.0018 | 0.9597 | |
| | 4 | 2.8 | 1.0234 | 0.0070 | 0.9172 | |

Note: The table reports results from 500 simulations of 40 treated states over 10 years, with at least two treated states in each of those years. See the note accompanying Table 11 for further information.

Table 16: Simulations (i.i.d. data, heterogeneous treatment effects): 30 states treated over 10 years (at least 2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 6.4 | 0.7218 | 0.0322 | 0.7409 | 0.124 |
| | 1 | 5.8 | 0.7167 | 0.0408 | 0.7234 | |
| | 2 | 5.6 | 0.7153 | 0.0249 | 0.7123 | |
| | 3 | 8.0 | 0.7150 | 0.0605 | 0.7425 | |
| | 4 | 3.8 | 0.7181 | -0.0103 | 0.6902 | |
| BJS | 0 | 13.2 | 0.6052 | 0.0322 | 0.7409 | 0.302 |
| | 1 | 11.2 | 0.5990 | 0.0408 | 0.7234 | |
| | 2 | 11.0 | 0.5983 | 0.0249 | 0.7123 | |
| | 3 | 12.2 | 0.6000 | 0.0605 | 0.7425 | |
| | 4 | 8.8 | 0.5995 | -0.0103 | 0.6902 | |
| BJS (leave out) | 0 | 2.6 | 0.8866 | 0.0322 | 0.7409 | 0.298 |
| | 1 | 2.8 | 0.8729 | 0.0408 | 0.7234 | |
| | 2 | 3.0 | 0.8715 | 0.0249 | 0.7123 | |
| | 3 | 3.6 | 0.8687 | 0.0605 | 0.7425 | |
| | 4 | 1.6 | 0.8663 | -0.0103 | 0.6902 | |
| CS | 0 | 5.0 | 0.9932 | -0.0125 | 0.9292 | 10.552 |
| | 1 | 6.4 | 0.9886 | -0.0161 | 0.9645 | |
| | 2 | 5.0 | 0.9833 | -0.0248 | 0.9618 | |
| | 3 | 6.0 | 0.9899 | 0.0165 | 0.9909 | |
| | 4 | 5.4 | 0.9825 | -0.0664 | 0.9753 | |
| SA | 0 | 2.2 | 1.0781 | -0.0104 | 0.9234 | 4.982 |
| | 1 | 3.0 | 1.0721 | -0.0143 | 0.9533 | |
| | 2 | 3.8 | 1.0677 | -0.0219 | 0.9504 | |
| | 3 | 3.8 | 1.0732 | 0.0189 | 0.9736 | |
| | 4 | 2.6 | 1.0672 | -0.0648 | 0.9667 | |

Note: The table reports results from 500 simulations of 30 treated states over 10 years, with at least two treated states in each of those years. See the note accompanying Table 11 for further information.

Table 17: Simulations (i.i.d. data, heterogeneous treatment effects): 20 states treated over 10 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 4.6 | 0.8376 | 0.0619 | 0.8310 | 0.208 |
| | 1 | 5.2 | 0.8418 | -0.0798 | 0.8649 | |
| | 2 | 5.8 | 0.8307 | -0.0278 | 0.8738 | |
| | 3 | 8.0 | 0.8482 | 0.0140 | 0.8942 | |
| | 4 | 7.8 | 0.8443 | 0.0365 | 0.8736 | |
| BJS | 0 | 17.8 | 0.6085 | 0.0619 | 0.8310 | 0.512 |
| | 1 | 18.6 | 0.6131 | -0.0798 | 0.8649 | |
| | 2 | 20.0 | 0.6087 | -0.0278 | 0.8738 | |
| | 3 | 21.4 | 0.6160 | 0.0140 | 0.8942 | |
| | 4 | 17.6 | 0.6217 | 0.0365 | 0.8736 | |
| BJS (leave out) | 0 | 1.0 | 1.1705 | 0.0619 | 0.8310 | 0.476 |
| | 1 | 1.4 | 1.1779 | -0.0798 | 0.8649 | |
| | 2 | 1.4 | 1.1673 | -0.0278 | 0.8738 | |
| | 3 | 1.8 | 1.1780 | 0.0140 | 0.8942 | |
| | 4 | 2.8 | 1.1903 | 0.0365 | 0.8736 | |
| CS | 0 | 5.4 | 1.1466 | 0.0431 | 1.1850 | 19.238 |
| | 1 | 6.4 | 1.1550 | -0.0973 | 1.2154 | |
| | 2 | 8.0 | 1.1415 | -0.0397 | 1.2404 | |
| | 3 | 7.4 | 1.1531 | -0.0053 | 1.2296 | |
| | 4 | 8.2 | 1.1479 | 0.0170 | 1.2015 | |
| SA | 0 | 3.6 | 1.2178 | 0.0409 | 1.1749 | 8.22 |
| | 1 | 4.2 | 1.2274 | -0.0991 | 1.1889 | |
| | 2 | 4.6 | 1.2136 | -0.0415 | 1.2100 | |
| | 3 | 4.2 | 1.2260 | -0.0072 | 1.2066 | |
| | 4 | 5.6 | 1.2225 | 0.0152 | 1.1805 | |

Note: The table reports results from 500 simulations of 20 treated states over 10 years, with two treated states in each of those years. See the note accompanying Table 11 for further information.

Table 18: Simulations (CPS wage data, heterogeneous treatment effects): Individuals from 40 states treated over 20 years (2 per year)

| Method | Period | Rejection rate | S.E. | Bias | RMSE | Speed (secs) |
|---|---|---|---|---|---|---|
| 2SDD | 0 | 6.6 | 0.0249 | 0.0000 | 0.0250 | 21.538 |
| | 1 | 5.6 | 0.0256 | 0.0024 | 0.0258 | |
| | 2 | 5.2 | 0.0259 | 0.0005 | 0.0261 | |
| | 3 | 7.0 | 0.0264 | -0.0003 | 0.0267 | |
| | 4 | 6.4 | 0.0274 | 0.0001 | 0.0271 | |
| dCDH | 0 | 5.6 | 0.0331 | 0.0004 | 0.0330 | 183.3125 |
| | 1 | 7.6 | 0.0337 | 0.0024 | 0.0338 | |
| | 2 | 5.6 | 0.0339 | -0.0007 | 0.0335 | |
| | 3 | 8.8 | 0.0345 | -0.0022 | 0.0353 | |
| | 4 | 3.8 | 0.0358 | -0.0003 | 0.0319 | |

Note: The table reports results from 500 simulations of individuals from 40 treated states over 20 years, with two treated states in each of those years. The sample size consists of 1,038,908 women between the ages of 25 and 50 from the CPS from 1979 to 2020. See the note accompanying Table 1 for further information.