Banc Ceannais na hÉireann
Central Bank of Ireland
Eurosystem

Research Technical Paper

Estimating effects of staggered intervention with count and binary outcomes: a simulation study

Anil Yadav, John McHale, Jason Harold & Stephen O'Neill

Vol. 2024, No. 4

Non-technical summary

Difference-in-Differences (DiD) is a popular method to estimate the effects of policy changes or interventions. DiD compares the differences in outcomes between groups that have been exposed to an intervention (treated) and those not exposed (control), both before and after the intervention. The difference in these differences attributed to be the effects of the intervention. When data is available for multiple periods before and after the intervention, dynamic effects of the intervention are estimated using an event-study (ES) approach. ES allows researchers to observe how the effect changes with varying length of exposure.

This paper addresses an important gap in the literature on DiD and ES approaches. Recent DiD and ES methodological developments have raised concerns about potential bias in estimated effects when these approaches are implemented using a two-way fixed effects (TWFE) estimator, especially in the context of heterogeneous treatment effects and staggered interventions. Several alternative estimators have been proposed that circumvent the so-called 'bad control' units problem and recover unbiased treatment effects. However, the focus has primarily been on linear outcome models. Our research extends this discussion to nonlinear outcome models such as count and binary outcomes, which are prevalent in many fields of economics. Finally, for empirical illustration, we apply the extended estimators to a prior published work.

In this study, we explore whether the issues associated with staggered designs also apply to nonlinear outcome models, and compare the relative performance of the standard TWFE-DiD/ES estimator to that five alternative estimators (proposed by Sun & Abraham, 2021; Callaway & Sant'Anna, 2021; Wooldridge, 2023; Borusyak et al., 2021; and Stacked regression), by employing Poisson quasi maximum likelihood estimator for count outcome and conditional logistic regression for binary outcome. Our simulation results confirm that the

issues associated with TWFE-DiD/ES with staggered designs also apply to nonlinear outcomes models for both DiD and ES designs. We also find that if alternative estimators are used straight of the shelf for count and binary outcomes, some of them produce biased estimates. In summary, the simulation study provides a comprehensive analysis of estimating treatment effects in the context of staggered interventions with count and binary outcomes.

Our findings have significant implications for applied research involving limited dependent variables. These results provide valuable insights into the performance of various estimators in the context of count and binary outcomes. They offer guidance for applied researchers, with the caveat that the extensions of the estimators employed in this paper were not proposed in the original papers cited, nor had their properties formally studied (except Wooldridge (2023)). Moreover, the paper emphasizes the need for further research to explore the (asymptotic) properties of the extended alternative estimators.

Estimating effects of staggered intervention with count and binary outcomes: a simulation study

Anil Yadav
Central Bank of Ireland; University of Galway, Ireland

John McHale
University of Galway, Ireland

Jason Harold
University of Galway, Ireland

Stephen O'Neill
London School of Hygiene & Tropical Medicine, London, UK

Abstract

Difference-in-Differences and Event-study methods with staggered intervention may provide biased estimates when these approaches are implemented using a two-way fixed effect (TWFE) estimator in the presence of heterogeneous effects. Recent literature proposed alternative estimators that are unbiased, however to date, attention has primarily focused on linear outcome models. This study addresses this gap by extending five of these alternative estimators to count and binary outcomes and assessing their accuracy against the TWFE estimator in Monte Carlo simulations. While unbiased for linear models, some of the estimators yield biased estimates for nonlinear outcomes. An application revisits the statistical association between citations and star coauthorship.

JEL Codes: C13; C18; C22; C23; C35
Keywords: Nonlinear difference-in-differences and Event-study; Staggered intervention; Count and Binary outcomes; Treatment effect heterogeneity;

* Corresponding author: Anil Yadav. Email: anil.yadav@centralbank.ie; North wall quay, Dublin 1, Central Bank of Ireland; Disclaimer: The views expressed are those of the authors and do not necessarily represent the views of the Central Bank of Ireland.

Acknowledgements
We gratefully acknowledge funding from Science Foundation Ireland under the SFI Science Policy Research Programme, Grant Number 17/SPR/5329. For helpful comments and discussions, we are indebted to participants of International Panel Data conference 2023, Irish Economic Analysis conference 2023, Scottish Economic Society annual conference 2024, and audiences at several seminar presentations.

1. Introduction

Difference-in-Differences (DiD) is one of the most widely used quasi-experimental designs employed to estimate causal effects (Abadie, 2005). The difference in outcomes between treated (those exposed to treatment) and control (those not exposed) units is compared before and after a policy intervention, with the difference in these differences attributed to be the effects of the intervention (Athey and Imbens, 2006). In the canonical DiD set-up, effects of the intervention are estimated over the period following the policy intervention, implicitly assuming that effects are homogenous or that variation in effects is not of direct interest and does not bias the overall treatment effect estimates. Where data is available for multiple periods before and after the intervention, effects in the literature are frequently estimated using an event-study (ES) design (see de Chaisemartin & d'Haultfoeuille, 2020). An ES provides treatment effect estimates in each period both before (during which effects should be null when the parallel trend and no-anticipations assumptions hold) and after the intervention. ES therefore conveniently allows researchers to explore the dynamic effect of a policy intervention, which is how the effect varies across different lengths of exposure to the treatment.

Recent developments in the literature on both DiD and ES have raised important concerns that when these approaches are implemented using two-way fixed effects (TWFE) models, they may provide biased effect estimates if effects are heterogeneous and the intervention is adopted or implemented in a 'staggered' fashion, that is at different points in time for different units or cohorts (Goodman-Bacon, 2021; de Chaisemartin & d'Haultfoeuille 2020; Borusyak et al., 2021). In essence, this literature shows that TWFE estimators rely on a weighted average of unit-level treatment effects and that earlier treated cohorts implicitly act as 'controls' for comparisons involving later treated units within TWFE estimators. Under heterogeneous treatment effects (HTE), these units may be 'bad' control units for such comparisons since the

outcomes for these earlier treated units will reflect their treatment effects, leading to biased effect estimates for the later treated units in comparisons that include them. If ignored, this leads TWFE to provide biased estimates and, in extreme cases, can even result in effect estimates having the incorrect sign.

A number of alternative approaches have been proposed to overcome this limitation of standard TWFE regressions (Borusyak et al., 2021; Gardner, 2022; Sun & Abraham, 2021; Callaway & Sant'Anna, 2021; Wooldridge, 2021; De Chaisemartin & d'Haultfoeuille 2020). The appropriateness of these alternative approaches for staggered interventions has been primarily demonstrated in the context of linear outcome models. However, limited dependent variables (e.g., count or binary) and hence nonlinear outcome models are prevalent in many fields of applied economics and policy evaluation and thus, it is important to understand how these recently developed approaches perform in such nonlinear settings.

Moving from linear to nonlinear outcome models raises a number of issues for DiD /ES, even in non-staggered designs. For instance, Taddeo et al. (2022) highlight that, unlike in linear models, the true effect and the parameter on an interaction between indicators for the treatment group and post-treatment period do not coincide in nonlinear models (Ai and Norton, 2003). Also, the Parallel Trends (PT) assumption underlying DiD may hold for the observed outcome (measured on the 'natural' scale) or for an underlying latent variable (measured on a 'transformed scale') (Barkowski, 2022). An additive effect on one scale may imply a multiplicative effect on another (Ciani and Fisher, 2019). However, these studies have not considered the additional complexities raised by staggered treatment designs. Similarly, the staggered DiD/ES literature has almost exclusively focused on linear TWFE-DiD/ES estimators with a few notable exceptions, such as Wooldridge (2021, 2023).

Given the prevalence of limited dependent variables in applied research, this is an important knowledge gap, which we address in this paper by rigorously comparing the performance of the standard TWFE-DiD/ES estimator to that of a number of recently proposed alternative estimators, focussing on count and binary outcomes. The five estimators considered here as they are frequently cited in the recent DiD/ES econometric literature, and include: the Interaction-weighted estimator (Sun & Abraham, 2021), an Inverse Probability Weighting (IPW) estimator (Callaway & Sant'Anna, 2021), a Stacked Regression (Cengiz et al., 2019; Deshpande and Li, 2019), the Extended-TWFE estimator (Wooldridge, 2023), and the Imputation estimator (Borusyak et al., 2021). Currently, each of these proposed alternative estimators uses linear fixed effect models, with the exception of Wooldridge (2023).

Our simulation results show that standard TWFE-DiD and TWFE-ES model for count and binary outcomes produces biased estimates under staggered interventions when effects are heterogeneous. This result confirms that the negative weighting problems arising from 'bad controls' units also occur for limited dependent variables and nonlinear outcome models. Importantly the simulation shows that under homogeneity in effects all estimators (extended to account for the non-linearity of the outcome) estimate unbiased coefficients. Turning to the case of heterogeneous treatment effects, we find that the form of heterogeneity is important. With nonlinear outcome models, TWFE-DiD produces biased estimates for any form of heterogeneity and the TWFE-ES design produces unbiased estimates when effects vary across time, which aligns with staggered designs for DiD and ES for linear outcome models (Baker et al., 2022). However, when effects vary across cohorts or across both cohort-period, only the interaction-weighted, IPW, and extended-TWFE estimators produce unbiased estimates for both count and binary outcomes. The performance of interaction-weighted, IPW and extended-TWFE estimators for count and binary outcomes is parallel with the estimators' performance

for linear outcome model, which implements ordinary least squares (OLS) to estimate the effects.

This study makes a number of contributions. First, we explore whether the issues associated with staggered designs also apply to nonlinear outcome models and find (as expected) that they do for both DiD and ES design. Second, we extend the alternative estimators to account for non-linearity in outcomes by applying them using nonlinear outcome models. For example, we use the Poisson quasi-maximum likelihood estimator (QMLE) for count outcomes, and the Conditional Logit Fixed Effect (CLE) estimator for binary outcomes for each of the recent estimators except the IPW estimator. Third, we show the extension of IPW estimator to count and binary outcomes for DiD and ES approaches. Fourth, we compare the performance of the (extended) methods, using % bias and root mean squared error, within a carefully designed Monte Carlo simulation study. Finally, we apply TWFE and the alternative estimators in an empirical case study, revisiting Yadav et al. (2023)'s empirical analysis that examined how co-authorship with a co-located star scientist affects the co-author's productivity. Our estimations complement Yadav et al. (2023) and suggest that co-authoring with a star scientist has a positive effect on the co-author's productivity, with moderate differences found in the magnitude of the effects across the methods.

The remainder of the paper is outlined as follows. In section 2, we briefly introduce the canonical DiD and PT assumptions under non-linearity, discuss the issues related to DiD, and extend to the event-study (ES) framework. Section 3 discusses the alternative estimators to estimate the causal effect under heterogeneity. Section 4 presents the Monte Carlo simulation design. We present the results in section 5. Conclusions are provided in section 6.

2. Methods

2.1 The DiD estimator

We first introduce the traditional (canonical) 2 × 2 DiD where there are two time periods, $t = \{1,2\}$, with treatment occurring only in the second period for units in the treated group, $D_i = 1$, while other units remain untreated in both periods (control group), $D_i = 0$. We denote $Y_{it}(0)$ and $Y_{it}(1)$ as unit i's the potential outcomes under each treatment status. The observed outcome is given by $Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$. In the absence of covariates, we can write the two-way fixed effect (TWFE) DiD regression as follows:

$$Y_{it} = \alpha_i + \theta_t + \beta D_{it} + \varepsilon_{it} \qquad (2.1)$$

Where D_{it} is an indicator for whether unit i is exposed to the intervention in period t, α_i and θ_t are unit and period fixed effects that account for time invariant and period specific unobserved confounders respectively.[1] It is straightforward to show that the estimated coefficient $\hat{\beta}$ obtained from equation (2.1) is equivalent to $\hat{\tau}$,an estimate of the estimand of interest, which is the average treatment effect on the treated (ATT) in period t in the canonical DiD setup (Roth et al., 2023):

$$\tau_t = \mathrm{E}[Y_t(1) - Y_t(0) | D = 1]$$

2.2 DiD estimator for staggered interventions

The TWFE model in equation (2.1) can be readily extended to account for many periods or units. However recent advances in the DiD literature (Roth et al., 2023; Baker et al., 2022) have highlighted the potential for bias in staggered designs. Violation of the parallel-trends and the no anticipation (i.e. absence of pre-effects) assumptions leads to bias. However, even where these assumptions hold, bias may ensue. TWFE regression yields unbiased estimates of the

1. In the simple 2 periods, 2 groups (2x2) case, this can be simplified to

$$Y_{it} = \alpha + \delta D_i + \gamma P_t + \beta D_i * P_t + \varepsilon_{it}$$

where P is a post-intervention indicator equal to 1 for period 2 and 0 in period 1, and δ captures time-invariant differences between the groups' outcomes, γ captures differences between the two periods common to both groups.

average treatment effect for the treated (ATT) when there is homogeneity in treatment effects across time and units (Goodman-Bacon, 2021; Borusyak et al., 2021). However, issues arise when treatment effects are heterogeneous and units are exposed to the intervention at different time points.

In staggered designs with HTEs, estimated counterfactual outcomes for later treated cohorts based on the standard TWFE model implicitly rely on outcomes for earlier treated units and thus are biased by variation in effects overtime, making these units 'bad' controls. Numerous studies have shown the TWFE-DiD estimator (β) is implicitly a weighted average of a number of different 2×2 DiD treatment effects, where the weights can be negative in the presence of HTEs since TWFE implicitly differences out some of these heterogeneous effects (Goodman-Bacon, 2021; de Chaisemartin & d'Haultfoeuille, 2023). The magnitude of weights depends on a number of factors including the time of treatment, the relative size of each treatment cohort, and the number of time periods (Goodman-Bacon, 2021). The implicit counterfactual underlying the TWFE estimate picks up changes in effects affecting earlier treated units, making them 'bad control' units for later comparisons, which feed through into the overall TWFE-DiD estimate. Several approaches have been proposed which seek to avoid using 'bad control' units when constructing counterfactuals and hence avoid biased effect estimates (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021; Cengiz et al., 2019; Borusyak et al, 2021; Wooldridge, 2021; *inter alia*). We briefly describe five of these approaches in section 3 below, however this is an area of very active research[2].

2.3. Event study design and potential outcome framework

Event-study (ES) designs extend DiD designs by estimating effects for each pre- and post-intervention period (Schmidheiny and Siegloch, 2019). In current practice, researchers use the

2. see Roth et al., 2023, for a recent review

classical TWFE-ES specification (Borusyak et al., 2021). This specification allows treatment effects to vary over time, and can be written as:

$$Y_{it} = \alpha_i + \delta_t + \sum_{j=-k}^{-2} \beta_j D_{it}^{j} + \sum_{j=0}^{l} \beta_j D_{it}^{j} + \varepsilon_{it} \qquad (2.2)$$

Assume g_i captures the first period in which the unit is exposed to the intervention, then $D_{i,t}^{j}$ is a relative-period treatment indicator that takes a value of 1 for unit *i* in the period *j* periods since unit *i* was first exposed to the intervention, and zero otherwise (i.e. $D_{i,t}^{j} = 1\{t - g_i = j\}$), k and l are positive constants where k is the maximum number of leads and l is the maximum number of lags. We exclude a relative period ($D_{i,t}^{-1}$) to avoid multi-collinearity, in which case effects are expressed relative to the effect in this period (which should be 0 in the absence of anticipation effects). In this specification, the researchers are interested in the coefficient of β_j for periods $j \geq 0$, and interpret these coefficients as the ATT at different periods of exposure since the treatment (Callaway & Sant'Anna, 2021).

Since the approach includes TWFEs, the combination of staggered intervention and HTEs also biases TWFE-ES (Baker, 2022). The standard TWFE-ES specification (equation 2.2), unlike specification (2.1), yields a sensible causal estimand when there is staggered adoption and homogeneity in treatment effects or heterogeneity across time. However, when there is heterogeneity across adoption cohorts, the coefficients, β_j, are difficult to interpret for two reasons. First, the ES suffers from the negative weighting issues discussed above and, second, β_j coefficients can be contaminated from the effects of other periods which influence the estimated counterfactual if unaccounted for (Sun & Abraham, 2021).

2.4 DiD and ES estimators for limited dependent variables

When Y_t is restricted in some way (e.g. bounded at zero or binary), the linear PT assumption may be unrealistic (Wooldridge, 2023). When implementing nonlinear DiD models for limited

dependent variables, the applied researcher must determine whether it is plausible that the PT assumption holds either on the natural scale or on a transformed scale (Barkowiski, 2021) based on intuitive knowledge of the empirical study. In this study, we focus on the scenario where the PT assumption holds on the transformed (latent) scale and perform the simulations accordingly[3].

Deb et al. (2017) show that nonlinear models are more appropriate for estimating causal effects when the outcome is bounded (e.g. count or binary). Researchers commonly use TWFE-DiD/ES within nonlinear models such as Poisson or Logit, to estimate treatment effects for nonlinear outcomes e.g. count[4] or binary outcomes[5] (Taddeo et al., 2021; Wooldridge, 2023). However, nonlinear TWFE-DiD/ES estimators have not been thoroughly examined within staggered DiD/ES designs in the context of heterogeneous treatment effects. Recent alternative estimators have primarily focused on linear TWFE-DiD/ES models, with the exception of (Wooldridge, 2023).

3. Alternative estimators

We employ five estimators in this study: Interaction-weighted estimator (Sun & Abraham, 2021), Inverse-Probability Weighting (IPW) estimator (Callaway & Sant'Anna, 2021), Stacked Regression (Cengiz et al., 2019), Extended-TWFE estimator (Wooldridge, 2021), and Imputation estimator (Borusyak et al., 2021). We briefly describe the approach taken by each estimator to identify the treatment effect (see appendix B for a more detailed description of the estimators).

3. In appendix A, we discuss the PT assumption in the case of 2 × 2 DiD for nonlinear DiD model.
4. For count outcomes, a common approach is to assume an exponential mean function, where the nonlinear DiD specification can be written as:

$$E[Y_{it}] = \exp(\alpha_i + \theta_t + \beta D_{it})$$

5. For binary outcomes, a nonlinear DiD model that respects the bounded nature of the outcome variable is appropriate such as logit model. The nonlinear DiD for binary outcome specification with $\Lambda(.)$ representing the logistic function can be written as:

$$Y_{it} = \Lambda(\alpha_i + \theta_t + \beta D_{it} + U_{it})$$

3.1 Interaction-weighted estimator:

The *Interaction-weighted* estimator (Sun & Abraham 2021) is a three-step approach. In the first step, a model is estimated that includes interactions between cohort and relative time indicators allowing us to estimate effects for each cohort in each relative time period using the last pre-intervention period as the comparison period by default. Then in the next step, a set of weights is calculated based on the sample share of cohorts under treatment in each relative time period. In the final step, the weights are then multiplied by the cohort-specific treatment effects to estimate the overall (i.e. for all units across all periods) or dynamic (i.e. relating to a particular duration of exposure) effects pre- and post- treatment.

3.2 Inverse-probability weighting (IPW) estimator:

The *Inverse-probability weighting* estimator (Callaway & Sant'Anna, 2021) is also a two-step approach that estimates ATT parameters.[6] In the first step, ATTs for each group (defined by the period in which they are first treated) in each time period ($ATT(g,t)$) are estimated using separate 2x2 comparisons using the last pre-intervention period for comparison by default. In the second step, a weighted aggregate of $ATT(g,t)$ is calculated to determine overall and/or dynamic pre- and post- treatment effects, where the weights are equal to each cohort's sample share in the relative period.

3.3 Stacked Regression:

In *stacked regression* (Cengiz et al., 2019; Deshpande and Li, 2019), cohort-specific datasets are created for each treatment cohort g and only "clean controls" (i.e., observations in which no exposure has occurred at the relevant time point) are included. These are chosen separately for each treated cohort g over a specific window from k_a periods before the treatment to k_b periods after the treatment, and then each dataset is stacked together. This stacking re-centers

6. The authors also propose a doubly robust estimator, however this would need to be adjusted to account for a nonlinear outcome model. We focus on the IPW estimator here.

each treatment date such that the data is balanced in relative time. While this estimator does not suffer from bias from differential timing of treatment adoption, it might still suffer from bias under heterogeneous effects across calendar-time or cohorts (Gardner, 2022; Wing, 2024). In the stacked dataset, some observations appear multiple times, therefore, standard errors are clustered at the unit-cohort level.

3.4 Extended Two-way Fixed Effect estimator:

The two-step *extended-TWFE* estimator (Wooldridge, 2021) makes the standard TWFE model more flexible to account for heterogeneity. First, the estimator estimates an interaction specification with interactions between cohort and period dummies to estimate effects for each cohort and each period using all pre-intervention periods for the comparison group. Then, the overall effect or dynamic effects are obtained from the weighted-aggregate of each cohort and period effects, where weights are defined by the sample share of cohorts under treatment. Wooldridge (2021) proposes the *extended-TWFE* estimator focusing on linear outcome models, with a brief sub-section on the extension of the estimator for nonlinear outcomes. Wooldridge (2023) formally modifies the *extended-TWFE* estimator to allow for nonlinear outcomes. In the paper, we use the nonlinear estimator proposed in Wooldridge (2023).

3.5 Imputation estimator:

The *imputation estimator* (Borusyak et al., 2021) involves three steps. First, the potential outcomes for treated units are estimated from a fixed effect (unit and time fixed effect; include covariates if available) parametric regression using data for all units (treated and control) but excluding any observations in which the unit has already been exposed to the intervention. Second, we impute the 'never treated' potential outcome for each unit using the prediction from the model in step one and estimate the individual treatment effect for treated units in each post-intervention period as the difference between the observed outcome and the 'never treated' group's potential outcome. Third, calculate the weights for each unit for each period under

treatment, with weights corresponding to the estimation target, and estimate the weighted average of individual level effects based on target estimand – overall effect and dynamic effect in our case.

While each of the five estimators explicitly focuses on using only the "good controls" in their approaches, there are differences across estimators in terms of: (i) data requirements, (ii) how they incorporate covariates, (iii) trade between robustness and efficiency, (iv) appropriate control group, and (v) the assumption of PT. For example, for the PT assumption, by default the interaction-weighted estimator imposes PT in the period right before treatment, the extended-TWFE estimator imposes PT assumption in all time periods, and the IPW estimator imposes PT conditional on covariates.

4. Simulations

We conduct simulation studies where the true ATT is known to assess the relative performance of standard TWFE-DiD/ES estimators and the alternative estimators outlined above in staggered treatment designs. We generate panel data for 1000 units for 20 time periods, which we refer to as 'years'. The units are assigned into six treatment cohorts, indexed by g representing the periods in which they are first treated and one 'never treated' control group[7]. The treatment cohorts are $g \in \{3, 6, 9, 12, 15, 18\}$. The corresponding number of units in each treated cohort are $N_g = \{120, 70, 140, 40, 60, 170\}$ with 400 units assigned to the control group[8].

We focus here on two common types of nonlinear outcome types, count and binary. The data generating process (DGP) satisfies the parallel-trends, no anticipation, and common shock assumptions. The parallel-trend holds on the transformed scale, that is, the underlying latent

7. In our simulations, we use only never-treated units as controls.
8. We examined the sensitivity of the results to different choices of the size of cohorts and found that results are robust.

linear scale for the binary outcome and the log-scale for the count outcome. The true treatment effects ($ATT_{g,t}$) are held constant for count and binary outcomes in the latent scale and are described in section 4.4 below.

4.1 Count outcome

For the count outcome, the DGP for the potential outcome variable with exponential mean function is[9]:

$$Y_{i,t} = \exp(\alpha_i + \alpha_t + ATT_{g,t}D_{i,t}) \qquad (4.1)$$

where. $Y_{i,t}$ is drawn from a Poisson distribution. α_i and α_t are unit and time fixed effects, where α_i are drawn from $\sim N(0,1) + \mu_{cohort} + e_{it}$ and μ_{cohort} is a cohort-specific difference[10] , i.e. each cohort is generated with different intercept, e_{it} is the error term for each unit, α_t are time fixed effects that do not vary by individual and are drawn from $\sim N(0,1) + 0.2 * t$. $ATT_{g,t}$ is true known effect for cohort g at time t. D_{it} is an indicator of whether unit i is exposed to the intervention in period t. $ATT_{g,t}$ is defined by scenarios (discussed below) i.e., how the ATT evolves for each cohort post-treatment.

4.2 Binary outcome

Following Fernández-Val and Weidner (2016), the DGP for binary outcome $Y_{i,t}$ is:

$$Y_{i,t} = 1[\alpha_i + \alpha_t + ATT_{g,t}D_{i,t} + U_{i,t} > 0] \qquad (4.2)$$

where $\alpha_i \sim N(0,1/16) + \mu_{cohort} + e_{it}$, $\alpha_t \sim N(0,1) + 0.2 * t$, $U_{i,t}$ is the logistic error term, and $ATT_{g,t}$ is the true effect.

9. We have also considered non-negative and skewed fractional (continuous) outcomes using the following DGP:

$$Y_{i,t} = \exp(\alpha_i + \alpha_t + ATT_{g,t}D_{i,t} + e_{i,t})$$

The result are similar to Count outcome. Due to space limitations, we limit our reporting to Count outcome.

10. In appendix C1, we discuss the cohort-specific imbalance and provide the simulation parameters.

4.3 Modelling

We extend the alternative estimators (with the exception of the IPW estimator (see below)) described above by employing fixed effect Poisson quasi-maximum likelihood estimation (QMLE) for count outcomes and the Conditional Logit fixed effect (CLE) for binary outcomes in place of the ordinary least squares (OLS) estimator used for linear models by these approaches. For illustration purpose, we show the bias in the TWFE-DiD and alternative estimators if a linear model using OLS is used to estimate effects for count and binary outcomes (see appendix D1). For count outcomes, Poisson QMLE has many desirable properties: the coefficient estimates remain consistent as long as the mean of the dependent variable is correctly specified independent of any assumption on the conditional variance (Wooldridge, 1999), while standard errors are consistent even if the underlying data generating process is not Poisson (Gourieroux et al., 1984). While CLE is consistent and asymptotically normal for binary outcome (Allison, 2009). Units whose outcomes do not vary are dropped when using this approach, altering the population to which estimates apply[11]. The Poisson QMLE and CLE regressions provide coefficient estimates interpreted as log rate-ratio and log odds-ratio for count and binary outcomes. Our simulation results (section 5) are based on log rate-ratio for the count outcome and log odds-ratio for the binary outcome.

The fixed effect Poisson QMLE and CLE cannot be used directly for the IPW estimator as the IPW estimator is not a regression-based estimator but a weighting based technique. We extend the IPW estimator to account for the distribution of the outcome variable for both count and binary outcomes (see appendix D2 for details). The estimates from the IPW estimator are on the same scale – log rate-ratio for the count outcome and log odds-ratio for the binary outcome – as alternative estimators.

11. In our simulations, the DGP for the binary outcome is designed such that the number of units dropped is minimized.

4.4 Scenarios for treatment effects evolution

In this paper, a key focus is on overall and dynamic effects in the DiD and ES settings. Therefore, we consider different treatment effects scenarios can vary across intervention cohorts, time since first treatment exposure, or both, which yields four main scenarios for the generation of $ATT_{g,t}$[12] (see appendix C2 for the treatment effect evolution by scenario):

- **Scenario A (Homogeneous treatment effects)**: $ATT_{g,t}$ is a constant. In this scenario, the treatment effect is constant for all cohorts in post-treatment period.
- **Scenario B (Heterogeneity across time since exposure)**: $ATT_{g,t}$, differs only by exposure duration. This type of heterogeneity can occur, for instance, if treated units learn about improving outcomes over the period since exposure and the effects compound over time.
- **Scenario C (Heterogeneity across intervention cohorts)[13]:** $ATT_{g,t}$ differ only by cohort (defined as a group that is exposed for the first time in the same period). For instance, early adopters may be those that expect larger gains from the intervention. Calendar effects or Selective-timing effects can be considered an example of this scenario of treatment effect when the target estimand is overall treatment effect or dynamic effect.
- **Scenario D (Heterogeneity across time since exposure and interventions cohorts):** $ATT_{g,t}$ differs *both* by exposure duration and intervention cohorts.

12. In appendix C3, we provide a stylized example for each scenario of treatment effects evolution.
13. In this scenario, the treatment effects vary across cohorts treated in different time points. For example, a time step can be monthly, quarterly, by-annually, or as in our simulations yearly. Therefore, the scenario captures this heterogeneity across cohorts treated in different time steps, in other words, captures the heterogeneity in calendar time. Calendar effect or Selective timing effects can be considered as an example of this scenario of treatment effect evolution, where the "intervention across cohorts" tries to capture all effects that could vary across time based on treatment adoption, assuming the effect is homogenous across duration since treatment.

Our analysis is based on 500 simulations for each scenario. We evaluate the performance of the standard TWFE-DiD and TWFE-ES model and each of the five alternative estimators (described in section 3) to examine the performance of each estimator based on bias and root mean squared error (RMSE). For the DiD model we compare the approaches according to % bias, while we use bias for the ES model. In the ES model, the true effect in pre-treatment periods is zero, making it impossible to calculate % bias in pre-treatment periods. Thus, the bias for each relative pre and post-treatment period in ES model is computed as:

$$Bias_j = \sum_{j=-k}^{j=l} (\beta_j - \widehat{\beta_j})$$

where k and l are maximum number of leads and lags, β_j is the true effect, and $\widehat{\beta_j}$ is the average estimated effect for relative period j across 500 simulations.

5. Simulation Results

5.1 Count outcome

5.1.1 Difference-in-Difference (DiD)

Figure 1 presents the boxplots of % bias in treatment effects from the simulations for DiD estimand for count outcome for each scenario by estimator, while Table 1 reports the corresponding mean % bias and RMSE. Appendix E shows the boxplots of estimated effects from simulations. The effects are interpreted as log rate-ratio. We begin with the homogenous effect scenario (Figure 1, Panel A). Standard TWFE-DiD regression performs well and shows very low bias. The other estimators also perform very well for this scenario. When effects are heterogeneous (scenario B, C, and D; panel B, C, and D in Figure 1), standard TWFE-DiD performs very poorly and does not correspond to a weighted average of causal effects. Interaction-weighted, IPW, and extended-TWFE estimators continues to perform well with low

bias and low RMSE under heterogeneous effects, whereas stacked regression and Imputation estimator report biased estimates (see appendix E). For standard TWFE-DiD, stacked regression, and imputation estimator, the bias is attributable to heterogeneity in effects across time and cohorts, and the estimators are unable to capture this heterogeneity. While stacked regression estimator circumvents the issue of bad control units, however, suffers from nonlinearity of outcome which leads to biased estimates even in case when effects are heterogeneous across time (scenario B). Furthermore, Baker et al. (2022) pointed that stacked regression implicitly assigns the weights that may not correspond to cohort-proportional weights which could lead to biased estimates when effects are heterogeneous.

5.1.2 Event-Study (ES)

For the homogeneous effect scenario (scenario A) Figure 2 shows the ES plots of estimated coefficients from the simulations for count outcome for each estimator[14]. The point estimates represent the mean of the estimated coefficients from the simulations. The 95% confidence interval is computed using the standard deviation in the point estimates across 500 simulations. Therefore, the confidence interval shows the variability in point estimates across simulations. When discussing the event-study results, the panels in the graph represent each estimator (Panel A: Standard TWFE-ES; Panel B: Interaction-weighted, Panel C: IPW; Panel D: Stacked regression; Panel E: Extended-TWFE; Panel F: Imputation estimator)[15]. When there are homogeneous effects, (as expected) the standard TWFE-ES regression estimates the true dynamic path of effects with high accuracy. All the alternative estimators perform well for scenario A.

When treatment effects vary across time but are the same across the length of exposure since treatment (scenario B), standard ES regression reports unbiased estimates. Panel A in Figure 3

14. For stacked regression we choose a window of -9 to 10 years for the estimation of the model.
15. Figure E.2 in appendix E shows the mean bias for all estimators and scenarios in the event-study setting.

shows the results for standard ES model. Sun and Abraham (2021) showed that when the effects of each lag (relative year) across cohorts are the same, the TWFE estimation provides unbiased ATT estimates in the standard ES setting. The reason is standard ES model allows the effects to vary across time in relative years since treatment exposure. Figure 3 demonstrates that the alternative estimators perform well for this scenario and provide unbiased ATT estimates.

Figure 4 presents the results for scenario C, where the treatment effects heterogeneity varies across intervention cohorts. Panel A shows that the standard ES estimator reports biased estimates for both lead and lags. This is similar to the cross-lag contamination highlighted in Sun and Abraham (2021) even when the parallel-trends and the no anticipation assumptions hold. Together with the standard TWFE-ES estimator, the stacked regression, and the imputation estimator also produce biased estimates (shown in Panel D and F in Figure 4). The ES graph for TWFE regression and stacked regression follow very similar patterns for the leads and lags. In stacked regression, the weights are implicitly assigned by the regression estimator, therefore, it produces biased estimates even controlling for bad control units. In contrast, the interaction-weighted, IPW, and extended-TWFE estimators produce unbiased ATT estimates with both pre and post treatment periods. In scenario D, when treatment effects vary across intervention cohorts and time, we get results similar to scenario C, where the standard ES model, stacked regression, and the imputation estimator give biased estimates(Panel A, D, and F in Figure 5). Imputation estimator produces low bias in pre-treatment period, however, the bias rises significantly in post-treatment periods. Interaction-weighted, IPW, and extended-TWFE estimators provide unbiased ATT estimates with low bias (Panel B, C, and E in Figure 5).

5.2 Binary outcome

5.2.1 Difference-in-Difference (DiD)

Table 2 reports the mean % Bias and RMSE from the simulations for DiD estimand for binary outcome for each scenario by estimator except for the imputation estimator[16]. Figure 6 presents the corresponding boxplots of % bias, and Figure F.1(appendix F) shows the boxplots of estimated coefficients from the simulations. The coefficient estimates are interpreted as log odds-ratio. When effects are homogenous (Panel A of Figure 6; Panel A of Figure F1.1), each estimator performs well with low bias and low RMSE. In the scenarios with heterogeneous effects (scenarios B, C, and D), interaction-weighted and extended-TWFE estimators continue to perform well, providing estimates with low bias and low RMSE (Table 2; Panel B, C, and D in Figure 6). Standard TWFE-DiD and stacked regression all report biased estimates.

5.2.2 Event-Study (ES)

Same as count outcome, when discussing the event-study results for binary outcome, the panels in the graph represent each estimator (Panel A: Standard TWFE-ES; Panel B: Interaction-weighted, Panel C: IPW; Panel D: Stacked regression; Panel E: Extended-TWFE)[17]. When the treatment effects are homogeneous (scenario A) or vary only across time (scenario B), the standard ES model as well as other estimators perform well and produce estimates with low bias for pre- and post-intervention periods. The ES results presented in Figures 7 and 8 show the results for each estimator for all leads and lags. For scenario C (heterogeneity across cohort; results shown in Figure 9), the interaction-weighted, IPW, and extended-TWFE estimators provide unbiased estimates. However, for this scenario, the other methods – the standard TWFE-ES and stacked regression – perform poorly with high bias produced in both pre and

16. Currently we are not aware of any extension of the imputation estimator for use with binary outcomes. Therefore, in our simulations for binary outcomes, we consider only the standard TWFE-DID/ES model, interaction-weighted, IPW, stacked regression, and extended-TWFE estimators.
17. Figure F.2 in appendix F shows the mean bias for all estimators and scenarios in the event-study setting.

post treatment periods. In scenario D (heterogeneity across cohort and time), again only the interaction-weighted, IPW, and extended-TWFE estimators provide estimates with low bias (see Panel B and E in Figure 10). The other methods suffer from the same problems as in scenario C (heterogeneity across cohort), and give biased estimates (Panel A, C, and D in Figure 10). Overall, for binary outcome, interaction-weighted, IPW, and extended-TWFE are robust to the different scenarios of treatment effect, similar to the results for count outcome.

6. Case Study: Empirical application

To provide an illustration of the results presented above we now present an empirical application. We revisit Yadav et al. (2023)'s (YMO hereafter) empirical analysis that examines how co-authorship with a co-located star scientist affects the co-author's productivity. YMO first implements coarsened exact matching to construct a treated and control group that are comparable in terms of observed characteristics. Then, they employ standard TWFE-ES regression using Poisson QMLE on the matched treated and control group to find that, following coauthorship with a star scientist, a co-author's research productivity increases both including and excluding the output of the star. The authors use unbalanced panel for the main analysis. However, our simulations are based on balanced panel. Therefore, for the application to be parallel with our simulations, we focus on the balanced panel results represented in figure 3 in YMO. The authors presented the results for output including and excluding star's output. For parsimony, our replication focuses only on the results including the star's output.

We use the data provided by YMO, which contains individual authors-level data for 2,458 authors from 1996 to 2017. Forming a co-authorship relationship with a star is used as the treatment in the study. All authors who co-authored with the star for the first time in the same year belong to a cohort. YMO has used the treatment start time since 1996. In our replication, we use treatment start time from 1997 onwards, as most of the alternative estimators drop the

observations for the treated cohort for whom no pre-treatment observations are available. Therefore, to be consistent across the estimators we use the treatment start time from 1997. In total we have 21 cohorts, g = {1997,1998,…,2017}.

6.1 Count outcome

YMO's identification strategy relies on staggered implementation of the treatment (co-authorship with the star) from 1997 to 2017. They implemented a DiD model with staggered exposure:

$$Y_{it} = exp(\alpha + \beta_1 star_{it} + \delta_t + \mu_i + \epsilon_{it}) \qquad (6.1)$$

and a standard ES model specified as:

$$Y_{it} = exp(\alpha + \beta_{\leq -4} star_{i,-4} + \sum_{j=-3}^{-2} \beta_j star_{i,j} + \sum_{j=0}^{3} \beta_j star_{i,j} + \beta_{\geq 4} star_{i,4} + \delta_t + \mu_i + \epsilon_{it}) \qquad (6.2)$$

where $Y_{i,t}$ is the measure of productivity (i.e. field normalized citations) for author i in year t. δ_t is a vector of time-fixed effects, μ_i is the unit-fixed effects and ϵ_{it} is the idiosyncratic error term. The variable $star_{i,j}$ is a star co-authorship indicator variable equal to 1 if, as of year t, an author co-authored with a star j years ago. Equation (6.2) uses 3 leads and 3 lags, with periods more than 3 preceding (proceeding) co-authorship, the authors combine the indicators into a single indicator, $star_{i,-4}$ ($star_{i,+4}$). YMO also implemented extended-TWFE estimator in their analysis, imposing PT in all pre-treatment periods until the last time period. Thus, we remain consistent with the original paper for extended-TWFE replication.

The estimated effects, using Poisson QMLE, of star's co-authorship on co-author's productivity using different estimators are shown in Figure 11 (DiD model) and Figure 12 (ES model). TWFE-DiD shows a positive effect of star co-authorship. While alternative estimators

produce qualitatively similar results, we find moderate differences in the magnitude of effects across the estimators. The interaction-weighted, IPW, and extended-TWFE (most accurate estimators in simulation) estimates show that, after accounting for heterogeneity in treatment effects, the treatment effect is slightly lower than the TWFE-DiD estimates. TWFE-ES graph suggests that following a co-authorship with a star, co-authors experience a statistically significant increase in their quality-adjusted productivity. Alternative estimators also show similar qualitative trend in the results, both in pre and post treatment periods. Our replication complements YMO and suggests that the co-authoring with star scientist positively affects the co-author's productivity, however, it indicates the presence of heterogeneous effects across time and cohorts.

6.2 Binary Outcome

YMO do not use a binary outcome in their analysis. However, for empirical illustration, we use YMO's dependent variable, field-normalized citations, to create a binary outcome. We took log of the dependent variable, then standardise, and assign a value 1 for author *i* in year *t* if the field-normalized citations are above zero (a single median that is defined across all observations), and 0 otherwise. The DiD and ES models are specified as:

$$Y_{it} = \Lambda(\alpha + \beta_1 star_{it} + \delta_t + \mu_i + U_{it}) \qquad (6.3)$$

$$Y_{it} = \Lambda\left[\alpha + \beta_{\leq -4} star_{i,-4} + \sum_{j=-3}^{-2} \beta_j star_{i,j} + \sum_{j=0}^{3} \beta_j star_{i,j} + \beta_{\geq 4} star_{i,4} + \delta_t + \mu_i + U_{it}\right] \qquad (6.4)$$

where $\Lambda(.)$ is the logistic function, $Y_{i,t}$ is the binary outcome, and other parameters are same as equation (6.1).

Figure 13 and 14 shows the results for DiD and ES models using Conditional Logit fixed effect estimator. The TWFE-DiD and the alternative estimators indicate a positive effect from star co-authorship on an author's productivity, however, we find differences in the magnitude of effects across the estimators. Interaction-weighted, IPW, and extended-TWFE estimates are lower than the TWFE-DiD estimates, indicating the presence of heterogeneous effects. The Event-study graph in Figure 14 presents the dynamic results estimated by TWFE-ES and the alternative estimators, and suggests a similar qualitative trend in the results, both in pre and post treatment periods. Both the DiD and ES results indicate the presence of heterogeneous effects in overall and pre and post treatment periods.

7. Conclusion

DiD and ES designs are the most popular approaches in the literature to identify the effect of a treatment/intervention on a treated group (Lee & Lee, 2021). Effects in these designs are commonly estimated using two-way fixed effects models. More recent literature suggests that standard TWFE-DiD/ES regressions are susceptible to producing biases under staggered treatment adoption and HTEs (Baker et al., 2022). While a variety of strategies (estimators) have been proposed to circumvent the bias arising in standard TWFE-DiD/ES regression, the focus of these estimators has been primarily on a linear outcome setting with very few exceptions. Thus, there exists a gap in the literature on the performance of these recent estimators in a nonlinear setting. We aim to contribute to this gap in the literature by focusing on the performance of these recent estimators in a nonlinear outcome setting (count and binary outcomes).

We examine, using simulations, the relative performance of the standard TWFE-DiD/ES model and five alternative estimators (Interaction-Weighted, IPW, Stacked Regression, Extended-TWFE, and Imputation Estimator) under different scenarios of HTEs evolution for count and

binary outcomes. Our simulation results indicate that the negative weighting problems arising from bad control units persist in the standard TWFE-DiD/ES model with staggered interventions and heterogeneous effects for count and binary outcomes. To our knowledge, this study is the first to highlight this finding in the case of nonlinear outcomes[18]. Our findings suggest that applied researchers interested in staggered DiD/ES model using count or binary outcomes should be careful while using standard TWFE-DiD/ES estimators when they suspect the presence of treatment effect heterogeneity. These approaches could lead to biased estimates of the underlying treatment effects and could therefore lead to erroneous conclusions regarding the effectiveness of an intervention.

We, additionally, extend the alternative estimators that are proposed by the recent literature to circumvent the limitations in standard ES model to account for non-linearity in outcomes and examine their relative performance. We use Poisson QMLE for count outcome and Conditional Logit fixed effect estimator for binary outcome for each estimator except for the IPW estimator, for which we have shown the extension for count and binary outcomes. The simulation results show that estimators such as the stacked regression and the imputation estimator that produce unbiased estimates for linear models (continuous outcome) are biased and fail to recover the true treatment effect under non-linearity in outcome in the presence of HTEs. Stacked regression produces biased estimates for all scenarios of heterogeneous effect for the DiD model and for scenarios C and D of treatment effect evolution for the ES model, with both count and binary outcomes[19]. Interaction-weighted, IPW, and Extended-TWFE estimators are found to be most robust in producing unbiased estimates when 'extended straight of the shelf' under staggered interventions for each scenario of heterogeneous treatment effects. Finally, for

18. Studies such as Baker et al (2022), Roth et el., (2023), Borosyak et al (2021), Barrios (2021), Linder & McConnell (2022) uses simulations to show the bias arising from staggered treatment and heterogeneous effects, however, the focus in these paper is on linear (continuous) outcome models.
19. Stacked Regression can produce biased estimates in the presence of HTEs even in the case of linear outcome (Gardner, 2022).

empirical illustration, we revisited YMO (Yadav et al., 2023), where we applied standard TWFE-ES and the alternative estimators to their data. The results from the alternative estimators are qualitatively similar when compared with the results from the original analysis. This could be due to comparable treated and control groups obtained from matching before implementing Poisson QMLE, indicating they strategically obtained similar groups and accounted for outcome distribution carefully in the regression or due to effects being more homogenous than in our simulation study, however, there were moderate differences in the magnitude of effects across the estimators[20].

Overall, our findings reveal that not all alternative estimators that circumvent the limitations of standard TWFE-ES model in linear models recover the true ATT in the case of nonlinear outcome models. In particular, if used straight of the shelf without appropriately accounting for the nature of the dependent variable bias can ensue. Based on these findings, we recommend, while being *careful* that the extensions employed in this study were not proposed by the original papers cited and have not had their properties formally studied (expect extended-TWFE by Wooldridge (2023)), researchers employ the interaction-weighted, IPW, or extended-TWFE estimator (or modify the alternative methods before use to acknowledge non-linearities). Both, interaction-weighted and extended-TWFE, of estimators are regression based and easy to implement, even for nonlinear outcomes, whereas IPW estimator is a weighting-based technique, that can be extended easily to account for non-linearity. The estimators do differ in some key factors such as interaction-weighted estimator impose PT assumption from the last pre-intervention period until the last period, the extended-TWFE imposes a PT assumption in all pre-intervention period until the last period, and IPW imposes a PT assumption conditional on covariates. We recommend that a researcher employ the

20. Chiu et al (2023) in a replication analysis find similar results to our study. The authors replicated 37 studies using TWFE estimator and six alternative estimators for linear outcome model and find that, in general, results from the alternative estimators are qualitatively similar to original study.

estimator most suitable for their empirical analysis based on the setting under investigation. Furthermore in Table G.1 (appendix G), we provide a brief comparison of the estimators examined in this paper.

Our study examined the performance of alternative estimators (initially developed for continuous outcome and linear model) for count and binary outcomes under staggered adoption and HTEs in the context of DiD and ES design. We believe the results provided in our paper can improve the credibility of staggered DiD/ES studies with nonlinear outcomes. However, a formal study of the properties of the extensions of the original five alternative estimators presents an interesting avenue for future research. Moreover, this study is focused on a balanced panel where PT and NA assumptions hold. Often, researchers encounter data with unbalanced panel, time variant/invariant covariates. In future work, we will extend the analysis of the performance of these alternative estimators to more complicated cases, such as different potential violation of PT assumption, fundamentally different control and treated group, covariates.

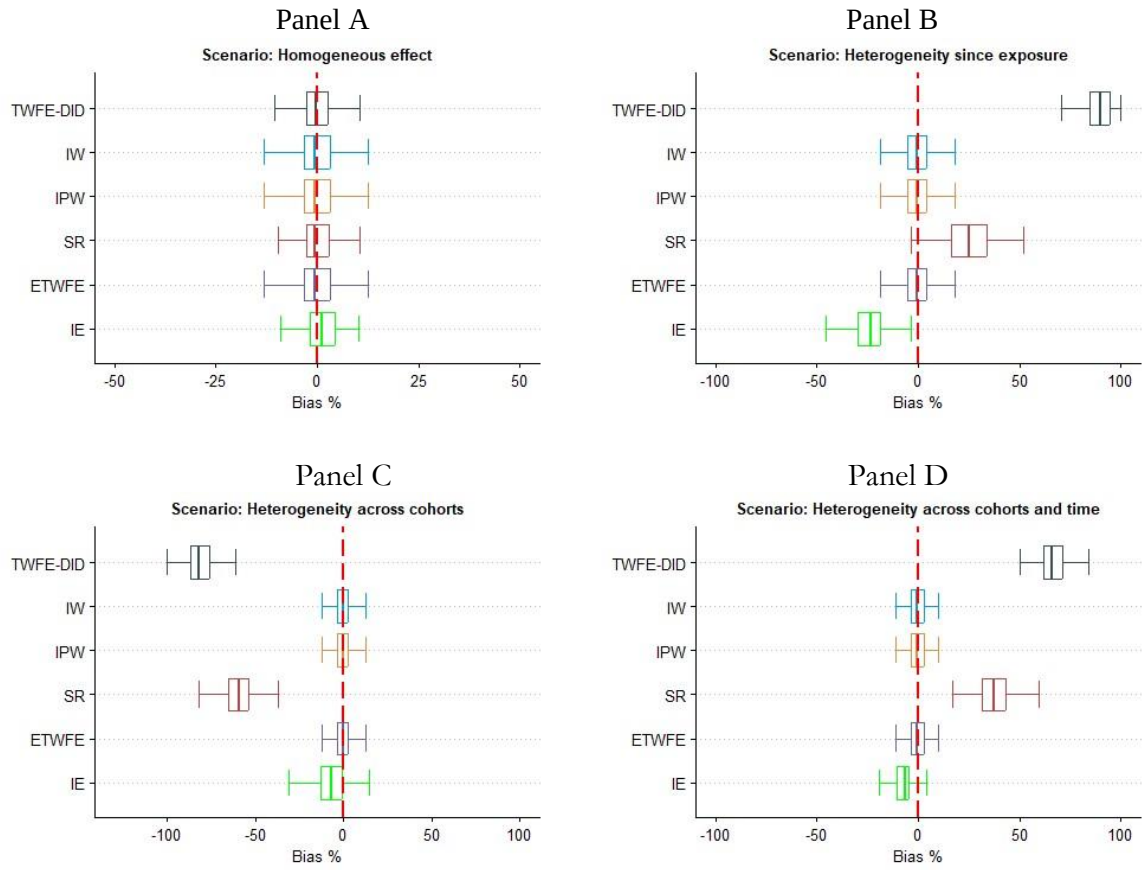

Figure 1: Boxplots of bias % in treatment effects for **count outcome**

Note: IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

Table 1: Mean % Bias and RMSE for DID model for **Count outcome**

	Mean % Bias					
Scenario	TWFE-DID	IW	IPW	SR	ETWFE	IE
Homogeneous effect	-0.080	-0.193	-0.193	-0.040	-0.193	1.024
Heterogeneity across time	92.094	-0.613	-0.613	25.372	-0.613	-24.297
Heterogeneity across cohorts	-81.584	-0.420	-0.420	-59.445	-0.420	-7.236
Heterogeneity across time and cohorts	66.385	-0.355	-0.355	37.634	-0.355	-7.238
	Root Mean Squared Error (RMSE)					
Scenario	TWFE-DID	IW	IPW	SR	ETWFE	IE
Homogeneous effect	0.045	0.052	0.052	0.044	0.052	0.046
Heterogeneity across time	0.772	0.062	0.062	0.164	0.062	0.217
Heterogeneity across cohorts	0.730	0.045	0.045	0.585	0.045	0.107
Heterogeneity across time and cohorts	0.800	0.061	0.061	0.400	0.061	0.103

Note: IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

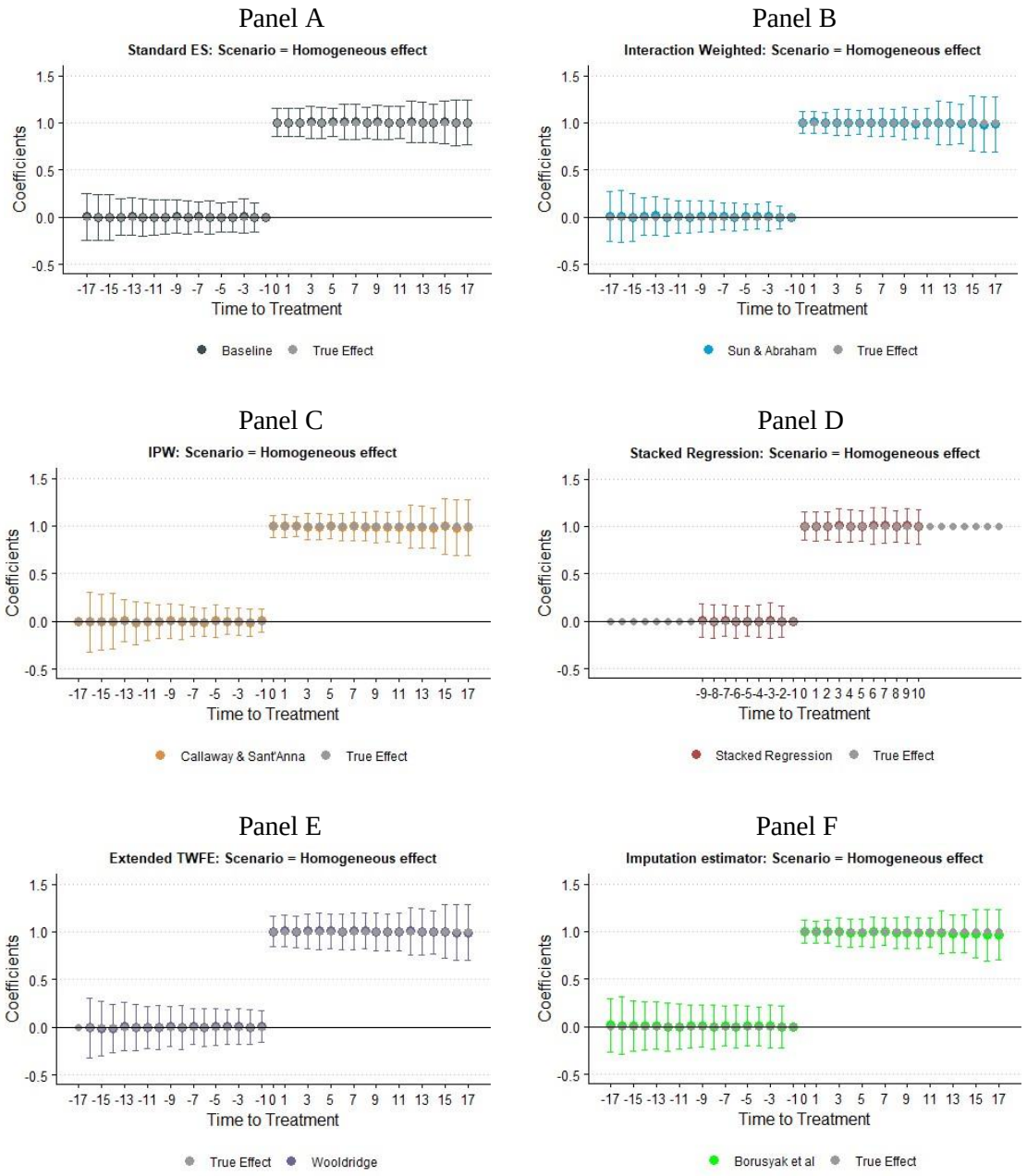

Figure 2: Event-study graphs from simulations for **Count outcome** for **Scenario A (Homogeneous effect)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulations. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd).

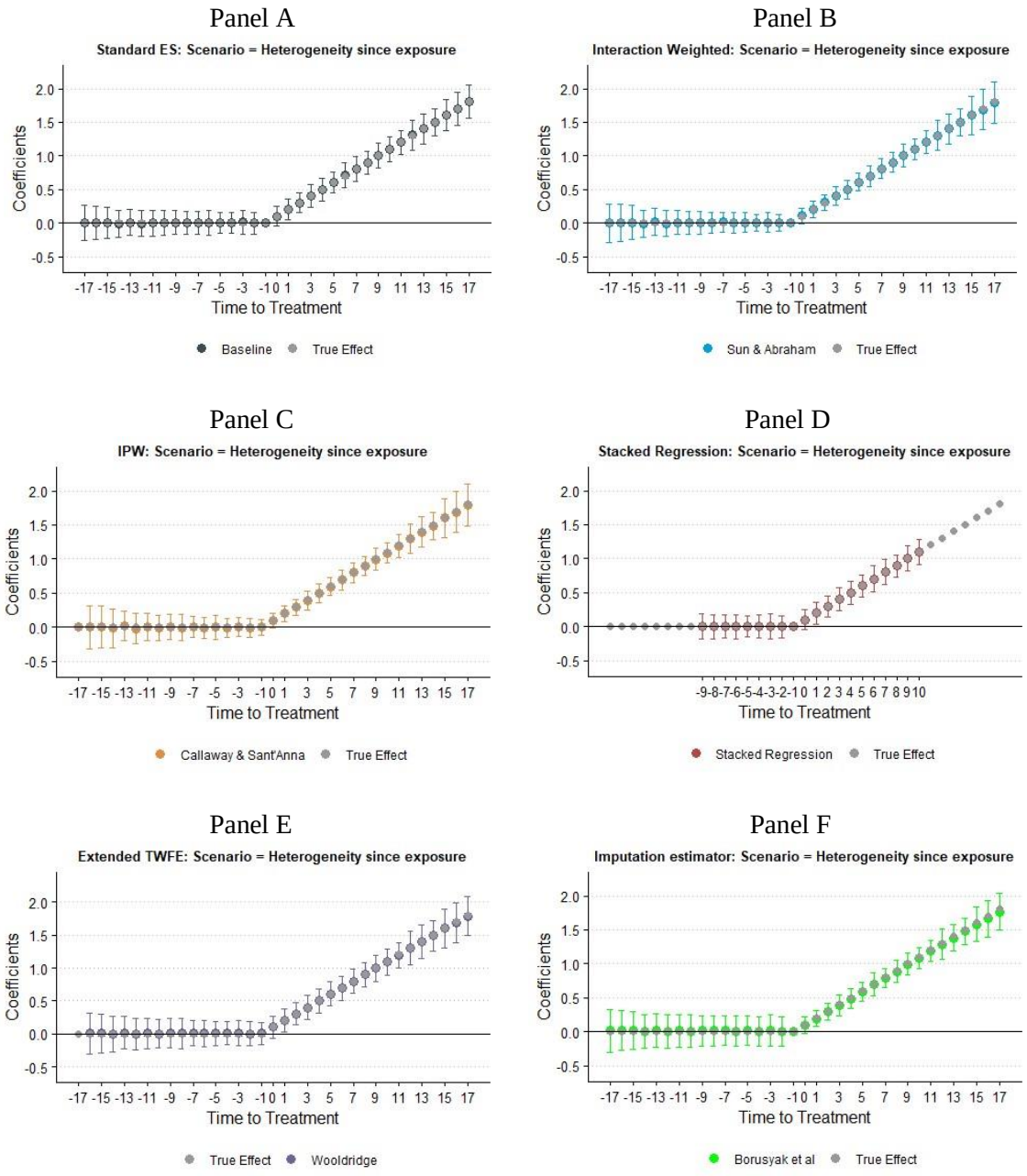

Figure 3: Event-study graphs for simulations for **Count outcome** for **Scenario B (Heterogeneity since exposure)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd).

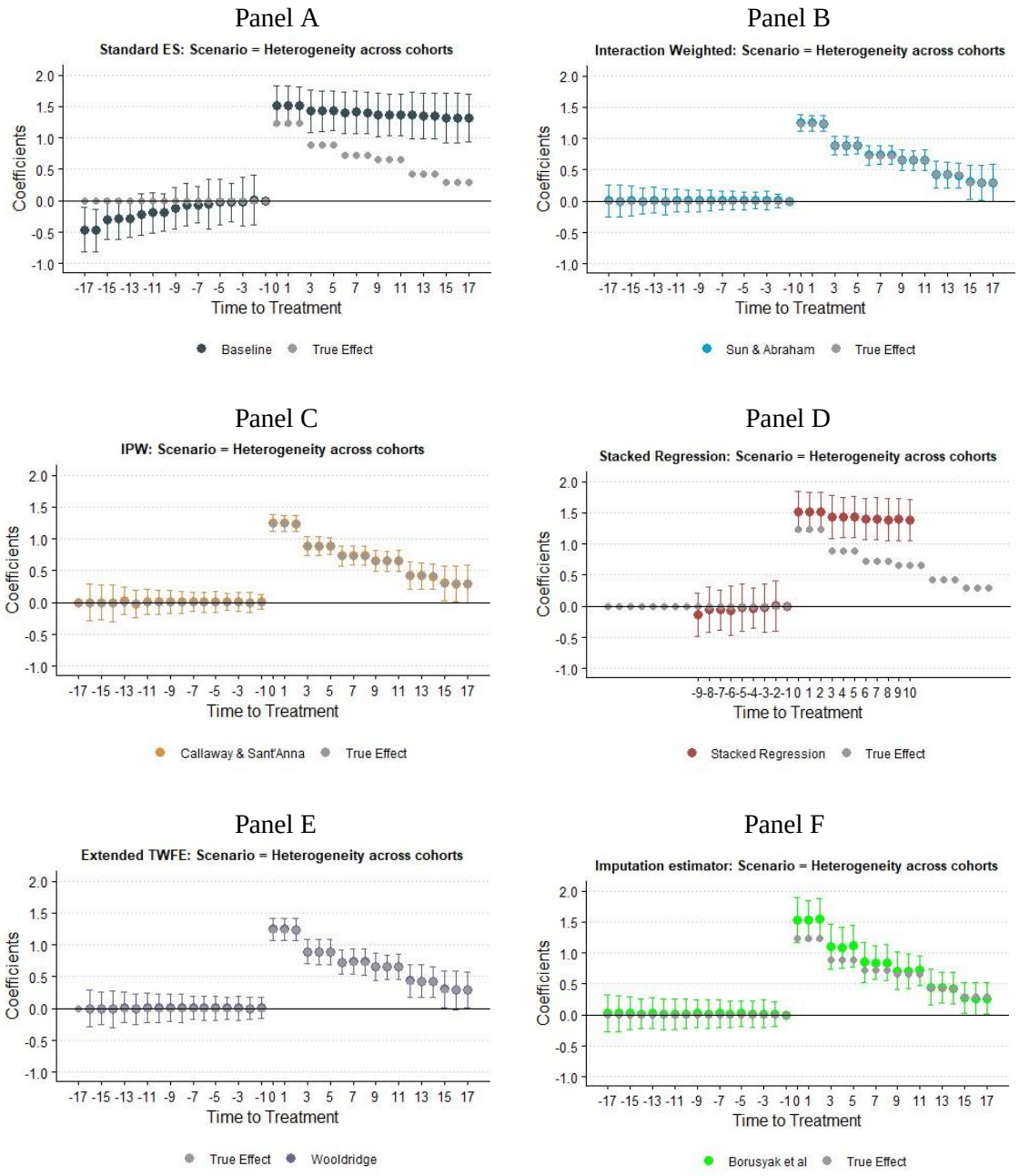

Figure 4: Event-study graphs for simulations for **Count outcome** for **Scenario C (Heterogeneity across cohorts)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd).

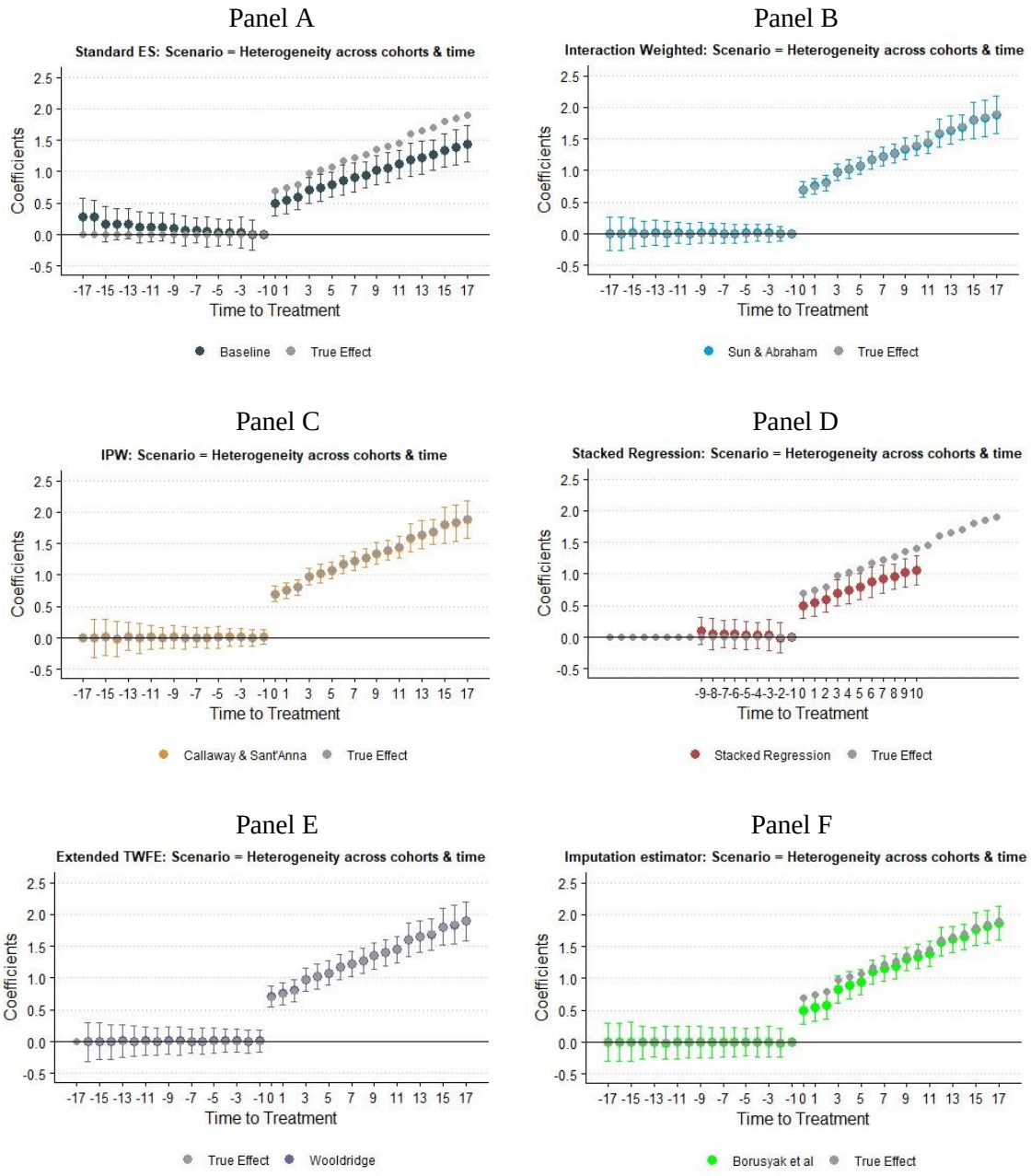

Figure 5: Event-study graphs for simulations for **Count outcome** for **Scenario D (Heterogeneity across cohorts and time since exposure)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd).

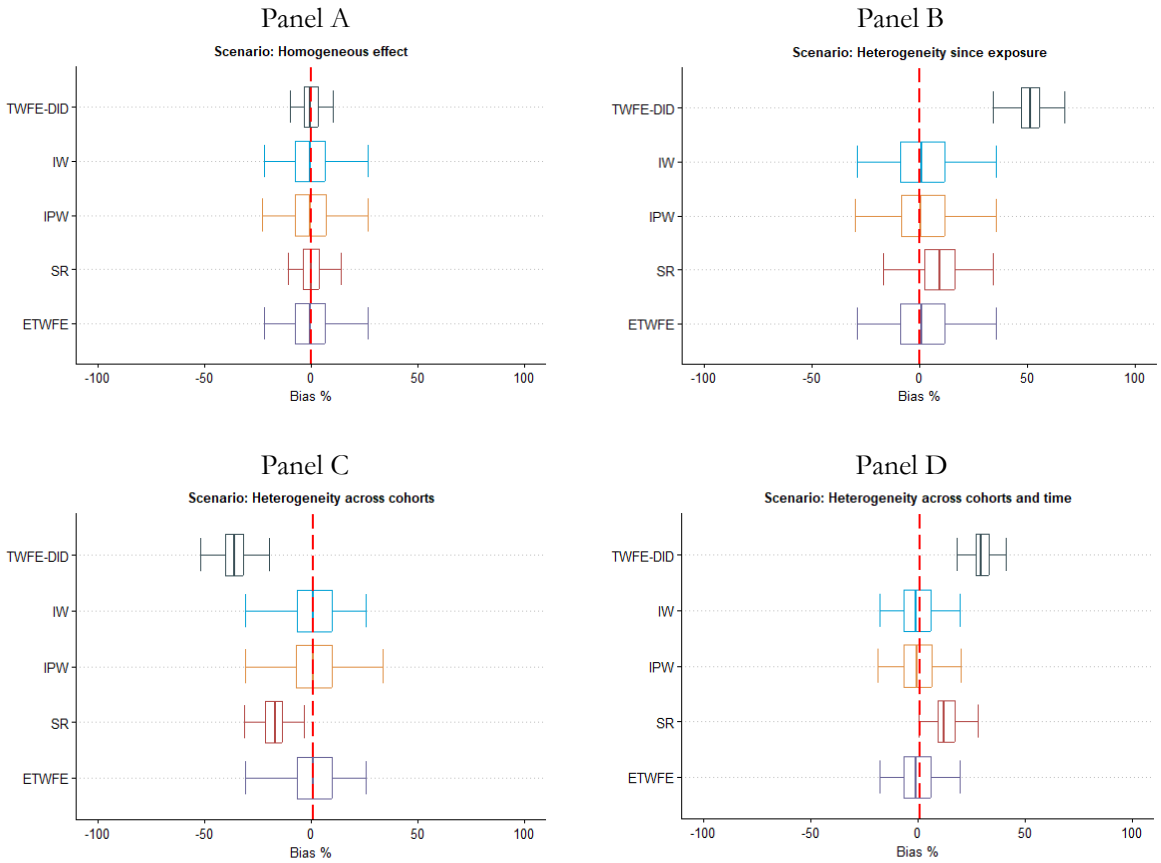

Figure 6: Boxplots of % bias in treatment effects for **Binary outcome**

Note: IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

Table 2: Mean % Bias and RMSE for DID model for **Binary** outcome

	Mean % Bias				
Scenario	TWFE-DID	IW	IPW	SR	ETWFE
Homogeneous effect	0.558	0.237	0.380	0.566	0.237
Heterogeneity across time	52.130	0.898	1.050	10.290	0.898
Heterogeneity across cohorts	-35.135	1.152	1.285	-17.231	1.152
Heterogeneity across time and cohorts	30.495	-0.078	0.099	13.622	-0.078
	Root Mean Squared Error (RMSE)				
Scenario	TWFE-DID	IW	IPW	SR	ETWFE
Homogeneous effect	0.054	0.102	0.101	0.058	0.102
Heterogeneity across time	0.377	0.099	0.098	0.076	0.099
Heterogeneity across cohorts	0.274	0.099	0.098	0.159	0.099
Heterogeneity across time and cohorts	0.349	0.102	0.100	0.146	0.102

Note: IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

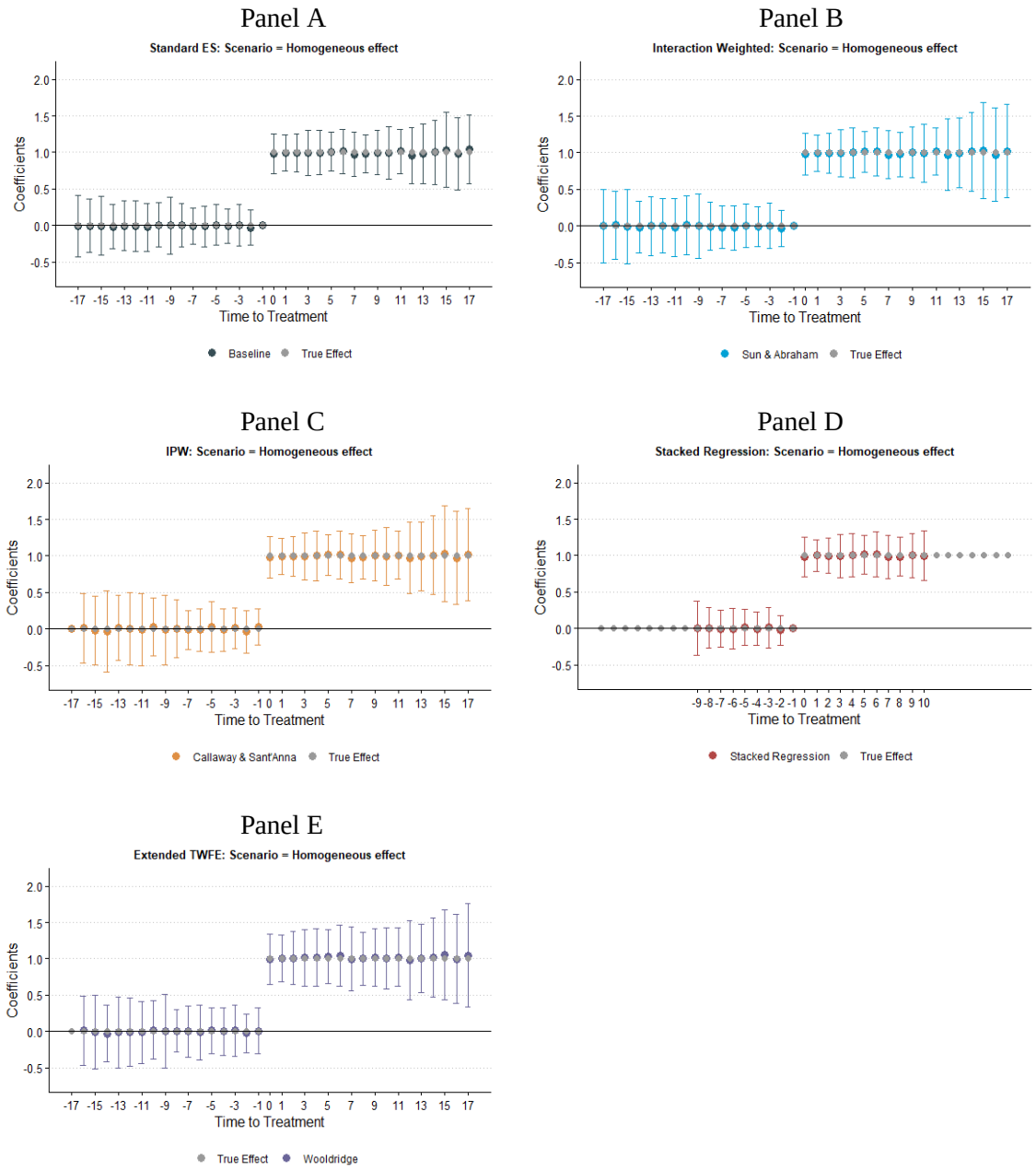

Figure 7: Event-study graphs for simulations for **Binary outcome** for **Scenario A (Homogeneous effect)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd). Currently there is no implementation of IE estimator extension for binary outcome.

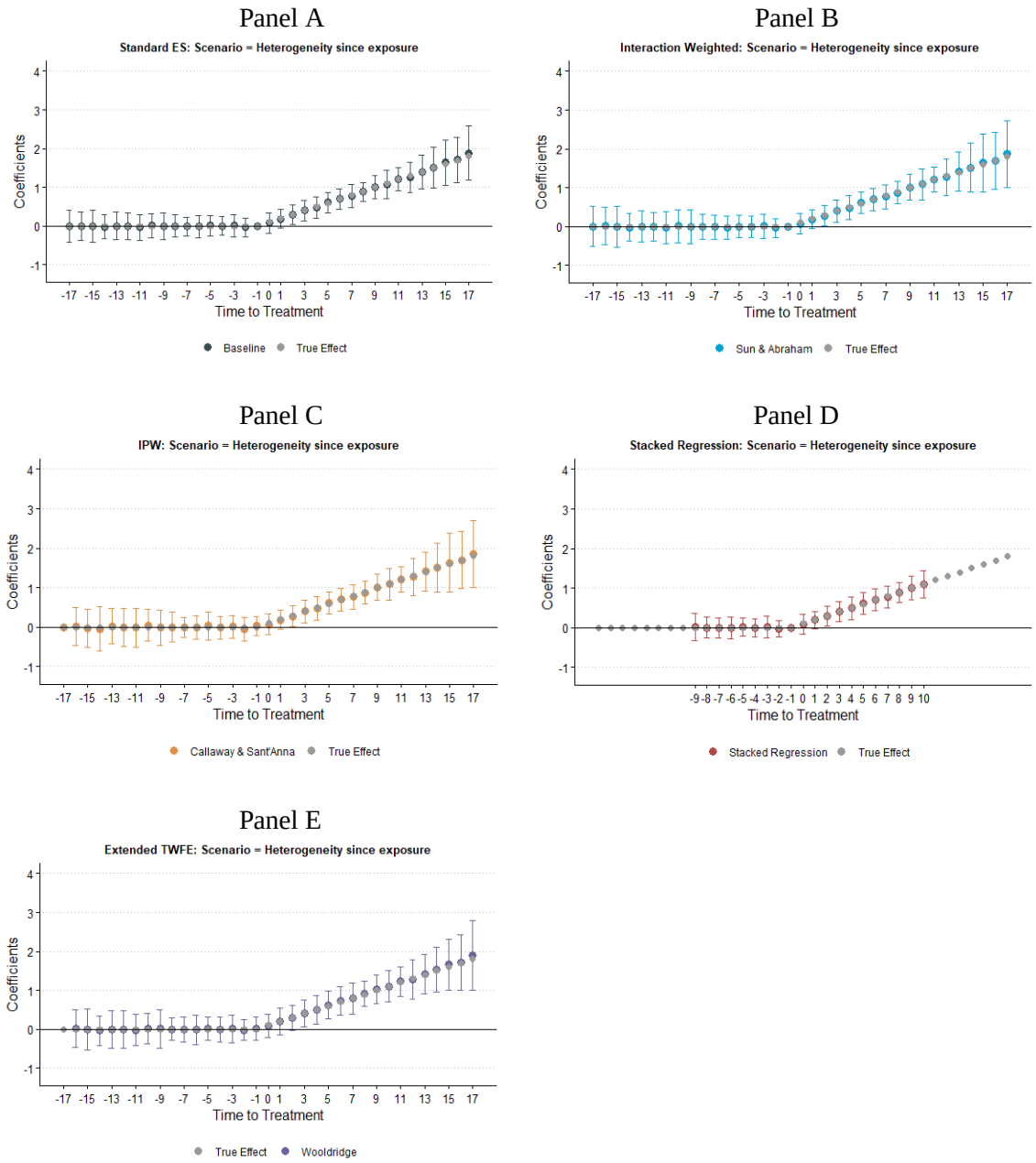

Figure 8: Event-study graphs for simulations for **Binary outcome** for **Scenario B (Heterogeneity since exposure)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd). Currently there is no implementation of IE estimator extension for binary outcome.

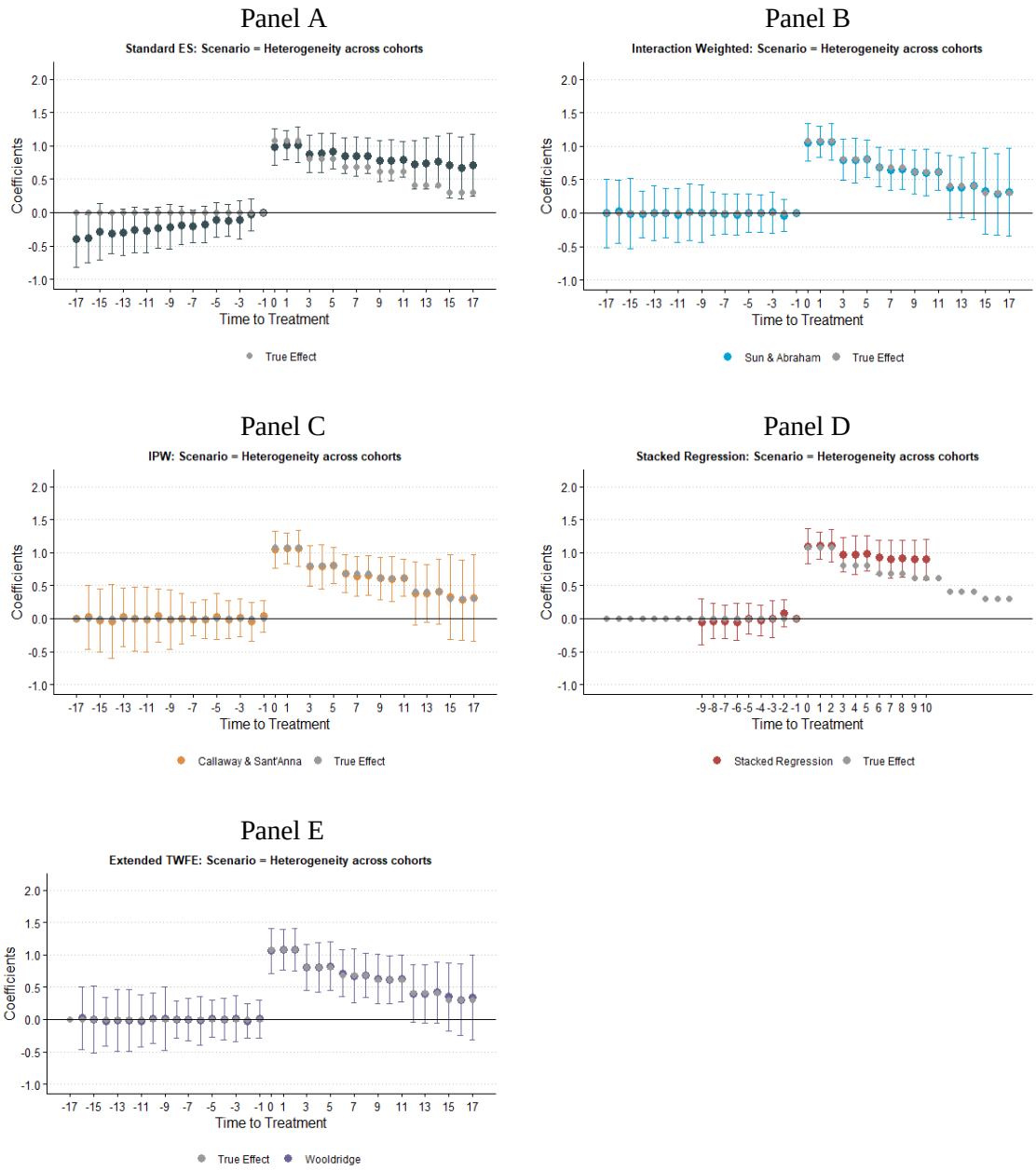

Figure 9: Event-study graphs from simulations for **Binary outcome** for **Scenario C (Heterogeneity across cohorts)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd). Currently there is no implementation of IE estimator extension for binary outcome.

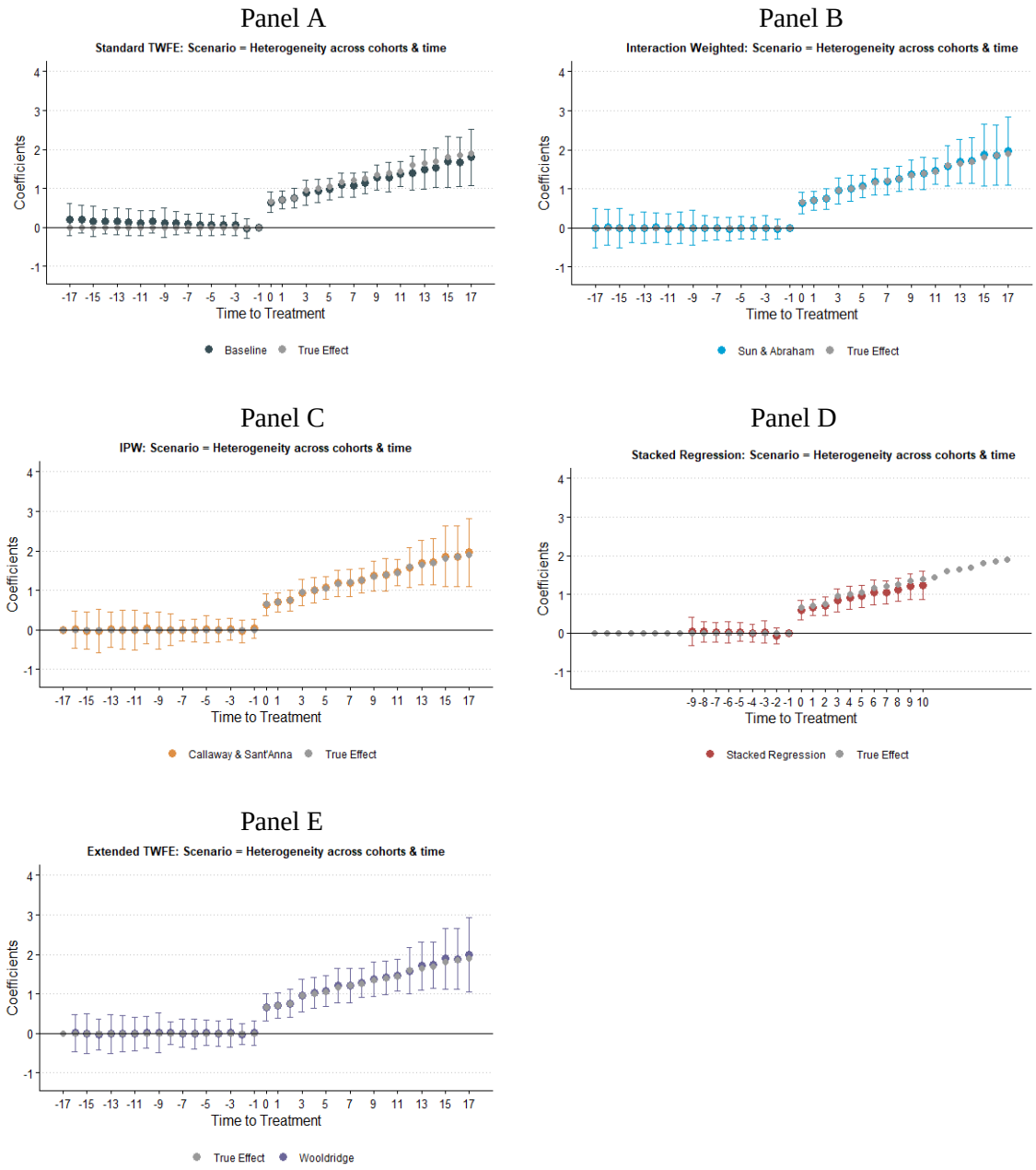

Figure 10: Event-study graphs for simulations for **Binary outcome** for **Scenario D (Heterogeneity across cohorts and time since exposure)**

Note. The point estimate (shown on y-axis) is the average of all point estimates from the simulations. The closer the point estimate is to the true effect evolution (true effect shown in grey), the lower the bias and vice-a-versa. The confidence intervals are generate using the standard deviations in points estimates from simulation. We target 95% confidence interval (CI), therefore, the confidence interval is generated by standard deviation (sd) in point estimate (e) across simulations (CI = e $\pm$ 1.96*sd). Currently there is no implementation of IE estimator extension for binary outcome.

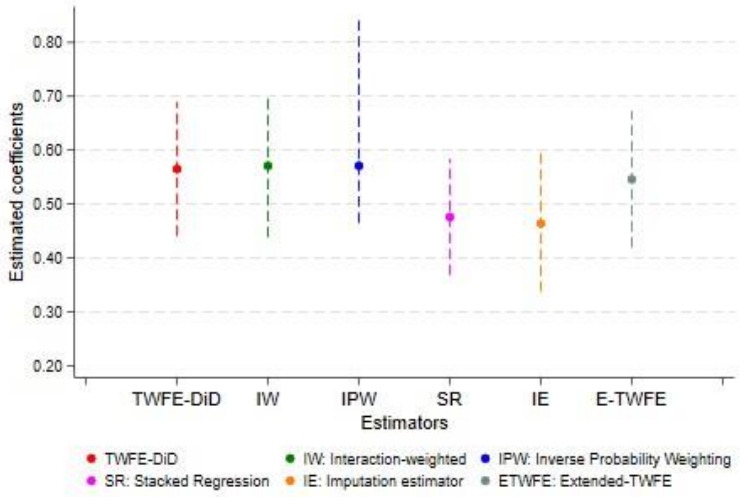

Figure 11: YMO: Standard TWFE-DiD and alternative estimators DiD estimates plot for Count Outcome

Note: The figure plots standard TWFE-DiD and alternative estimators coefficient estimates and 95% confidence interval. For standard TWFE-DiD, Interaction weighted, Stacked Regression, and Extended-TWFE, the confidence intervals are generated through regression, whereas for Inverse probability weighting and Imputation estimator the confidence interval are obtained through bootstrapping.

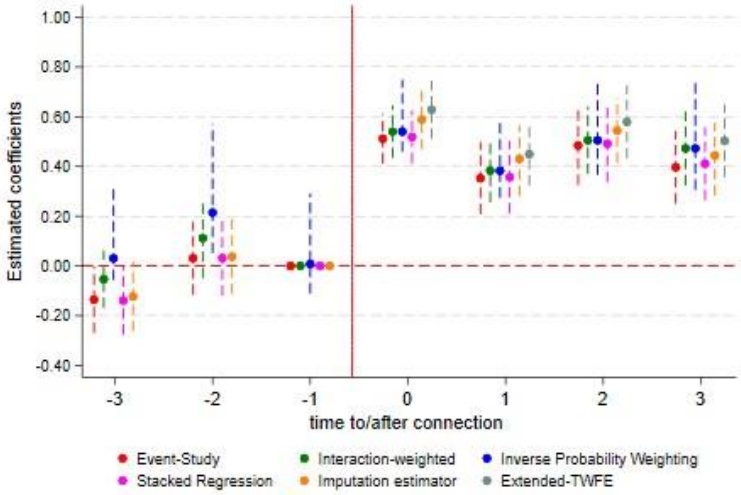

Figure 12: YMO: Standard TWFE-ES and alternative estimators event-study plot for Count Outcome

Note: The figure plots standard TWFE-ES and alternative estimators coefficient estimates and 95% confidence interval for relative-time periods 3 years before the treatment to 3 years after the treatment. Except IPW, all other estimators are implemented using Poisson QMLE to estimate the effect. For standard TWFE-ES, Interaction weighted, Stacked Regression, and Extended-TWFE, the confidence intervals are generated through regression, whereas for Inverse probability weighting and Imputation estimator the confidence interval are obtained through bootstrapping.

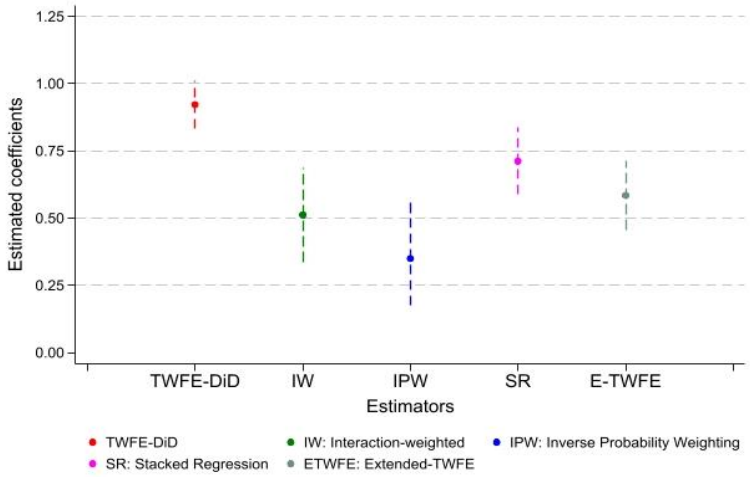

Figure 13: YMO: Standard TWFE-DiD and alternative estimators DiD estimates plot for Binary Outcome

Note: The figure plots standard TWFE-DiD and alternative estimators coefficient estimates and 95% confidence interval. For standard TWFE-DiD, Interaction weighted, Stacked Regression, and Extended-TWFE, the confidence intervals are generated through regression, whereas for Inverse probability weighting and Imputation estimator the confidence interval are obtained through bootstrapping.

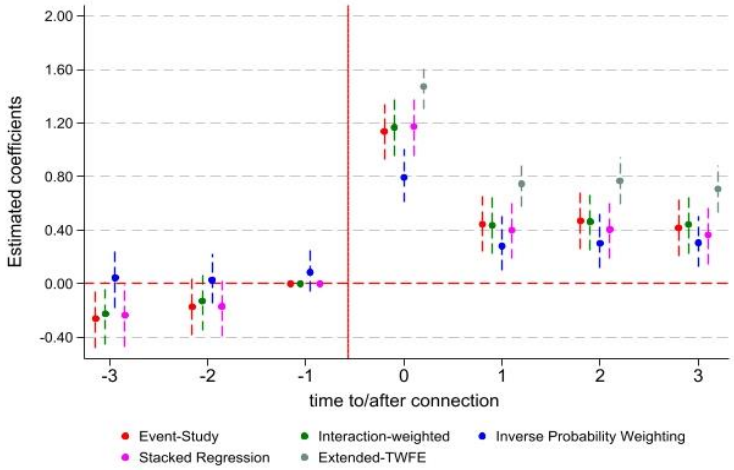

Figure 14: YMO: Standard TWFE-ES and alternative estimators event-study plot for Binary Outcome

Note: The figure plots standard TWFE-ES and alternative estimators coefficient estimates and 95% confidence interval for relative-time periods 3 years before the treatment to 3 years after the treatment. Except IPW, all other estimators are implemented using Poisson QMLE to estimate the effect. For standard TWFE-ES, Interaction weighted, Stacked Regression, and Extended-TWFE, the confidence intervals are generated through regression, whereas for Inverse probability weighting and Imputation estimator the confidence interval are obtained through bootstrapping.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, *72*(1), 1-19.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, *80*(1), 123-129.

Allison, P. D. (2009). *Fixed effects regression models*. SAGE publications.

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431-497.

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, *144*(2), 370-395.

Barkowski, S. (2021). Interpretation of nonlinear difference-in-differences: the role of the parallel trends assumption. *Available at SSRN 3772458*.

Barrios, J. M. (2021). Staggeringly problematic: A primer on staggered DiD for accounting researchers. *Available at SSRN 3794859*.

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

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

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

Chiu, A., Lan, X., Liu, Z., & Xu, Y. (2023). What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study. *arXiv preprint arXiv:2309.15983*.

Ciani, E., & Fisher, P. (2019). Dif-in-dif estimators of multiplicative treatment effects. *Journal of Econometric Methods*, *8*(1), 20160011.

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, *110*(9), 2964-2996.

De Chaisemartin, C., & d'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, *26*(3), C1-C30.

Deb, P., Norton, E. C., & Manning, W. G. (2017). *Health econometrics using Stata* (Vol. 3). College Station, TX: Stata press.

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

Fernández-Val, I., & Weidner, M. (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics*, *192*(1), 291-312.

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.

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

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica: Journal of the Econometric Society*, 701-720.

Lee, M. J., & Lee, S. (2021). Difference in differences and ratio in ratios for limited dependent variables. *arXiv preprint arXiv:2111.12948*.

Lindner, S., & McConnell, K. J. (2021). Heterogeneous treatment effects and bias in the analysis of the stepped wedge design. *Health Services and Outcomes Research Methodology*, 1-20.

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218-2244.

Schmidheiny, K., & Siegloch, S. (2019). On event study designs and distributed-lag models: Equivalence, generalization and practical implications.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, *225*(2), 175-199.

Taddeo, M. M., Amorim, L. D., & Aquino, R. (2022). Causal measures using generalized difference-in-difference approach with nonlinear models. *Statistics and Its Interface*, *15*(4), 399-413.

Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). *Stacked Difference-in-Differences* (No. w32054). National Bureau of Economic Research.

Wooldridge, J.M., 1999. Quasi-likelihood methods for count data. *Handbook of applied econometrics volume 2: Microeconomics*, pp.321-368.

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

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, *26*(3), C31-C66.

Yadav, A., McHale, J., & O'Neill, S. (2023). How does co-authoring with a star affect scientists' productivity? Evidence from small open economies. *Research Policy*, *52*(1), 104660.

Appendix

Appendix A: Parallel trend assumption for nonlinear DiD

To obtain PT assumption for nonlinear DID models, we assume a strictly increasing, continuously differentiable function $G(\cdot)$ and can state the assumption in two parts,

$$\mathrm{E}[Y_1(0)|D] = G(\alpha + \delta D)$$

$$\mathrm{E}[Y_2(0)|D] = G(\alpha + \delta D + \gamma)$$

where the key restriction is that $G(\cdot)$ is monotonically increasing. Combining the above two gives:

$$G^{-1}(\mathrm{E}[Y_2(0)|D]) - G^{-1}(\mathrm{E}[Y_1(0)|D]) = \gamma$$

where the PT assumption applies to a nonlinear transformation of the means. The linear PT assumption holds true for the indices contained with the function $G(\cdot)$ (This type of assumption is more in line with the one discussed in Puhani (2012)), for example, the linear PT assumption holds for the underlying latent variable in case of binary dependent variable but fails generally for $\mathrm{E}[Y_t(0)|D]$ (Wooldridge, 2023). Our choice of $G(\cdot)$ affects how we estimate the ATT, τ_2 (Barkowski (2021) provide a clear explanation of how the choice of PT assumption based on scale can impact the final inference of the results).

Using the conditional expectations for nonlinear model, the counterfactual outcome is defined as:

$$\mathrm{E}[Y_2(0)|D = 1] = G(\alpha + \delta + \gamma)$$

Therefore, the true ATT, $\tau_2 = \mathrm{E}[Y_2(1) - Y_2(0)|\, D = 1]$ is identified in regression form as

$$\tau_2 = G(\alpha + \delta + \gamma + \beta) - G(\alpha + \delta + \gamma)$$

Appendix B: Alternative estimators

We provide a discussion on the alternative estimation techniques that have been proposed to deal with estimation in staggered DiD/ES designs with HTEs:

B.1. Interaction-weighted estimator - Sun and Abraham (2021)

Sun and Abraham (2021) examined heterogeneous effects in the context of ES design and staggered treatment timing. To examine the dynamic effect in ES, a researcher includes leads and lags of treatment. The author showed in ES, with staggered treatment timing, the TWFE estimates of coefficients on lead and lag indicators will be contaminated by treatment effects from other relative periods in the presence of heterogeneous effects. The key focus of the interaction-weighted estimator is the "Cohort-specific average treatment effect on treated" l periods away from the initial treatment for the cohorts first treated at event-time e, $(CATT_{e,l})$. The theoretical finding in their paper is that TWFE coefficients on lead and lag are biased because coefficients of l periods relative to treatment can be seen as linear combination non-convex average of $CATT$ of that period and $CATT$ from other periods. They proposed a three-step estimation strategy that is robust to treatment effect heterogeneity and calculates weighted average of $CATT_{e,l}$.

Sun and Abraham (2021) proposed an estimation technique to estimate weighted average $CATT_{e,l}$ using an interacted specification that interacts with relative time indicators $D_{i,t}^{l}$ and cohort indicators $1.\{E_i = e\}$, which they labelled as "interaction-weighted" estimator. To estimate $CATT_{e,l}$ interact relative time dummies with group dummies as follows:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e} \sum_{l \neq -1} \delta_{e,l} \left(1.\{E_i = e\}.D_{i,t}^{l}\right) + \varepsilon_{i,t} \qquad (B1.1)$$

where $\delta_{e,l}$ is the DiD estimator for $CATT_{e,l}$. The units used as controls are those units that are never treated or the last treated group[21] (this group is then never used as a treated group). The weights are defined equal to each cohort's sample share in the relative period l. The interaction-weighted estimator is formed by taking weighted average over all $CATT$ from equation (B1.1) multiplied by the weights for relevant cohort and relative period. The interaction-weighted estimator is given by:

$$\hat{v}_g = \frac{1}{|g|} \sum_{l \in g} \sum_{e} \hat{\delta}_{e,l} \, \widehat{Pr}(E_i = e | E_i \in [-l, T-l]) \qquad (B1.2)$$

where $\hat{\delta}_{e,l}$ is estimated from equation (B1.1) and $\widehat{Pr}(E_i = e | E_i \in [-l, T-l])$ is the weights equal to share of each cohort e in relative period l. $\hat{v}_g$ is similar to $\theta_{es}(e)$ in interpretation, and is a parameter

21. Sun and Abraham (2021) argues that when the last treated group is used as control then the researcher needs to exclude more than one relative period (possibly the farthest lead) to avoid multi-collinearity issues.

of interest in many applied studies. Thus, we can create standard event-study plots across relative periods l to understand the dynamic effect.

B.2. Inverse Probability Weighting (IPW) Estimator - Callaway & Sant'Anna (2021)

Callaway & Sant'Anna (2021) proposed an outcome regression estimator, IPW estimator, and Doubly-Robust estimator. In this paper, we focus on the IPW estimator. IPW estimator is a way to deal with weighting problems of TWFE regression and estimate treatment effect parameters using DiD with multiple time periods, variation in treatment timing, and assuming parallel trends hold only after conditional on time-invariant covariates. The key concept in the estimator is the "group-time average treatment on the treated", $ATT(g,t)$. $ATT(g,t)$ is a unique ATT for each cohort of units treated at the same time point g, and time period t. The authors rely on a non-parametric approach to develop their estimator and propose a two-step estimation strategy with a bootstrap procedure to conduct valid inference for $ATT(g,t)$. IPW estimator (under the assumptions) identifies $ATT(g,t)$ non-parametrically as:

$$ATT(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right] \qquad (B2.1)$$

where $p_g(X)C$ is the propensity score conditional on covariates X and being a unit from treated group g or control group C. The $p_g(X)C$, normalized to one, determines the weights. G_g is the binary variable equal to one for units treated at time g, and C is the binary variable for control group units. $ATT(g,t)$ is simply the weighted average of the "long-difference" between the outcome of treated group g and control group C. The intuition of this approach is to take observations from control group C and treated group g, and drop observations from other groups. Then assign higher weights to observations in control group that have characteristics analogous to treated group g, and low weights to observations in control group that are different. The reweighting ensures that control and treatment groups are balanced conditional on covariates[22]. The author also provides a simple way to aggregate $ATT(g,t)$ into simpler parameters such as by relative time, by calendar time, etc.

We focus on aggregation by overall ATT and relative time, i.e., ES, as the paper focuses on DiD and ES design. The author provides a simple way to aggregate $ATT(g,t)$ across cohorts and time:

$$\theta_O = \frac{1}{K}\sum_{g\in G}\sum_{t=2}^{T} 1.\{t \geq g\} ATT(g,t)\, P(G=g|G\leq T) \qquad (B2.2)$$

22. The IPW estimator allows to choose control units between the two: "never-treated units" or "not-yet treated units".

and across different lengths of exposure to treatment as:

$$\theta_{es}(e) = \sum_{g \in G} 1.\{g + e \leq T\}\, P(G = g | G + e \leq T)\, ATT(g, g + e) \qquad (B2.3)$$

θ_O is the overall effect of participating in the treatment and $\theta_{es}(e)$ is the average effect of participating in the treatment for e time periods since the treatment was implemented across all groups that have ever been observed to have participated in the treatment for exactly e time periods, where e is event time i.e., time from the treatment. T is the last time in the data. $\theta_{es}(e)$ is the target parameter in many studies and can be used to generate standard ES plots across different e's to view dynamic treatment effects.

B.3. Stacked regression Estimator – Cengiz et al. (2019); Deshpande & Li (2019)

Stacked Regression estimate treatment effects with staggered treatment timing and heterogeneity in effects. Cengiz et al., (2019) employed stacked regression to examine the effect of minimum wages on low-wage jobs. The author use stacked regressions as a robustness check against the problems with aggregating DiD into a single parameter. Stacked regression can easily be extended to ES.

In stacked regression, a new event-specific dataset is created for each treated cohort g. The dataset contains the observations for that cohort g over a specific window from k_a to k_b, where k_a periods before and k_b periods after the treatment is administered, along with "clean controls" (units that did not receive treatment during the estimation window, k_a to k_b)[23]. Choosing the length of k_a and k_b is specific to research design. Then assign a specific event indicator w_g for each event time. We then stack each event-specific dataset and regress the outcome on treatment using the following DiD and ES specification:

$$Y_{itg} = \alpha_{ig} + \lambda_{tg} + \beta D_{itg} + \varepsilon_{itg} \qquad (B3.1)$$

$$Y_{itg} = \alpha_{ig} + \lambda_{tg} + \sum_{j=-k_a}^{-2} \beta_j\, (1.\{t - g_i = j\}) + \sum_{j=0}^{k_b} \beta_j\, (1.\{t - g_i = j\}) + \varepsilon_{itg} \qquad (B3.2)$$

Cengiz et al. (2019) noted that stacked regression is an attractive alternative approach to standard TWFE as it incorporates more strict criteria to create the control group. In stacked regression, we need to saturate dataset-specific unit and time fixed effects. This is the only difference between standard TWFE and stacked regression functional forms. However, stacked regression gives freedom to choose if we want multiple copies of same observation. We account for this fact by clustering standard errors at unit-group level to make sensible interpretations. The stacking creates a setting where all units are treated

23. k_a and k_b are the length of pre and post event window. For instance, k_a is the number of years before the treatment that is required for estimation, and k_b is the number of years after the treatment that is required for estimation.

simultaneously, thus prevents the past treated units acts as control. This is more robust to problems arising in the presence of heterogeneous effects in staggered treatment design. Similar to other estimators, we can create a lead/lag ES plot for relative time periods.

B.4. Extended-TWFE (ETWFE) estimator – Wooldridge (2021)

Wooldridge (2021) proposed an Extended-TWFE estimator to estimate cohort-period specific treatment effects, τ_{rs}, under staggered design in the presence of heterogeneous effects. He pointed out that the problem with TWFE is that the model is too restrictive. If we explicitly allow treatment effects to vary across cohorts and periods in the regression, the FE estimator captures the treatment effect heterogeneity across cohorts and periods. Wooldridge (2021) shows the equivalence between TWFE and Two-Way Mundlak (TWM) regression and argues that various estimators for policy intervention analysis can be computed using TWFE or pooled ordinary least square (POLS) regression. Wooldridge (2021) proposed TWM and TWFE regression that is robust to treatment effect heterogeneity, and estimates τ_{rs} – both regression approaches are equivalent.

The proposed estimation technique in Wooldridge (2021) uses an interacted specification with interaction between cohort and period dummies to estimate τ_{rs}, which he called "Extended-TWFE" (ETWFE) estimator (if a researcher uses fixed effect estimator) or "TWM" estimator (if a researcher uses POLS estimator). The effects using the TWFE estimator is estimated by[24]:

$$Y_{i,t} = \alpha_i + f_t + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}\ (w_{it} \cdot d_{ir} \cdot f_{st}) + \varepsilon_{it} \qquad (B4.1)$$

where τ_{rs} is the cohort-period specific effect. τ_{rs} captures the heterogeneity in effect for unit i in cohort r for period s. w_{it} is the treatment indicator variable that switches to one when the treatment is administered, d_{ir} and f_{st} are the cohort and time dummy, and α_i and f_t are unit and time fixed effect. Wooldridge (2021) encourages to use TWFE estimator over POLS due to its advantages for the unbalanced panel, as TWFE allows correlation between sample selection and unobserved heterogeneity. We can also include covariates in the regression as controls. Also, incorporating covariates interaction can capture the moderating effect of the covariate on sub-population. τ_{rs} can then be aggregated based on the researcher's main question, such as cohort effects, exposure effects, etc., to get the aggregate effect of the treatment. We aggregate the effect by exposure to treatment to obtain dynamic treatment effects (as this paper focuses on ES design).

24. The TWM estimator is:

$$Y_{i,t} = \eta + \sum_{r=q}^{T} \lambda_r d_{ir} + \sum_{s=2}^{T} \theta_s f_{st} + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}\ (w_{it} \cdot d_{ir} \cdot f_{st}) + \varepsilon_{it}$$

We explicitly include time(f_{st}) and cohort (d_{ir}) dummies and estimate τ_{rs} using the POLS regression.

B.5. Imputation estimator (IE) – Borusyak et al. (2021)

Borusyak et al. (2021) proposed the imputation estimator[25]. They suggest that good estimations of the true effects are possible using the three-step imputation estimator if we could impute potential outcomes for treated units using the data of not-yet treated (including control) units, and simultaneously avoiding the issues arising in standard TWFE model. The imputation estimator uses average outcome over all pre-treatment periods that could potentially improve the efficiency of the estimator, however, the estimator depends on strong PT assumption as it requires PT holding over all pre-treatment periods (Roth et al., 2023). The imputation estimator can be more susceptible to bias if there is a monotonic violation of pre-trends. (Roth, 2023; de Chaisemartin & d'Haultfoeuille, 2023)

To understand the approach, it is useful to define Z^0 as the set of observations for not-yet treated units (in particular, Z^0 includes all observations for the control group and pre-treatment observations for the units that eventually become treated) and define Z^1 as set that includes post-treatment observations for treated units. In the first-step, using the set of observations Z^0, fit the TWFE regression as follows to estimate the untreated potential outcomes for observations in Z^0:

$$Y_{i,t}(\mathrm{Z}^0) = \alpha_i + \lambda_t + \varepsilon_{it} \qquad (B5.1)$$

We can also add time-varying covariates in the above equation. Then in the second-step, use the predicted value from the above regression, $\hat{Y}_{it}(\mathrm{Z}^0)$, to impute the never-treated potential outcome for each unit in the whole data, and, get an estimate of treatment effects for each unit,

$$\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it} \qquad (B5.2)$$

where $\hat{Y}_{it}$ is the untreated potential outcome for each unit, and $\hat{\tau}_{it}$ is the individual treatment effect which can be used to form aggregate measures. In the third-step, estimate the weights for each unit under treatment and estimate the weighted average of these individual level effects based on target estimand.

25. See Gardner (2022), Liu et al. (2021) for similar approaches. In this paper, we focus on the estimator proposed in Borusyak et al. (2021).

Appendix C

C1. Cohort-specific differences

Cohort-specific differences:

In the DGP, α_i the unit fixed effects are drawn from $\sim N(0,1) + \mu_{cohort} + e_{it}$ where μ_{cohort} is cohort-specific differences, in other words, imbalance between the cohorts. In the simulation, we have 7 cohorts, $g \in \{3, 6, 9, 12, 15, 18, \infty\}$ where ∞ for untreated units. We add cohort-specific differences in the DGP to make data more realistic. For example, when a tax-policy is introduced within a country, different states adopt the policy at different times indicating some unobserved differences within those states. Table C1.1 shows the value the data is generated for each cohort-specific difference.

Table C1.1: mean for cohort-specific difference

Cohort	mean
3	-0.1
6	-0.2
9	-0.3
12	-0.4
15	-0.5
18	-0.6
∞	0

C2. True value of treatment effects

The average treatment effect on treated for each and period, $ATT_{g,t}$, for each scenario of treatment effect evolution are given by:

Scenario A (Homogeneous effect):

$$ATT_{g,t} = 1 \times 1\,[t \geq g]$$

Scenario B (Heterogeneity across time):

$$ATT_{g,t} = 0.1(t - g + 1) \times 1[t \geq g]$$

Scenario C (Heterogeneity across cohorts):

$$ATT_{g,t} = 0.1(g) \times 1\,[t \geq g]$$

Scenario D (Heterogeneity across time and cohorts):

$$ATT_{g,t} = 0.05(t - g + 1) \times 1[t \geq g] + 1.15 - 0.05(g) \times 1\,[t \geq g]$$

C3. Stylized example for each scenario of treatment effects evolution

We present a stylized example of for each scenario treatment effect evolution for true known ATT for cohort g at time t, $ATT_{g,t}$. The purpose of this example is to clarify the different types of heterogeneity in effects used in simulations. For clarity we focus on three treated cohorts, {X,Y,Z}, over five periods. In this example, cohort X receives treatment at time 2; cohort Y receives treatment at time 3; and cohort Z received treatment at time 4. The following is the $ATT_{g,t}$ for each scenario in this example:

Scenario A: Homogeneous treatment effect. For this scenario, the true effects are constant. We assume the constant is 0.5 (Note. This value 0.5 is used as an example. It do not represent the true effect values used in simulation). Table C3.1 shows the treatment effect evolution for scenario A. The bold values represent the periods under treatment for each cohort.

Table C3.1: $ATT_{g,t}$ evolution scenario A

	Cohorts		
Time	X	Y	Z
1	0	0	0
2	**0.5**	0	0
3	**0.5**	**0.5**	0
4	**0.5**	**0.5**	**0.5**
5	**0.5**	**0.5**	**0.5**

Scenario B: Heterogeneity across time since exposure. In this scenario the treatment effect varies since exposure but for each cohort. Table C3.2 shows the evolution of $ATT_{g,t}$ for scenario B. We assume the true effect is increases with the length of exposure since treatment.

Table C3.2: $ATT_{g,t}$ evolution scenario B

	Cohorts		
Time	X	Y	Z
1	0	0	0
2	**0.1**	0	0
3	**0.2**	**0.1**	0
4	**0.3**	**0.2**	**0.1**
5	**0.4**	**0.3**	**0.2**

Scenario C: Heterogeneity across intervention cohorts. For this scenario, the effect varies across treated cohorts but is constant with time since first exposure. Table C3.3 shows the $ATT_{g,t}$ evolution for scenario C. In this example, we assume the earlier treated cohort experience lower treatment effect compared to later treated cohort.

Table C3.3: $ATT_{g,t}$ evolution scenario C

	Cohorts		
Time	X	Y	Z
1	0	0	0
2	**0.4**	0	0
3	**0.4**	**0.7**	0
4	**0.4**	**0.7**	**0.9**
5	**0.4**	**0.7**	**0.9**

Scenario D. Heterogeneity across time since exposure and interventions cohorts. In this scenario, the true treatment effect varies across both intervention cohorts and time since exposure. Table C3.4 shows the $ATT_{g,t}$ evolution both across time and cohorts. In this example, $ATT_{g,t}$ values for scenario D are the combination for scenario B and scenario C. For instance, at time 2 for cohort X, true effect is summation of exposure effect (from scenario B) and cohort-specific effect (from scenario C) i.e. 0.1 + 0.4 = 0.5. At time 3 for cohort X, true effect is 0.2 (exposure effect at time 3 from scenario B) + 0.4 (cohort-specific effect at time 3 from scenario C) = 0.6. This way $ATT_{g,t}$ is generate for each cohort g at time t.

Table C3.4: $ATT_{g,t}$ evolution scenario D

	Cohorts		
Time	X	Y	Z
1	0	0	0
2	**0.5**	0	0
3	**0.6**	**0.8**	0
4	**0.7**	**0.9**	**1.0**
5	**0.8**	**1.0**	**1.1**

Appendix D

D1. Linear DiD for Count and Binary outcomes

For illustration purposes, we show the bias in the TWFE-DiD and alternative estimators if OLS (in other words, linear DiD model)[26] is used for Count and Binary outcomes. Figures D1.1 and D1.2 compare the nonlinear and linear DiD model for Count and Binary outcomes for homogeneous effect (Scenario A). The graphs show that all estimators are very biased if the linear DiD model, compared to nonlinear DiD model, is employed for Count and Binary outcome. Therefore, we need to employ Poisson QMLE for count outcomes and the Conditional logit Fixed Effect (CLE) for binary outcomes in place of the OLS estimator used for linear models by these approaches.

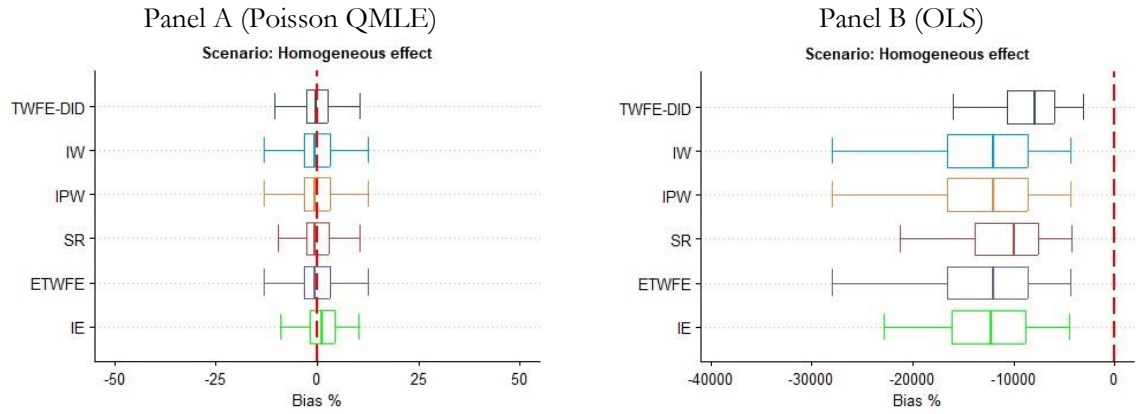

Figure D1.1 Boxplots of % bias in treatment effects for **Count outcome** using Poisson QMLE and OLS, for homogeneous effect (Scenario A)

Note: IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

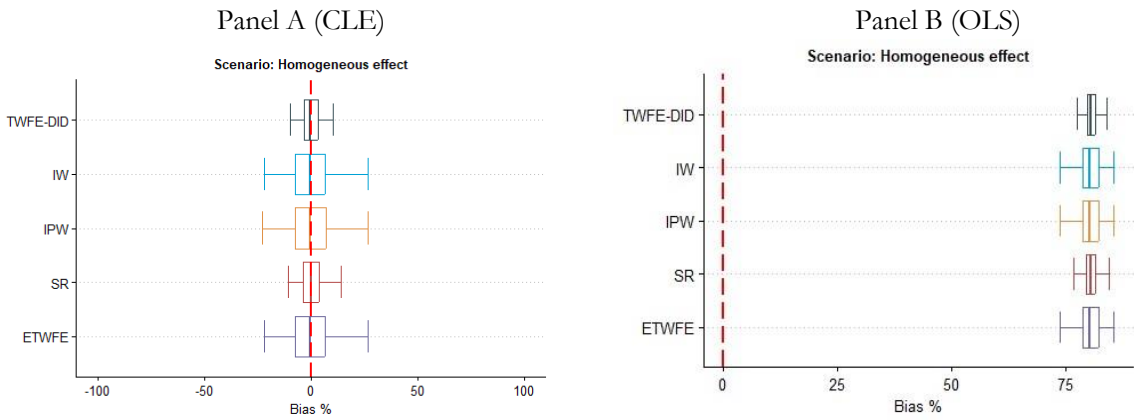

Figure D1.2 Boxplots of % bias in treatment effects for **Binary outcome** using CLE and OLS, for homogeneous effect (Scenario A)

Note. IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

26. The IPW estimator does not employ OLS but linear version of IPW estimator for linear model. Thus, here we use the linear version of the IPW estimator for the linear DiD model.

D2. Extension of IPW estimator to account for outcome distribution for count and binary outcomes

Equation B2.1 is the IPW estimator for staggered intervention (can be used for single intervention) proposed by Callaway & Sant'Anna (2021). The estimator estimates $ATT(g,t)$, ATTs for each group in each time period, using separate 2×2 comparisons. For each 2×2 model, equation B2.1 takes the following form (Sant'Anna & Zhao, 2020):

$$ATT_t = \tau_t = \frac{1}{\mathrm{E}[D]}\mathrm{E}\left[\frac{D - p(X)}{1 - p(X)}\,(Y_{t+k} - Y_t)\right] \qquad (D2.1)$$

where Y_{t+k} is the post-treatment outcome and Y_t is the pre-treatment outcome, D is the treatment indicator, and $p(X)$ indicates the probability of being treated conditional on pre-treatment covariates X.

For the 2×2 model, another way to approach the above equation:

$$\tau = \theta_1 - \theta_0$$

where $\theta_1 = \mathrm{E}\left[Y_{i,t+k}(1) \,\middle|\, D_i = 1\right]$, the expected outcome for the treated group is consistently estimated by (Li and Li, 2019):

$$\widehat{\theta_1} = \sum_{i=1}^{N} D_i Y_{i,t+k} \Big/ \sum_{i=1}^{N} D_i \qquad (D2.2)$$

and $\theta_0 = \mathrm{E}\left[Y_{i,t+k}(0) \,\middle|\, D_i = 1\right]$ is their expected counterfactual outcome. We use an IPW estimator to estimate $\widehat{\theta_0}$:

$$\widehat{\theta_0} = \frac{\sum_{i=1}^{N} D_i Y_{i,t} w_i}{\sum_{i=1}^{N} D_i} + \frac{\sum_{i=1}^{N} (1 - D_i)(Y_{i,t+k} - Y_{i,t})w_i}{\sum_{i=1}^{N} D_i} \qquad (D2.3)$$

Where w_i equals 1 for the treated group and $w_i = \hat{p}(X)/(1 - \hat{p}(X))$ for the control group.

Thus, the ATT for the 2×2 model is estimated using the following for continuous outcome, which is the mean difference:

$$\widehat{ATT} = \tau = \widehat{\theta_1} - \widehat{\theta_0} \qquad (D2.4)$$

Equation (D2.1) and (D2.4) are equivalent for continuous outcome for the 2×2 case.

Count outcomes:

For count outcomes, the effect for the 2×2 model is interpreted as causal rate ratio, where the estimand is given by $\frac{\mathrm{E}\left[Y_{i,t+k}(1) \,|D_i=1\right]}{\mathrm{E}\left[Y_{i,t+k}(0) \,|D_i=1\right]} = \frac{\widehat{\theta_1}}{\widehat{\theta_0}}$ (Li and Li, 2019). In our simulations, we assume the

PT assumption holds in the latent scale; therefore, we take the log of the rate ratio. Taking logs produces coefficient estimates comparable to those obtained from alternative (regression-based) estimators. On the log scale, the ATT can be identified using the following specification for the 2 × 2 case:

$$\widehat{ATT} = \tau = ln\left(\frac{\mathrm{E}\left[Y_{i,t+k}(1)\,\middle|D_i = 1\right]}{\mathrm{E}\left[Y_{i,t+k}(0)\,\middle|D_i = 1\right]}\right) = ln\left(\frac{\widehat{\theta_1}}{\widehat{\theta_0}}\right) \qquad (D2.5)$$

where $ln\left(\widehat{\theta_0}\right) = ln\left(\frac{\sum_{i=1}^{N} D_i Y_{i,t} w_i}{\sum_{i=1}^{N} D_i}\right) + ln\left(\frac{\sum_{i=1}^{N}(1-D_i)\left(Y_{i,t+k}\right)w_i}{\sum_{i=1}^{N} D_i} \Big/ \frac{\sum_{i=1}^{N}(1-D_i)(Y_{i,t})w_i}{\sum_{i=1}^{N} D_i}\right)$

Binary outcomes:

For binary outcome, the coefficient estimates are interpreted as causal odds-ratio. For the 2 × 2 case, the estimand is given by $\left\{\frac{\mathrm{E}[Y_{i,t+k}(1)|D_i=1]}{1-\mathrm{E}[Y_{i,t+k}(1)|D_i=1]}\right\} \Big/ \left\{\frac{\mathrm{E}[Y_{i,t+k}(0)|D_i=1]}{1-\mathrm{E}[Y_{i,t+k}(0)|D_i=1]}\right\}$. Same as the count outcome, we take the log of the odds ratio as our simulations assume the PT assumption holds in the latent scale. This provides estimates on a similar scale as estimates for alternative estimators obtained using logistic regression. The ATT can be identified using the following specification for the 2 × 2 case:

$$\widehat{ATT} = \tau = ln\left(\frac{\left\{\frac{\mathrm{E}\left[Y_{i,t+k}(1)\,\middle|D_i = 1\right]}{1 - \mathrm{E}\left[Y_{i,t+k}(1)\,\middle|D_i = 1\right]}\right\}}{\left\{\frac{\mathrm{E}\left[Y_{i,t+k}(0)\,\middle|D_i = 1\right]}{1 - \mathrm{E}\left[Y_{i,t+k}(0)\,\middle|D_i = 1\right]}\right\}}\right) = \ln\left(\frac{\left\{\frac{\widehat{\theta_1}}{1-\widehat{\theta_1}}\right\}}{\left\{\frac{\widehat{\theta_0}}{1-\widehat{\theta_0}}\right\}}\right) \qquad (D2.6)$$

where $ln\left(\frac{\widehat{\theta_0}}{1-\widehat{\theta_0}}\right) = ln\left(\frac{\left(\frac{\sum_{i=1}^{N} D_i Y_{i,t} w_i}{\sum_{i=1}^{N} D_i}\right)}{1-\left(\frac{\sum_{i=1}^{N} D_i Y_{i,t} w_i}{\sum_{i=1}^{N} D_i}\right)}\right) + ln\left(\frac{\frac{\sum_{i=1}^{N}(1-D_i)(Y_{i,t+k})w_i}{\sum_{i=1}^{N} D_i}}{1-\frac{\sum_{i=1}^{N}(1-D_i)(Y_{i,t+k})w_i}{\sum_{i=1}^{N} D_i}} \Big/ \frac{\frac{\sum_{i=1}^{N}(1-D_i)(Y_{i,t})w_i}{\sum_{i=1}^{N} D_i}}{1-\frac{\sum_{i=1}^{N}(1-D_i)(Y_{i,t})w_i}{\sum_{i=1}^{N} D_i}}\right)$

Equations D2.5 and D2.6 provide ATT estimates for each group in each time period, $ATT(g,t)$, for count and binary outcomes. Then, in the next step, we follow the aggregation scheme suggested by Callaway & Sant'Anna (2021) to obtain overall and dynamic treatment effects. We aggregate $ATT(g,t)$ to produce overall and dynamic effects of the treatment using equations B2.2 and B2.3. Therefore, our extension of the IPW estimator accounts for the outcome distribution in the first step of the estimation procedure, and in the second step, we follow the same procedure as in Callaway & Sant'Anna (2021) to aggregate $ATT(g,t)$ into parameters of interest.

Appendix E

E.1 Count outcome estimated effects boxplots for DiD from simulations

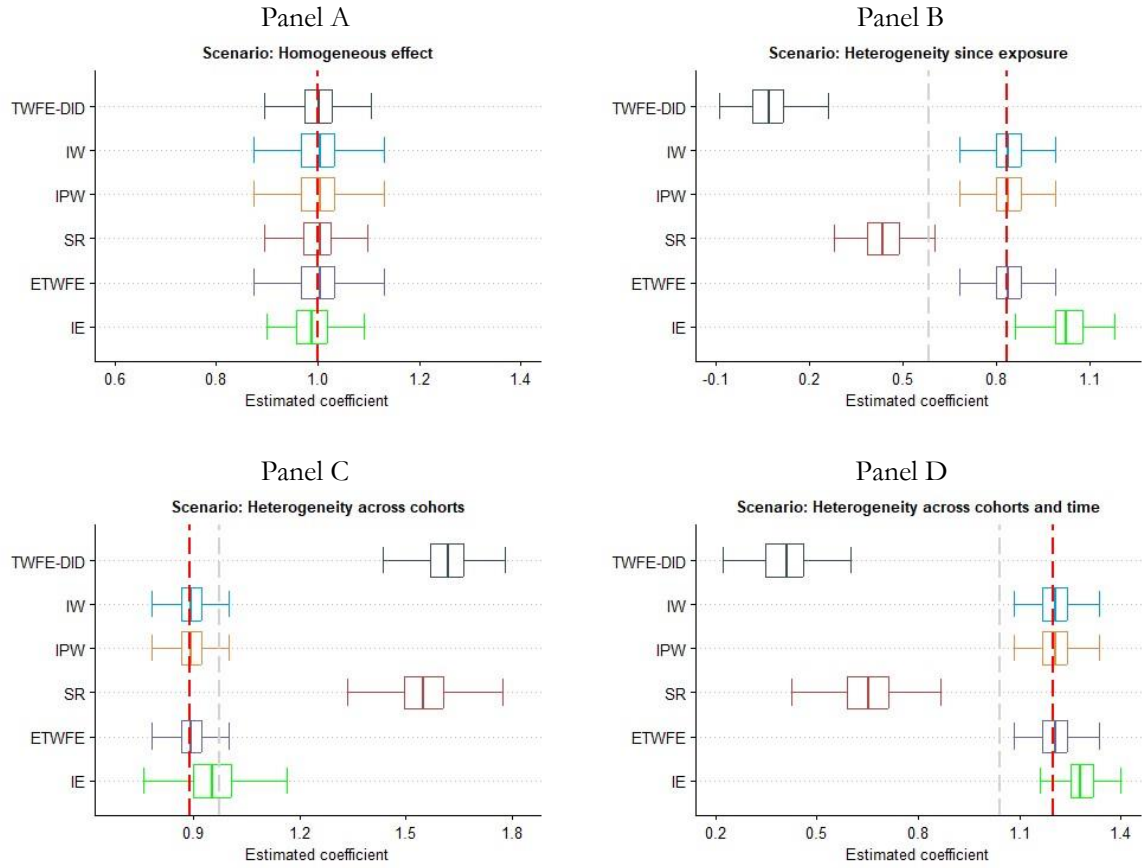

Figure E.1: Boxplots of estimated effects from simulations for **Count outcome**

Note: The red vertical line represents the true effect for the particular scenario, whereas the grey vertical line represents the true effect for stacked regression. IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

E.2 Count outcome mean bias for event-study from simulations

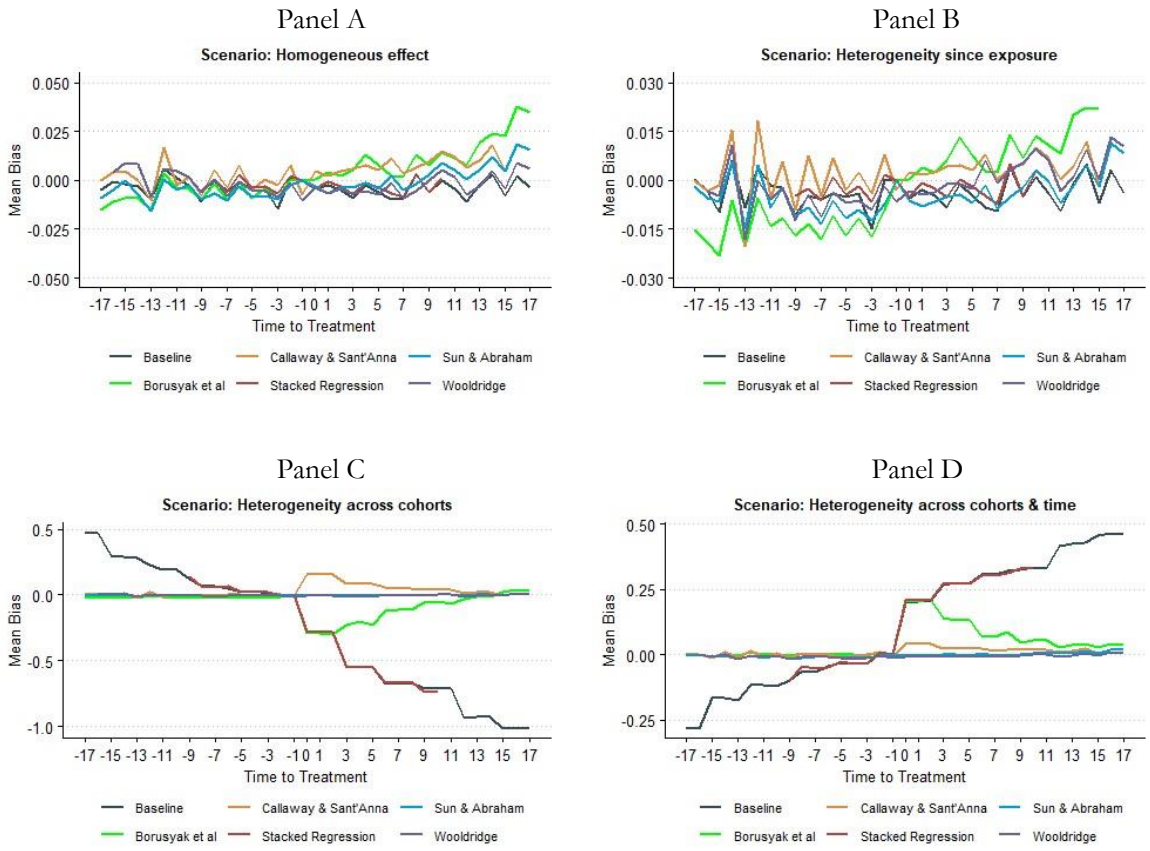

Figure E.2: Mean bias in the event-study model for **Count outcome**

Note: The line graph shows the mean bias across 500 simulations for each estimator and scenario. Closer to 0 represents low mean bias and vice-a-versa. Baseline: TWFE-ES estimator; Callaway & Sant'Anna: Inverse Probability Weighting, Sun & Abraham: Interaction-weighted; Wooldridge: Extended-TWFE; Borusyak et al: Imputation estimator.

Appendix F

F.1 Binary outcome estimated effects boxplots for DiD from simulations

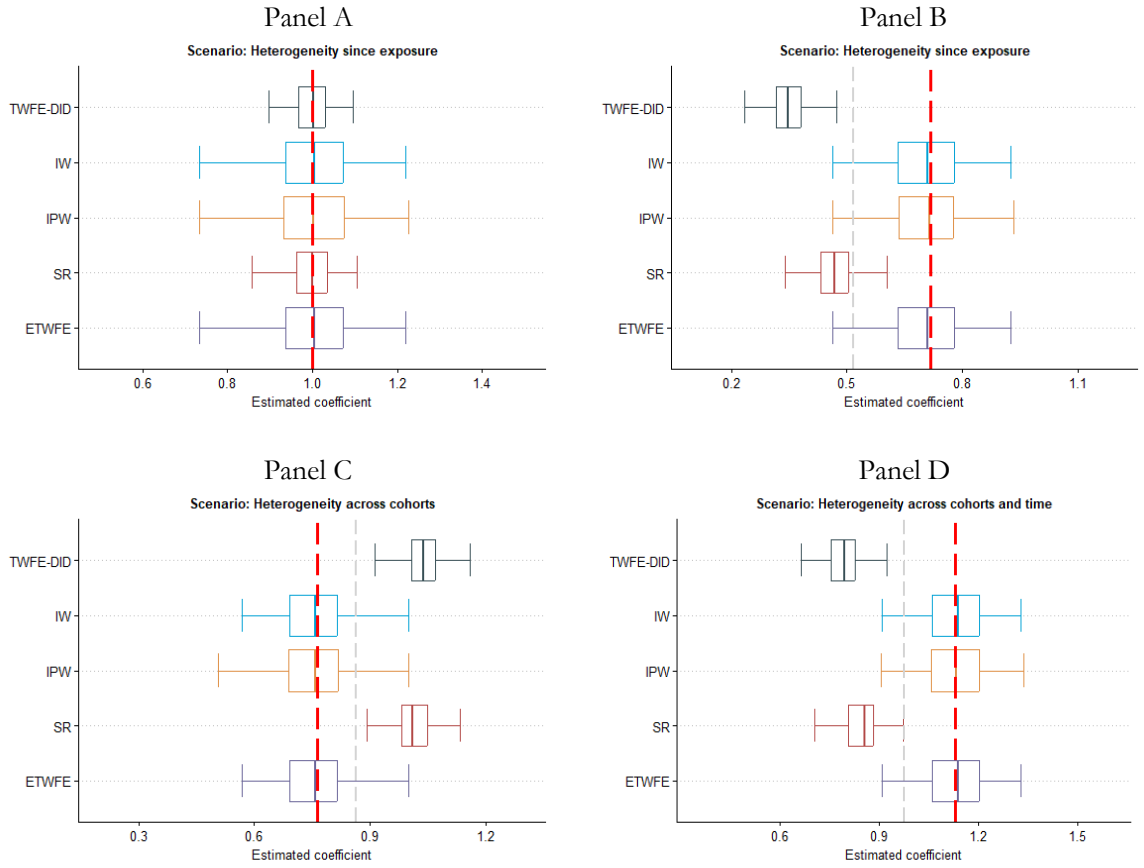

Figure F.2: Boxplots of estimated effects from simulations for **Binary outcome**

Note: The red vertical line represents the true effect for the particular scenario, whereas the grey vertical line represents the true effect for stacked regression. IW: Interaction-weighted, IPW: Inverse Probability Weighting, SR: Stacked Regression, ETWFE: Extended-TWFE, IE: Imputation estimator

F.2 Binary outcome mean bias for event-study from simulations

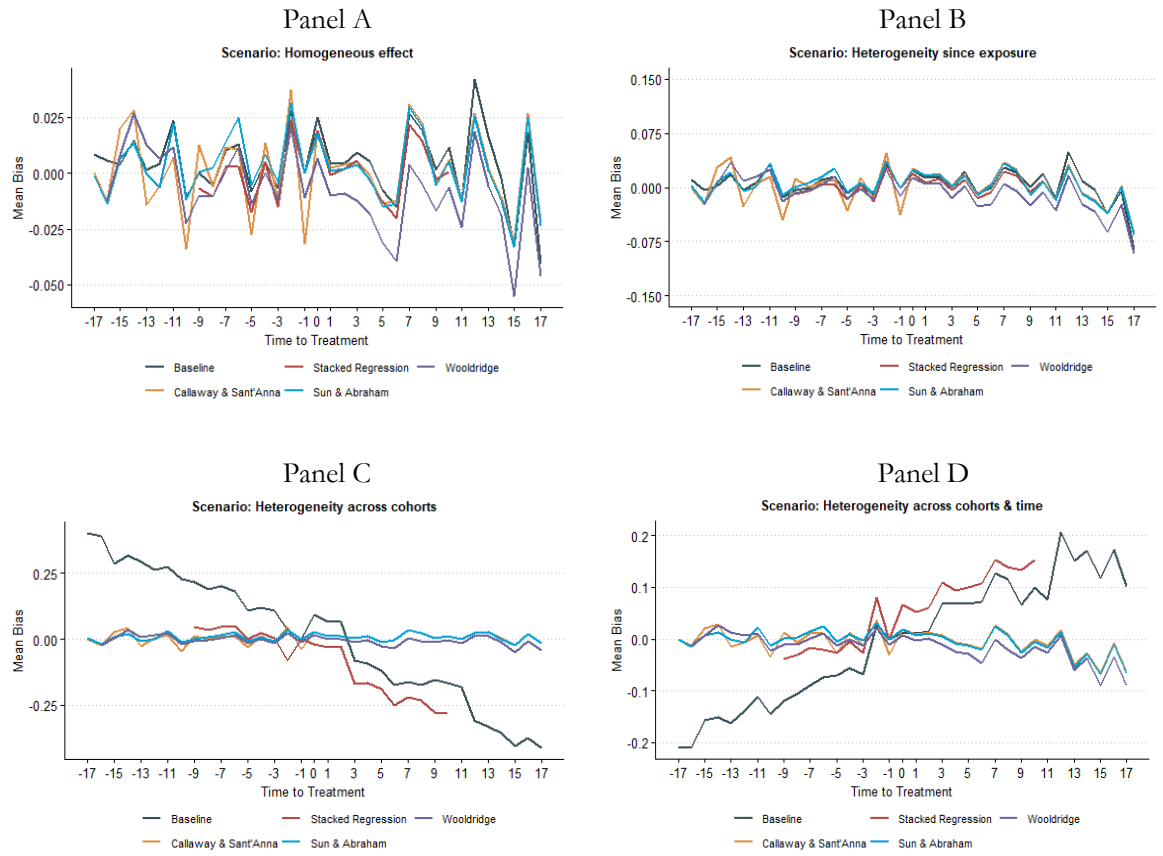

Figure F.2: Mean bias in the event-study model for **Binary outcome**

Note: The line graph shows the mean bias across 500 simulations for each estimator and scenario. Closer to 0 represents low mean bias and vice-a-versa. Baseline: TWFE-ES estimator; Callaway & Sant'Anna: Inverse Probability Weighting, Sun & Abraham: Interaction-weighted; Wooldridge: Extended-TWFE; Borusyak et al: Imputation estimator.

Appendix G: Comparison of the estimators

Table B1: Comparison of the estimators

	Clean controls issue/ comparison group	Negative weighting issue	Parallel Trend (PT) assumption reference year	Data requirements	Finest estimand
standard TWFE estimator	Create counterfactuals using never treated, not-yet treated, and already-treated units. Using already treated creates "bad comparisons" that may lead to biased estimates.	Under heterogeneous treatment effects, "bad comparisons" may put negative weight on the parameter of interest. Also, the estimates suffer from "cross-lag contamination" discussed in Sun & Abraham (2021)	Imposes PT right before the treatment starts until the last time period	Does not require pre-treatment periods. Estimates will be provided even when pre-treatment periods are not available for some units	ATT = produces single treatment effect
interaction weighted estimator by Sun and Abraham (2021)	Explicitly creates "good comparisons" using never treated and/or not-yet treated units	The estimation uses fully interacted regressions to recover an estimate of group specific ATT that avoids the limitation of TWFE regression	Imposes PT right before the treatment starts until the last time period	Strict requirement of at least one pre-treatment period for treated units; otherwise unit is dropped from estimation. And, a set of never treated units (if available, otherwise last-treated cohort) is used as control cohort	$ATT(g,t)$ = Treatment effect for each cohort g for each period t
Inverse-probability weighting by Callaway & Sant'Anna (2021)	Explicitly creates "good comparisons" using never treated and/or not-yet treated units	The estimator uses separate 2x2 comparisons to estimate effects for each group for each period using the last pre-intervention period for comparison to avoid the limitation of TWFE regression	Imposes PT right before the treatment starts until the last time period; however option to choose arbitrary pre-treatment period	Strict requirement of at least one pre-treatment period for treated units; otherwise unit is dropped from estimation. And, a set of never treated units, if available, otherwise not-yet treated units are used as control	$ATT(g,t)$ = Treatment effect for each cohort g for each period t
stacked regression	Explicitly creates "good comparisons" using never treated and/or not-yet treated units	Under heterogeneous treatment effects, the effects are biased as weights are assigned by the regression estimator are not proportional to cohort share under treatment	Imposes PT right before the treatment starts until the last time period	Requires a common window for pre-treatment and post-treatment periods for all treated units.	ATT = produces single treatment effect
imputation estimator by Borusyak et al (2021)	Explicitly creates "good comparisons (imputed counterfactuals)" using not-yet treated observations	The estimator used not-treated observations to create counterfactuals and explicitly specifies individual weights that avoid the limitation of TWFE regression	Imposes PT in all pre-treatment periods until the last time period	Strict requirement of at least one pre-treatment period for treated units; otherwise unit is dropped from estimation. And, a set of never treated units (if available, otherwise last-treated cohort) used as control cohort	τ_{it} = Treatment effect for each unit i for each period t
Extended TWFE by Wooldridge (2023)	Explicitly creates "good comparisons" using never treated and/or not-yet treated units	The estimation uses post-treatment fully interacted regressions to recover an estimate of group specific ATT that avoids the limitation of TWFE regression	Imposes PT in all pre-treatment periods until the last time period	Strict requirement of at least one pre-treatment period for treated units; otherwise unit is dropped from estimation. And, a set of never treated units (if available, otherwise last-treated cohort) is used as control cohort	$ATT(g,t)$ = Treatment effect for each cohort g for each period t

Note: In this paper, we use only "never-treated" units as controls to create clean controls. A researcher should be careful when using "never treated" units versus "not-yet treated" units as controls, as choosing either has implications for the evolution of the parallel trend assumption. If researchers assume parallel trend holds based on both never treated and not-yet treated units as controls and use only never treated units as controls in estimation, the estimation will not estimate the causal effect due to violation of parallel trend assumption.

T: +353 (0)1 224 6000
www.centralbank.ie
publications@centralbank.ie

Bosca PO 559, Baile Átha Cliath 1, Éire
PO Box 559, Dublin 1, Ireland