Assessing Treatment Effect From Quasi-Experiment Design In Operations Management

ABSTRACT

Two-way-fixed-effects (TWFE) regression using Difference-in-Difference (DiD) is a workhorse to assess treatment effect from quasi-experiment designs in operations management field. However, latest advancement in econometrics field proves that TWFE is prone to yield biased estimates when the treatment timing is different in a multi-group multi-period setting – also known as a staggered DiD design. Given 47% of DiD studies we surveyed in the top three operations management journals adopt staggered DiD designs and are susceptible to biased results, we theoretically review the causes of the biases associated with TWFE in a staggered DiD design, systematically compare several new alternative estimators that can help overcome these biases, practically provide a framework to select an appropriate estimator, tactically highlight they key areas during implementation, and empirically illustrate the pitfalls associated with TWFE in a staggered DiD design through simulation and replication. We also provide replication code for easy implementation to increase research rigor for the whole research community in operations management.

KEY WORDS: difference-in-difference, event study, treatment effect, quasi-experiment

ASSESSING TREATMENT EFFECT FROM QUASI-EXPERIMENT DESIGN IN OPERATIONS MANAGEMENT

1. INTRODUCTION

Assessing the impact of an exogenous shock (i.e., policy intervention, rare event, and etc.) has drawn tremendous amount of attention from researchers in operations management (OM) field, such as buyer supplier relationship (Fan et al. 2022), Covid-19 impact (Cui et al. 2022, Han et al. 2022, Ge et al. 2023, Wang 2022), environmental issues (Lo et al. 2022, Scott et al. 2023, Song et al. 2023), healthcare management (Chun et al. 2022, KC and Kim 2022, Li et al. 2022, Qiu et al. 2022), hotel operations (Chen et al. 2023), international business operations (Chan et al. 2023, Jacobs et al. 2022, Lam et al. 2022, Klöckner et al. 2022), retail operations (Akturk and Ketzenberg 2022, Calvo et al. 2023, Ergin et al. 2022, Ren et al. 2023, Wang et al. 2023, Kokkodis et al. 2022, Lee et al. 2022), and ridesharing (Agarwal et al. 2023, Barrios et al. 2022, Gong et al. 2023, Li et al. 2022, Miao et al. 2022, Pan and Qiu 2022, Zhou and Wan 2022). In the extant econometrics literature, the exogenous shock is known as the "treatment" while the impact of the exogenous shock is commonly referred to as the "treatment effect" (Wooldridge 2010).

As empirical OM researchers normally investigate the impact of an exogenous shock post hoc, a quasi-experiment rather than a truly randomized experiment is utilized in their research design. Consequently, two way fixed effects (TWFE) difference-in-difference (DiD) is the most frequently used estimator to draw causal inference for the treatment effect. However, with the recent advancement in econometrics, TWFE DiD has been proven to produce biased estimates when there is variation in treatment timing (i.e., different units received treatment at different times) in a multi-group multi-period setting – also referred to a staggered DiD design. In reviewing a total of 51 DiD articles in three top OM journals (JOM, MSOM, POM) published in 2022 and 2023, we identify 24 articles (47%) that adopt TWFE staggered DiD designs and are subsequently *susceptible* to biased and misleading results (Appendix 1). Therefore, in the top three OM journals, approximately 1 out of 2 DiD articles is *susceptible* to biased and misleading results.

In the recent wake of conducting responsible operations management research (MSOM 2020 Issue 6), we accordingly call for responsible research in drawing causal inference for treatment effect in the OM field. Despite the recent call reminding researchers of the importance of correctly assessing treatment effect in the OM field (Shang and Rönkkö 2022, Mithas et al. 2022, Barrios et al. 2022), there is no systematic review and in-depth analysis on why staggered DiD design suffers from biased estimates and how to choose an appropriate estimator to avoid the bias. Accordingly, current study aims to fill this void for the OM research community – hence our key research objective. This is especially important for the business discipline as publications in flagship business journals draw tremendous amount of attention from the practice whereas research results were disseminated through public interviews, national and state journals,

professional conferences, and etc. To educate practitioners with biased research findings might result in catastrophic business consequences, harming both the stringency of academic research and the practicality of business practices.

We start our research objective by surveying empirical OM research published in three top OM journals (JOM, MSOM, POM) from January 2022 to May 2023 (at the time of writing). We define empirical OM research as those research "working with empirical data sets" following Simchi-Levi (2022, p.2). We identify a total of 181 empirical studies, 51 studies (28%) of which utilize TWFE DiD and/or event study to draw causal inference (Appendix 2) – approximately 1 out of 3 empirical articles uses TWFE DiD estimator. These 51 studies either investigate the average treatment effect on the treated (ATT) within a certain period of time after the treatment or examine how the treatment effect evolves over time following the treatment. We define the former as "static effect" and the latter as "dynamic effect" following recent econometrics literature (Sun and Abraham 2021, Baker et al. 2022). Dynamic effect is normally referred to as event study methodology in the OM literature (Dong et al. 2019, Cui et al. 2019, Li and Wu 2020, Cui et al. 2022). We then categorize the different DiD designs in these 51 studies into five general designs. The first design is a canonical 2×2 design where there are only two groups and two time periods. The second and third design examine static and dynamic effect respectively in a multi-group multi-period setting where the treatment timing is the *same* for the treatment groups. The last two designs also assess static and dynamic effect respectively in a multi-group multi-period setting. However, the treatment timing is *different* for the treatment groups – hence the name of staggered DiD design (Wooldridge 2010).

Among the five categories of DiD designs, only the last two designs, i.e., staggered DiD with static and dynamic effect, are prone to biased estimates when using TWFE as the estimator (more in Section 2). Therefore, we only focus on staggered DiD design in the current study. Research on the pitfalls associated with TWFE staggered DiD has exploded in recent years in the econometrics field (Deshpande and Li 2019, De Chaisemartin and d'Haultfoeuille 2020, Callaway and Sant'Anna 2021, Sun and Abraham 2021, Baker et al. 2022, Roth et al. 2023), making it challenging for empirical OM researchers to differentiate between the different alternative estimators and select an appropriate estimator for implementation in empirical research settings. To provide a practical guidance on this topic, we synthesize the recent econometrics research on staggered DiD design by theoretically explaining the sources of the biases associated with TWFE and systematically comparing the differences among the different alternative estimators. To facilitate the selection of the appropriate estimator, we provide a flowchart that researchers can use as a guidance to select the appropriate estimator based on their data and research settings. To facilitate the implementation process after selecting the right estimator, we also highlight the key areas that researchers should pay attention to during the implementation process of the selected estimator. Lastly, we use two examples – simulation and replication study – to empirically debunk the sources of the biases and show

how alternative estimators can help to produce unbiased estimates. We also provide replication code for these two empirical examples so researchers can easily adapt our code for easy implementation.

The validity of all the five DiD designs crucially rely on one important assumption – the parallel trend assumption (although this assumption is not testable in a 2×2 design), i.e., control group and treatment group should trend approximately at the same rate on the outcome variable before the treatment. A common practice to test the parallel trend assumption is to use TWFE event study methodology to test the coefficients on the leads (i.e., the time indicators before treatment). In the 51 studies we surveyed, 35 studies (69%) test the parallel trend assumption and 27 of them use the coefficients on the leads to test this assumption. However, using coefficients on the leads from TWFE event study to test the parallel pretrend can lead to significant distortion in causal inferences in a staggered DiD design (Roth 2020, de Chaisemartin and D'Haultfoeuille 2020, Callaway and Sant'Anna 2021). In our staggered DiD simulation example, we show that TWFE event study incorrectly estimates a pretrend when there is no pretrend (parallel pre-trend) in the data. We further show that the newly developed alternative estimators can not only correctly estimate the pretrend but also correctly estimate the treatment effect. We accordingly call for researchers to start using the newly developed alternative estimators to correctly test the parallel trend assumption as well as the treatment effect.

We make three major contributions to the OM field. First, our study is an extension to the recent methodological review of field experiment (Gao et al. 2023) as well as the recent reproducibility project of laboratory experiment in the OM field (Davis et al. 2023). We focus on the missing leg of quasi-experiment and remind researchers of the importance to correctly draw statistical inference from quasi-experiment designs to increase research validity. Second, despite DiD being used as a workhorse in empirical OM research and despite the call for rigorous treatment effect research (Shang and Rönkkö 2022, Mithas et al. 2022), when and how to use the right DiD estimator remain unanswered. Although recent econometrics literature presents numerous new alternative DiD estimators, which alternative to use for what kind of research settings also remains unanswered, only making it more challenging for empirical OM researchers to decide between different alternatives. Our study, accordingly, synthesizes the recent development of staggered DiD design in the econometrics field and provide a detailed step-by-step guide for OM researchers to use and choose an appropriate estimator. In addition, we highlight several key areas when implementing the alternative estimators, which were not explained in recent DiD literature. We also use simulation and replication examples to help researchers further understand the implementation process. Third, we provide researchers with implementation code of our simulation and replication examples for easy application. We empirically demonstrate that 1) the widespread practice of using TWFE event study methodology to test the parallel trend assumption can result in misleading inferences (incorrectly estimate a pre-trend when there is none); and 2) TWFE estimator produces biased result for both static and dynamic

effect in a staggered DiD design. We also show that alternative estimators can correctly estimate both the pretrend and the treatment effect. We accordingly call for researchers to adopt the alternative estimators to test the parallel pretrend as well as the treatment effect.

There are two important notes in our study we want to highlight. First, the current study does not aim to critique or negate any of the 51 studies surveyed as methodologies always evolve. Instead, our purpose is to remind researchers of the deficiencies associated with TWFE (at the time of writing) in a staggered DiD design and how to overcome these deficiencies. By doing so, we hope to promote research stringency and knowledge advancement for the whole OM research community. Second, we replicate one published OM study to facilitate the understanding of the "why" and "how". However, we do not aim to conduct an exhaustive replication study of all staggered DiD studies to document the status of bias, such as how many studies are positively/negatively biased. We aim to use the replication example to help researchers better understand the rationale of why the biases occurred and how to handle them.

2. WHY TWFE PRODUCES BIASED ESTIMATES IN STAGGERED DID DESIGN

Among the 51 studies we surveyed, only one study briefly discussed the issues associated with estimating static effect in a staggered DiD design (Barrios et al. 2022). Albeit some researchers claim that a TWFE staggered DiD design increases the validity of this design, we explain how TWFE produces biased estimates when used to estimate treatment effects in such a design.

First, violation of the parallel trend assumption will lead to biased estimates. This rule applies to all the five different DiD designs. In a nutshell, the parallel trend assumption states that the treatment group and the control group should trend at similar rates on their outcomes before the treatment takes place such that the pattern observed after the treatment cannot be simply explained by a linear violations of parallel trends (Rambachan and Roth 2023). Second, violation of the no-anticipation assumption can also lead to biased estimates. This rule also applies to all the five different DiD designs. In contrast to the well-known parallel trend assumption, no-anticipation assumption draws less attention in empirical OM research. No anticipation assumption basically states that the treatment has no causal effect prior to its implementation (Roth et al. 2023). In other words, all units in the treated group do not have prior information about when their treatment is going to start and, therefore, do not act in advance before treatment starts. The validity of this assumption largely depends on how well researchers know their data/subjects.

Even when both the parallel tread assumption and the no-anticipation assumption hold, there are still other causes that can lead to biases estimates. However, these biases only associate with the static and dynamic treatments effect in a staggered DiD design. The canonical 2×2 design and the single treatment in a multi-group multi-period design are not prone to these biases (see details in Online Supplement). In a staggered DiD design, OM researchers tend to estimate two different treatment effects: static effect and

dynamic effect. Static effect estimates a single treatment effect which is time invariant and answers the following question: what is the average treatment effect for all units who have participated in the treatment up to time period *T*? (Sun and Abraham 2021). Often referred to as event study methodology, dynamic effect allows treatment effects to be estimated at each time period both before and after the treatment and is a widespread practice to test how the treatment effect evolves over time (Sun and Abraham 2021, Baker et al. 2022). We explain why staggered DiD design are prone to biased estimates when using TWFE as the estimator even when the parallel tread assumption and the no-anticipation assumption hold.

In its simplest form, TWFE estimator in a staggered DiD design can be written as Equation 1 and Equation 2: a regression of an outcome variable on a set of unit and time dummies along with a single treatment indicator (for the static specification) or a treatment indicator interacted with relative time since treatment (in the dynamic specification). Throughout this article, we use i = 1 ... I to represent cross-sectional units, t = 1... T to represent measurement occasions, l = 1... L to represent time periods after the treatment, g = 1 ... G to represent different groups, and δ or δ^{DD} to indicate the treatment effect. For example, different individuals i coming from different income groups g have improved their earnings by δ in a time period l after receiving a treatment.

$$Y_{it} = \alpha_i + \lambda_t + \mu_g \sum_{l \geq 0} D_{i,t}^{l} + v_{i,t} \qquad \text{Equation 1}$$

$$Y_{it} = \alpha_i + \lambda_t + \sum_{l=-K}^{-2} \mu_l D_{i,t}^{l} + \sum_{l=0}^{L} \mu_l D_{i,t}^{l} + v_{i,t} \qquad \text{Equation 2}$$

In both Equations, $D_{i,t}^{l}$ is the indicator of being l periods after the treatment. In the static specification of Equation 1, a single treatment indicator ($\sum_{l \geq 0} D_{i,t}^{l}$) collects all treatment indicators into a single post-period. In the dynamic specification of Equation 2, the main parameters of interest are the μ_ls which capture the differences in the outcome Y_{it} between treated and untreated groups l time periods apart from the time when the treatment started. For both of these specifications, applied econometrics literature shows that when treatment timing is different and treatment effects are heterogeneous, both single (Equation 1) and interacted (Equation 2) treatment coefficients fail to consistently estimate a causally interpretable estimand. Ideally, the target parameter that naturally extends the Average Treatment Effect on the Treated (ATT) from the canonical 2×2 design is the Cohort-specific Average Treatment Effects on the Treated (CATT), which are defined as:

$$ATT(g,t) = \mathbb{E}\left[Y_t(g) - Y_t(0) \middle| G_g = 1\right]$$

Where G_g denotes the indicator for a particular treatment cohort. In particular, if treatment effects are heterogeneous across treatment cohorts, which almost always stays true in OM DiD studies, the static regression coefficient $\sum_{l \geq 0} D_{i,t}^{l}$ from Equation 1 recovers only a linear combination (with potentially

negative weights) of the CATTs above while the dynamic regression coefficient $\sum_{l=0}^{L} \mu_l D_{it}^{l}$ are a linear combination (also with potentially negative weights) of the average treatment effect from its own time period as well as from other relative time periods, both leading to biased estimates. Entire econometrics papers have been written to explain why TWFE is not robust to treatment effect heterogeneity in a staggered DiD design (de Chaisemartin and D'Haultfoeuille 2020, Goodman-Bacon 2021, and Sun and Abraham 2021, Borusyak et al. 2024). To sum the literature, the key is that when TWFE estimates treatment effect from a staggered design, OLS effectively conducts four different comparisons among different groups: early-treated (as treatment) compared to later-treated (as controls); later-treated (as treatment) compared to always-treated (as controls); later-treated (as treatment) compared to early-treated (as controls); and treated (as treatment) compared to un-treated or never-treated (as controls). If there are no never-treated control groups as is the case with some of the 51 empirical papers we surveyed, then OLS effectively conducts two comparisons: early-treated groups *vs* later-treated groups, and later-treated groups *vs* early-treated groups. OLS then computes the average treatment effect using variance based weights from each 2×2 comparison. The potential problematic comparisons are the *later-treated vs always-treated* and *later-treated vs early-treated* where always-treated and early-treated groups are used as controls. The reason is that when used as controls, always-treated groups and early-treated groups have received treatments already, so the changes in these two groups over time may have already reflected the treatment effect. Therefore, the comparison is not valid. In other words, always-treated and early-treated groups are not "clean controls" and the comparisons based on these two groups are not clean 2×2 designs.

Further, when OLS subtracts the changes of always-treated groups and early-treated groups from later-treated groups to compute weights, negative weights can occur if the changes in the early-treated groups are bigger than the changes in the later-treated groups (Sun and Abraham 2021). In an extreme case, Baker et al. (2022) demonstrated that even when ATT for every treated group is positive, the average treatment effect across all groups can be negative and statistically significant. This is because the treatment effect in TWFE is a weighted average of all possible simple 2×2 DiDs including using already/early-treated groups as effective controls for not-yet-treated groups (Goodman-Bacon 2021). In addition, if treatment effect extends beyond more than one period, the changes in the outcome of the always/early-treated units will be contaminated by changes in treatment effect itself (Goodman-Bacon 2021). Even in situations when the sign of the weights is not negative, the weights themselves can be driven by TWFE estimation methods and other factors such as number of time periods and group size (Goodman-Bacon 2021, Sun and Abraham 2021).

In addition to the biases caused by the above-mentioned unclean control groups (always-treated and already-treated), even when the parallel trend assumption and the no-anticipation assumption hold, so that the CATTs are, in principle, nonparametrically identified, the estimators in Equation 1 and Equation 2

are based on regressions that are not fully saturated and hence misspecified (more details in Online Supplement on saturated *vs* non-saturated regression). Theoretically, in a staggered DiD design, a fully saturated regression allows each group and each time period since the treatment to have potentially heterogeneous dynamic treatment effects (hence ATT(g,t)). However, in practical applications, the static and dynamic specification are usually non-fully saturated regressions (such as missing some of the potential interactions), hence resulting in the weighted average problem detailed out by recent econometrics literature (Sun and Abraham 2021). The role of this mis-specification is attenuated if treatment effects are homogeneous across cohorts, which unfortunately does not hold for most, if not all, staggered DiD designs in the OM field.

To sum up, in a staggered DiD design, the biases of the TWFE as an estimator for the treatment effect may come principally from each of the four sources discussed above (or any combination of them): 1. violation of the parallel trend assumption; 2. violation of the no-anticipation assumption; 3. always-treated and already-treated units are used as effective controls; and 4. misspecification bias due to lack of fully-saturated regression function when treatment effects are heterogeneous.

3. COMPARISON AMONG ALTERNATIVE ESTIMATORS

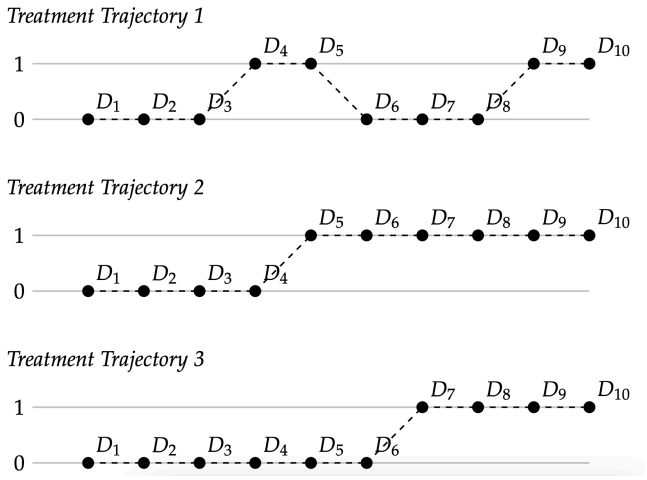
To solve the biases associated with TWFE estimator in a staggered DiD design, recent econometrics literature has made significant progress. At the time of writing, three prominent new estimators have been proposed (De Chaisemartin and d'Haultfoeuille 2020, Callaway and Sant'Anna 2021, and Sun and Abraham 2021). Despite the heated discussion on this topic in the econometrics field, little attention has been paid to the pragmatic application of these different estimators. Therefore, we synthesize the different estimators to provide empirical OM researchers a useful guide to select the right estimator based on idiosyncratic research settings.

Following Baker et al. (2022), we also add the method of Deshpande and Li (2019) to our portfolio of alternative estimators. Therefore, we compare four different estimators in our current paper: Deshpande and Li (2019), De Chaisemartin and d'Haultfoeuille (2020), Callaway and Sant'Anna (2021), and Sun and Abraham (2021). Before we establish which estimator to use under what kind of research settings, we first elaborate the differences between these four estimators in this section. We start with defining two fundamental concepts related to staggered DiD design.

3.1 Treatment Trajectory and Absorbing State

A treatment trajectory for individual i is simply a full history of the $(D_{it})^{T}{}_{t=1.}$ Figure 1 shows some examples of a treatment trajectory.

Figure 1 Treatment Trajectory



In treatment trajectory 1, an individual i is not treated for the first three time periods, then treated for 2 periods in $t = 4$ and $t = 5$, then becomes untreated again for another three time periods, and then becomes treated again in $t = 9$ and $t = 10$. Conversely, in treatment trajectories 2 and 3, the first treatment is in period $t = 5$ and period $t = 7$ respectively and individual stays treated thereafter. Treatment trajectory 1 is different from treatment trajectory 2 and 3. In treatment trajectory 2 and 3, once a unit is treated it stays treated. This is precisely the notion of treatment as an absorbing state. In treatment trajectory 2 and 3, treatment is an absorbing state as there is a t^* such that for all $t > t^*$, $P(D_{it} = 1|D_{it^*} = 1) = 1$. Whilst in treatment trajectory 1, treatment is not in an absorbing state as the treatment switches on and off for the same individual.

3.2 Differences Among Alternative Estimators

A first major difference among the four alternative estimators is that De Chaisemartin and d'Haultfoeuille (2020) allows for general treatment trajectories (i.e., non-absorbing state) while Sun and Abraham (2021) and Callaway and Sant'Anna (2021) require treatment to be in an absorbing state. Deshpande and Li (2019) also implicitly consider an absorbing state, as in their example of studying the effect of application costs on disability programs, Deshpande and Li (2019) identified these effects using the closings of Social Security Administration offices as an example. Office closings are treated as irreversible in their study hence is an absorbing state. In other words, Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Deshpande and Li (2019) are restricted to staggered DiD design while De Chaisemartin and d'Haultfoeuille (2020) is not restricted to staggered DiD design. In Figure 1, treatment trajectory 2 and 3 fit the staggered DiD design while treatment trajectory1 does not fit the staggered adoption design. However, if researchers redefine their treatment as "ever having received treatment before", then, treatment trajectory 1 can also be applied to a staggered design.

The second major difference among the four alternative estimators lies in the estimation mechanism of the treatment effect. Deshpande and Li (2019) use several versions of TWFE DiD regression design (that does not appear to be fully saturated) to estimate treatment effect. In contrast, the common approach of De Chaisemartin and d'Haultfoeuille (2020), Sun and Abraham (2021), and Callaway and Sant'Anna (2021) estimators is to separately estimate the treatment effect at various levels of disaggregation. For Callaway and Sant'Anna (2021) and Sun and Abraham (2021) who strictly consider a staggered adoption, this amounts to estimating each CATT at the cohort-relative time pair level, denoted as (g, t) in Callaway and Sant'Anna (2021) and (e, l) in Sun and Abraham (2021). As De Chaisemartin and d'Haultfoeuille (2020) focus on instantaneous effect as a target treatment effect, De Chaisemartin and d'Haultfoeuille (2020) estimator does not estimate a separate CATT for each time period. Instead, all suitable DiD comparisons are aggregated across cohorts and time periods into *only two groups* (leavers and joiners), each reflecting a change in the non-absorbing change of treatment, hence the instantaneous treatment effect.

Despite Deshpande and Li (2019) use non fully-saturated regressions to estimate treatment effect, Baker et al. (2022) show that Deshpande and Li (2019) method performs equally well as Callaway and Sant'Anna (2021) and Sun and Abraham (2021) in terms of correctly estimating the treatment effect, leading to the third difference between the four alternatives – how the estimation is developed and by using what kind of control groups. In Sun and Abraham (2021), each CATT is obtained by first identifying a suitable control group: the choice of an appropriate control group depends both on the form of the parallel trend assumption that is maintained and the degree of anticipation. Two different identification strategies are available in Sun and Abraham (2021): never-treated or not-yet-treated groups. In a setting where there is no never-treated group, Sun and Abraham (2021) use the last cohort to be treated (last-treated) as a control group. Callaway and Sant'Anna (2021) estimator is developed in three steps: in the first step the CATTs are obtained as DiD estimates on each subset of the data. In the second step, cohort treatment probabilities are estimated with sample proportions. Finally, the third step combines the two estimators by weighting the estimated CATTs according to their estimated proportions. Callaway and Sant'Anna (2021) also use never-treated units as effective controls. However, when there is no never-treated units, Callaway and Sant'Anna (2021) use not-yet-treated units as effective controls, instead of using last-treated units as in Sun and Abraham (2021). De Chaisemartin and d'Haultfoeuille (2020) also identify a suitable control group. For the joiners (who switch from 0 to 1), the suitable controls are the groups who stayed untreated from the previous period. For the leavers (who switch from 1 to 0), the suitable controls are the groups who stayed treated from the previous period.

Deshpande and Li (2019) method is not an estimator per se as Deshpande and Li (2019) still use TWFE estimator. Like the other three estimators that all modify the units that could serve as clean control groups, Deshpande and Li (2019) specifically use not-yet-treated units as clean control groups. In addition,

unlike the other three estimators where the selection of clean effective controls are constructed in the background by the estimator, Deshpande and Li's (2019) clean control group is achieved by slicing and reconstructing the original data such that only not-yet-treated units are used to compare to treated units for each treatment cohort. Then, all the slices of clean comparisons of all treatment cohorts are stacked together and TWFE estimator is applied on this stacked data. Therefore, Deshpande and Li (2019) method is also referred to as "stacked" regression.

The last major difference among the four alternative estimators lies in the flexibility and ease of application. Despite all four alternatives can easily accommodate covariates, Callaway and Sant'Anna (2021) estimator allows covariates to enter under a conditional parallel trend assumption, i.e., when covariates are predictive of treatment effect heterogeneity and/or of treatment value. Because the version of the parallel trend assumption conditional on covariates is weaker than the unconditional parallel trend assumption, Callaway and Sant'Anna (2021) estimator therefore is more robust. However, the inclusion of the covariates in parametric form opens up the possibility of misspecification in the way covariates enter the regression function or the propensity score. This limitation is partially mitigated by the doubly robust estimator in Callaway and Sant'Anna (2021), which does not suffer from misspecification bias as long as either the generalized propensity score or the outcome regressions are correctly specified. In addition, Callaway and Sant'Anna (2021) also allow researchers to incorporate institutional knowledge about treatment anticipation, therefore enlarging the scope of applications where the method is applicable, relative to a strict form of no anticipation assumption, which is implicitly or explicitly featured in all the other three studies. Next, Callaway and Sant'Anna (2021) also allow researchers to evaluate which control groups are most appropriate depending on the version of the parallel trend assumption that researchers wish to maintain. In applications where the never-treated units may have very different outcome trajectories from the units that are eventually treated, the not-yet-treated units form a more suitable control. This flexibility allows researchers to incorporate institutional knowledge in their choice of preferred estimator and increase the range of applications where it can be deployed. Finally, Callaway and Sant'Anna (2021) does not focus on a single estimand but offers strategies to recover a variety of target parameters that are all weighted averages of CATTs (more in Section 5.1). This makes the method suitable to answer a richer set of policy questions.

As the effective control groups used by Sun and Abraham (2021) are slightly different from Callaway and Sant'Anna (2021) as discussed above, Sun and Abraham (2021) rely on the unconditional form of parallel trend assumption – different from Callaway and Sant'Anna (2021). But otherwise, these two estimators are interchangeable. However, one advantage of Sun and Abraham (2021) is the easy implementation as it is regression-based, which might be the most familiar form to OM empirical researchers.

Turning to De Chaisemartin and d'Haultfoeuille (2020), De Chaisemartin and d'Haultfoeuille (2020) is not limited to application where the treatment is an absorbing state. Therefore, it can accommodate richer forms of treatment histories and their method is suitable for a wider range of applications. De Chaisemartin and d'Haultfoeuille (2020) also allows for a fuzzy version of DiD which broadens the set of examples to which the method can be applied. One of the caveats is that to obtain a new consistent estimator of the ATT parameter outside of the staggered adoption design, a parallel trend assumption for the treated potential-outcome is needed to guarantee that one can identify the counterfactual outcome for groups who *leave* the treatment. This assumption does not appear in the staggered design because treatment is in an absorbing state. One of the additional assumption requires a valid comparison group for each DiD comparison. These stable groups play the role of the not-yet-treated or never-treated group as in Callaway and Sant'Anna (2021) and Sun and Abraham (2021). We summarize the differences among the four alternatives in Table 1 for easy reference.

Table 1 Comparison Between Four Alternative Estimators

	De Chaisemartin and d'Haultfoeuille (2020)	Sun and Abraham (2021)	Callaway and Sant'Anna (2021)	Deshpande and Li (2019)
Restricted to Staggered?	No	Yes	Yes	Yes
Treatment in Absorbing State?	No	Yes	Yes	Yes
Need to Reconstruct Data?	No	No	No	Yes
Accommodate Covariates?	Yes	Yes	Yes	Yes
Allow No-anticipation Assumption?	No	No	Yes	No
Conditional Parallel Trend Assumption?	No	No	Yes	Yes
Effective Control Group	untreated (for joiners) and treated (for leavers) in the previous period	never-treated, last treated	never-treated, not-yet-treated	not-yet-treated
Focus of Treatment Effect	Instataneous	Static and Dynamic	Static and Dynamic	Static and Dynamic

4. SELECT AN ALTERNATIVE ESTIMATOR

Having discussed the differences among different alternatives, we now provide a step-by-step guide for empirical OM researchers to select an appropriate alternative estimator. Note that researchers need to use all information and knowledge they have about their data as well as their domain-specific expertise to choose an appropriate estimator.

In **Step 1**, researchers should consider if the research setting fits the staggered adoption design. While De Chaisemartin and d'Haultfoeuille (2020) estimator can be easily extended to the staggered adoption design, it stands out mostly due to its ability to accommodate treatments that are not an absorbing state (treatment switches on and off). However, it is always possible to redefine the treatment as "having been treated at least once", which mechanically induces the staggered adoption design in De Chaisemartin and d'Haultfoeuille (2020) estimator. However, this re-definition of treatment and encoding cannot distinguish two treatment paths after the first time the units were treated, but it may be sufficient to answer the research question at hand.

In **Step 2**, researchers consider whether all units in the treatment groups are subject to the same treatment. For example, if the groups coincide with counties or states subject to a policy change, this means that all units in the treated counties or states are treated, and all the units in the untreated states are untreated. If the treatment variable can be defined at the group level, all four estimators can be considered. If this is not the case, De Chaisemartin and d'Haultfoeuille (2020) can accommodate the so-called fuzzy-DiD, where treatment is not uniform at the group level. This is another feature De Chaisemartin and d'Haultfoeuille (2020) stands out from the other three estimators.

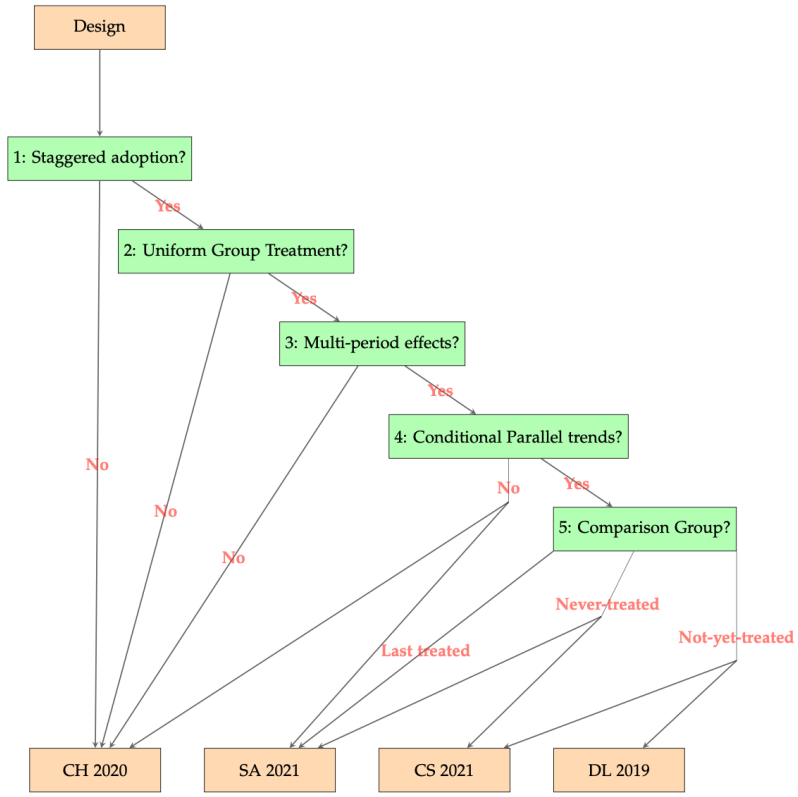
In **Step 3,** researchers specify which estimand is the target. If instantaneous treatment effect is of interest, then all estimators are suitable. If instead multi-period effect, which include both static and dynamic effect, or their corresponding different forms of aggregates, is of interest, De Chaisemartin and d'Haultfoeuille (2020) is not suitable and one of the other three estimators must be used.

In **Step 4,** researchers consider the information from the covariates. If covariates are predictive of treatment effects and the distribution of the covariates changes a lot across cohorts, the unconditional form of the parallel trend assumptions that appears in De Chaisemartin and d'Haultfoeuille (2020) and Sun and Abraham (2021) may not be credible. However, in such cases, it is still possible that after controlling for covariates, the conditional parallel trend assumption is valid for some comparison groups. If so, researchers can use the strategy in Callaway and Sant'Anna (2021). Researchers should note that because Callaway and Sant'Anna (2021) use a parametric propensity score, non-parametric control of the covariates is unlikely to hold and one must rely on a correct parametric specification.

Finally, in **Step 5,** the researchers consider which version of the parallel trend assumption is likely to hold in the data, and hence which control groups should be chosen for the appropriate DiD comparison. These choices largely depend on how well researchers know their data. The most obvious choice is the never-treated group, i.e. units that have never experienced the treatment. If the data from the researchers indeed consists of such never-treated groups, then both Sun and Abraham (2021) and Callaway and Sant'Anna (2021) estimators are available to use although the authors of both papers invite caution. This is because, sometimes, for example, in healthcare operations, the never-treated units (patients or subjects) are unlikely to be a valid comparison group, as the evolution of their outcomes is likely to be profoundly different (due to other underlying unobserved health shocks). If this is the case, the parallel trend assumption is unlikely to hold for the never-treated units. So, using these never-treated units as a comparison group introduces bias in the CATTs. Other times, the never-treated units only account for a very small sub-sample, so if this small sub-sample of never-treated units are used to predict the mean evolution of the untreated outcome for the untreated units, these never-treated units will introduce a very noisy estimate in the DiD comparison and inflate the variance of the CATTs. In either case, it may be better to rely on a different parallel trend assumption, such as using not-yet-treated group (Callaway and Sant'Anna 2021) and/or last-

treated group (Sun and Abraham 2021) as the comparison group, therefore, bypassing some of the problems discussed above. Further, if researchers believe that the parallel trend assumption holds for not-yet-treated group, then both Callaway and Sant'Anna (2021) and Deshpande and Li (2019) are available to use. Whether never-treated, not-yet-treated, or last-treated units form a more plausible control group largely depends on the empirical application. The choice eventually boils down to researcher's knowledge and understanding of their data and the associated research topics.

Figure 2 How to Choose An Alternative Estimator



These five steps are summarized in Figure 2. For illustrative purpose, we use an existing staggered DiD study (Wang and Overby 2022) to empirically show the selection process. Wang and Overby (2022) investigated how the availability of market lending impacts per capita consumer bankruptcy filing. Market lending is proxied by the availability of service from Lending Club – the biggest online platform that started its peer-to-peer loan services in 2007. Wang and Overby (2022) compiled 28 quarterly data from 2008Q1 to 2014Q4 to assess the impact at county-quarter level. Table 2 is a mini version of the original dataset. For county 1019, Lending Club service becomes available from the 5th occasion, hence "Availability of Lending Club" was coded as 0 before the 5th occasion and 1 after the 5th occasion. For other counties, the first available date of Lending Club service varies and was coded as different occasions. Therefore, this is a typical staggered design with multiple groups and multiple time periods. "Per Capita Bankruptcy Filing" is

the dependent variable. Control variables used in Wang and Overby (2022) are: population, number of employees, average monthly earnings per individual, size of the labor force, and median household income. Wang and Overby (2022) matched two different datasets: 50 states dataset (including the district of Columbia) where the service of Lending Club was approved since it started its business in 2007; and 9 states dataset where Lending Club service was approved after 2010. We show the estimator selection process in the following five steps.

Table 2 Example Data of Wang and Overby 2022

County	Occasion (Tt)	Availability of Lending Club (LC_{it})	Per Capita Bankruptcy Filing (Y_{it})
1019	1	0	1.35
1019	2	0	0.90
1019	3	0	1.07
1019	4	0	1.11
1019	5	1	1.02
1019	6	1	1.27
…	…	…	…
1019	19	1	0.92
1019	27	1	1.08
1019	28	1	0.58

1. Staggered Adoption? Yes. The approval of a Lending Club is in an absorbing state. If approval can be revoked in a substantial number of units in the sample, one could still use De Chaisemartin and d'Haultfoeuille (2020) estimator to assess the instantaneous treatment effect.
2. Uniform Group Treatment? Yes. Because the policy changes at the state level, all states in which it is approved are subject to the same treatment value.
3. Multi-period Effects? Yes. Wang and Overby (2022) investigated the long term effect of the availability of market lending on bankruptcy filings following the immediate period when the approval was granted (instantaneous effect in De Chaisemartin and d'Haultfoeuille 2020). Both static and dynamic effects may be used to answer this research question.
4. Conditional parallel trends? Yes. Because additional covariates like number of employees, average earnings per individual, size of the labor force, and median household income are available, one could use the regression adjustment, propensity score weighting, or doubly robust approaches featuring these variables. In addition, the underlying parallel trend assumption may only be plausible if we control for the number of employees and other covariates. For example, if average earnings per individual helps predict the likelihood of a filing for bankruptcy (mostly likely yes) and average earnings per individual varies across groups of states (highly likely yes), then trajectories of offered loans to earlier-treated *vs* later-treated *vs* never-treated individuals would be different. Hence, the conditional parallel trend assumption by incorporating these covariates.

5. Comparison group? Looking at the data structure of Wang and Overby (2022), there are two choices of comparison groups: never-treated and not-yet-treated. In the 50 states data, however, never-treated units are a very small group. The smaller the percentage of never-treated units, the more biased the results from the TWFE DiD regression (Callaway and Sant'Anna 2021). Therefore, using not-yet-treated group in the 50 states data would be a more plausible choice.

5. THINGS TO CONSIDER WHEN IMPLEMENTING ALTERNATIVE ESTIMATORS

After researchers have finalized which estimator to use following the flowchart in Figure 2, researchers now move on to the implementation stage. In this section, we discuss some practical strategies during the implementation stage.

5.1 Report Treatment Effects At Different Levels of Disaggregates

Callaway and Sant'Anna (2021) provide estimators for the most disaggregated CATTs at the (*g*,*t*) (group and relative time) level, one could choose to report a more parsimonious summary that averages over the cells with appropriate weights. In Callaway and Sant'Anna (2021), treatment effect θ takes the form of:

$$\theta = \sum_{g \in G} \sum_{t=2}^{T} w(g,t) \cdot ATT(g,t)$$

The weights should be chosen depending on the target parameter of interest, which is in turn, defined jointly by the research question that empirical researchers are addressing. If the research question emphasizes the time dynamics of treatment effects over the heterogeneity in treatment cohort, one may obtain a summary by aggregating over groups: $\theta_{es}(e) = \sum g \in \mathcal{G}\, \mathcal{W}_e^{es} ATT(g, g+e) = \mathbb{1}[g + e \leq T]\, \mathbb{1}\, [t - g = e]\, P\, (G = g | G + e \leq T)\, ATT(g, g + e)$. The instantaneous treatment effect studied in De Chaisemartin and d'Haultfoeuille (2020) is one special case of this parameter with $e = 0$. If researcher is interested to report a comparison between longer exposures *vs* shorter exposures to treatment, one must note that the change between $\theta_{es}(e)$ and $\theta_{es}(e')$ is not only due to a treatment effect dynamics but also due to a compositional change in the groups. Callaway and Sant'Anna (2021) provide a variation of this parameter which balances out the group composition so that only changes in treatment effects are isolated. If the research question focuses on how the particular circumstances of treatment timing affect the overall trajectory of outcomes, perhaps because certain macroeconomic conditions were in place at the time of treatment, then aggregating over time is the preferred choice: $\theta_{sel}(g) = \frac{1}{T-g+1} \sum t = g^T ATT(g,t)$ captures the trajectory effect of having been treated in group/cohort *g*. If one wants to further aggregate it across groups to obtain an overall average with an analogous interpretation as the ATT in the $G = 2$, $T = 2$ case, this is given by the weighting scheme: $\theta_{sel}^{O} = \sum_{g \in \mathcal{G}} \theta_s\, el(g) P(G = g | G \leq T)$. If the research question emphasizes the effect of treatment at a particular calendar time, without emphasizing differential length of exposure to treatment, on can weight according

to the scheme:$\theta_c(t) = \sum_{g \in G} \mathbb{1}t \geq g\, P(G = g|G \leq t)ATT(g,t)$. Similar to the overall group average, $\theta_C(t)$ can be further aggregated over time: $\theta_c^O = \frac{1}{T-1}\sum_{t=2}^{T} \theta_c(t)$.

In practice, we recommend that researchers define a target parameter that is most appropriate to answer their research questions, then, select the appropriate weighting scheme for the CATTs and compute the desired treatment effects. Fortunately, aggregating and reporting treatment effects at different levels as discussed above is very straightforward in implementation. Callaway and Sant'Anna (2021) provide easy-to-implement functions with which researchers can easily modify the corresponding arguments to select different weighting schemes.

5.2 Doubly Robust Procedure in Callaway and Sant'Anna (2021)

The key parameter to implement the doubly robust procedure of Callaway and Sant'Anna (2021) is the generalized propensity score $p_{g,s}(X) = P(G_g = 1|G_g + (1 - D_s)(1 - G_g))$. It is the conditional probability (as a function of the covariates X) of g's group indicator dummy $G_g = 1$ within the sample of not-yet-treated as of time s units. Typically, a researcher would first specify a generalized propensity score model that, while parametric, is flexible enough. For example, one could choose a logit model with linear and interaction terms for the observable characteristics of interest that are considered relevant for the treatment selection. Once that is specified, the researcher specifies a functional form for the outcome regression for each group g's baseline potential outcome. Finally one can combine them and follow the traditional doubly robust estimation.

5.3 Data Structure of Estimators

As previously discussed in Section 3.2, researchers must re-construct the original data to use the method proposed by Deshpande and Li (2019). However, if researchers intend to use the estimators of Sun and Abraham (2021) and Callaway and Sant'Anna (2021), researchers can use the original data directly for analysis without the need of reconstructing the data. This is because both Sun and Abraham (2021) and Callaway and Sant'Anna (2021) estimators do not require homogeneity of treatment effects over time (in fact, the time and group heterogeneity is the very cause of the negative weights issue discussed in recent econometrics literature). Both Sun and Abraham (2021) and Callaway and Sant'Anna (2021) estimators are already geared towards capturing the variation of each cohort's treatment effect over the relative time of treatment. So, no further modification of data is needed for Sun and Abraham (2021) or Callaway and Sant'Anna (2021) estimators. This is in stark contrast with the TWFE estimator that is still used by Deshpande and Li (2019), where time/group heterogeneity is one of the underlying causes of negative weights in the aggregation of the CATTs, hence, the need to modify the data into a clean stacked data in Deshpande and Li (2019).

5.4 Large Scale Applications

For large scale applications, the bootstrapping procedure for the doubly robust estimator in Callaway and Sant'Anna (2021) may be very computationally demanding. If there are many cells (g, t), it may be costly to compute uniform confidence bands. On the other hand, one may be interested in reporting only aggregated summaries as opposed to all (g, t) CATTs, in which case the computation cost may be reduced. Even when the cell CATTs are the key parameter of interest, one may report point-wise instead of uniform confidence bands, which should be less costly to compute. Finally, if one is willing to maintain the stronger assumption of unconditional parallel trends, the procedure in Callaway and Sant'Anna (2021) and its computational time are greatly simplified under this scenario.

5.5 Combining TWFE DiD with Matching

Among the four alternatives, only Callaway and Sant'Anna (2021) (building upon the estimator of Sant'Anna and Zhao 2020) explicitly incorporates inverse propensity score weighting (IPW) in their proposed estimators of CATTs. In fact their doubly robust estimator features an Augmented IPW (AIPW) structure. The original idea is motivated by conditional parallel trends and it has a precursor in Abadie (2005). The bulk of these estimators explicitly incorporate the idea that after selection on observables, the parallel trend assumption is more plausible. In situations where the parallel trend assumption holds only conditionally, the DiD estimator that relies on unconditional parallel trends and does not consider different treatment probabilities would be a biased estimator for the corresponding ATT. This source of bias is in addition to the negative weights problem already detailed out in Sun and Abraham (2021). So incorporating the propensity score weighting helps mitigating this issue. One key challenge is that if there are many covariates, a non-parametric estimator for the propensity score may not converge fast enough to obtain a semiparametric rate of convergence for the ATTs. Conversely, a parametric propensity score will open the door for mis-specification. An empirically reasonable choice is to use a flexible parametric specification like a logit with interactions.

5.6 Instantaneous Treatment Effect in a Staggered DiD Design

The instantaneous treatment effect is the main target of De Chaisemartin and d'Haultfoeuille (2020) estimator while in Sun and Abraham (2021) and Callaway and Sant'Anna (2021) instantaneous treatment effect is simply one of the parameters that one can extract from the whole collection of CATTs. In fact, Callaway and Sant'Anna (2021) show that one can recover the instantaneous treatment effect by weighting the CATTs with an appropriate weighting scheme (see Section 5.1). Because it is the only parameter of interest, De Chaisemartin and d'Haultfoeuille (2020) has relatively parsimonious assumptions to estimate it, whereas Callaway and Sant'Anna (2021) and Sun and Abraham (2021) cast their assumption for identification of the general cell level CATT. For this reason, whenever the instantaneous effect is the target, De Chaisemartin and d'Haultfoeuille (2020) approach has wider applicability to empirical contexts.

Conversely, if one cares about comparing the instantaneous *vs* long run effect of a policy, the De Chaisemartin and d'Haultfoeuille (2020) estimator is not immediately applicable while Callaway and Sant'Anna (2021) and Sun and Abraham (2021)'s estimators are eligible.

6. TEST FOR THE PARALLEL PRE-TREND IN STAGGERED DID DESIGN

6.1 Review of the Parallel Trend Assumption in Operations Management

The validity of all the five categories of DiD design, including the canonical 2×2 design, rely on the parallel/common trend assumption, i.e., the treatment group and the control group should trend at similar rates on their outcomes before the treatment takes place. The parallel trend assumption in a simple 2×2 design is not testable as there are only two time periods of observations. In a multiple time periods design, however, researchers can test this assumption. Two common practices were adopted in the extant OM literature to test the parallel trend assumption. The first approach is a preliminary approach using visualization tools where researchers plot the average outcomes of all the treatment groups and the average outcomes of all the control groups for each time period before the start of the treatment. Then, researchers visually compare the trendline of the treatment group against the trendline of the control group to detect if the two trendlines are parallel to each other. The second approach is a statistical approach using event study methodology (Equation 2) to test the differences between the trend of the treatment groups and the trend of the control groups on the pre-treatment time indicators. If there is no significant statistical difference found on the time indicators prior to the treatment, researchers consequently conclude that the parallel trend assumption holds.

Among the 51 studies surveyed, 35 studies (69%) test for the parallel trend assumption, among which 4 studies only rely on visualization tools to test for the parallel trend assumption. 14 studies use event study methodology to test the statistical difference on the pre-trend between treatment groups and control groups. 13 studies use both visualization tools and event study to conduct the parallel trend assumption test. 4 studies test on the coefficient of the interaction term (treatment*pretrend time indicator). In sum, 31 (88.5%) out of the 35 studies use TWFE DiD regression to test the parallel trend assumption.

In other research fields, Roth (2022) did a survey of the applied literature and found that checking the coefficients of the leads is also a widespread practice to test for parallel trends. A popular practice as it is, using TWFE DiD regression to test the parallel trend assumption has also been proved to be problematic. Roth (2022) showed that using pre-treatment coefficients from TWFE as a test for parallel trend can lead to significant distortions in causal inference. Callaway and Sant'Anna (2021), Sun and Abraham (2021), and de Chaisemartin and D'Haultfoeuille (2020) also advise against using coefficients from TWFE to test for parallel trends. The distortion or contamination arises from the fact that the coefficient estimate is a linear combination of group specific effects from both its own time periods and from other time periods

(Sun and Abraham 2021). As such, including effects from other time periods will distort the estimate of the coefficients on the leads as well as on the lags. In addition, the coefficient estimate is also affected by both pretrends and treatment effect heterogeneity, unless strong assumptions of treatment effect homogeneity hold (Sun and Abraham 2021).

6.2 Caution of Using TWFE to Gauge Parallel Pre-trend

In the context of a staggered DiD design, the identification of group and time specific CATT (as well as their aggregated parameters) relies on different versions of the parallel trend assumption, researchers need to investigate their empirical research content of this assumption to conduct the falsifiability placebo tests. As discussed, some authors (Roth 2022) have cautioned against using pre-trends to gauge the parallel trends assumption and have instead proposed sensitivity analysis exercises. However, using pre-trend tests as a falsification is still a dominant practice in empirical OM research and researchers can use this test given the right estimator is used (more in Section 7), so we discuss how the four main approaches differ in terms of their parallel trend assumption in this section.

First, as the parallel trend assumptions in De Chaisemartin and d'Haultfoeuille (2020), Sun and Abraham (2021), and Callaway and Sant'Anna (2021) are not nested, each of them may be holding in specific contexts while being violated in others. As such, we highlight that researchers' institutional knowledge should be the primitive to guide researchers on their choices of assumption to maintain. In addition, as the different forms of parallel trends and no-anticipation assumptions are related, as highlighted by Callaway and Sant'Anna (2021), their plausibility should be evaluated jointly.

In De Chaisemartin and d'Haultfoeuille (2020), the parallel trend assumption takes the form of a strong exogeneity assumption of the potential outcomes with respect to the treatment variable: for all periods and groups (g,t) we have $\mathbb{E}\left[Y_{g,t}(0) - Y_{g,t-1}(0)|D_{g1},\cdots, D_{g,T}\right] = \mathbb{E}\left[Y_{g,t}(0) - Y_{g,t}(0)\right]$. This assumption rules out the presence of groups that entered into treatment because they experienced negative shocks. It has testable implications: a suitably constructed placebo DiD estimator comparing the evolution from period $(t-2)$ to $(t-1)$ between the group with $D_{g,t-2} = D_{g,t-1} = 0$, $D_{g,t} = 1$ and the group $D_{g,t-2} = D_{g,t-1} = 0$, $D_{g,t} = 0$ must be centered at 0, under the parallel trend assumption. So one can run a falsification test using this auxiliary placebo regression and corroborate whether the parallel trend assumption holds in the specific setting.

In Sun and Abraham (2021), the valid comparison group is never-treated (or last-treated if there is no never-treated). In the staggered adoption setting, as the first time of treatment (i.e. the event $\{E_i = e\}$) completely captures the history of treatments (the $D_{g,i} = 0$ for all periods before and $D_{g,i} = 1$ for all periods after), the parallel trend assumption is cast as $\mathbb{E}\left[Y_{i,t}(0) - Y_{i,s}(0)|E_i = e\right]$. In the staggered adoption design, Sun and Abraham (2021) and De Chaisemartin and d'Haultfoeuille (2020) assumptions coincide. Sun and

Abraham (2021) focus on showing that in the general case, without arguably strong assumptions, the test of pre-trends based on non-fully saturated regressions is not valid in general. The discussion in Sun and Abraham (2021) (Proposition 3) immediately implies that a placebo test based on TWFE is not valid, as the estimates of the pre-trends are contaminated by estimates of the treatment effects or the excluded periods. That means that even under no-anticipation assumption, one could find a rejection in the placebo test even though parallel trends hold. While Sun and Abraham (2021) does not offer an alternative valid test, they refer to the paper by Callaway and Sant'Anna (2021). Incidentally, because it is not clear whether the regression in Deshpande and Li (2019) is fully saturated, the same discussion in Sun and Abraham (2021) leads to believe that the placebo test in Deshpande and Li (2019) may also suffer from the same criticism unless strong assumptions about treatment effect homogeneity and the excluded periods are maintained.

Callaway and Sant'Anna (2021) provides two types of parallel trend assumptions, each reflecting the choice of a comparison group for their DiD estimators. The first one is based on a never-treated group. In the no-anticipation case, for all $t \geq g$ we have $\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, C = 1]$ where the event $\{C = 1\}$ corresponds to $D_t = 0$ for all time periods t, which translates to that the evolution of the baseline potential outcome for all treated cohorts is the same as the one for the never treated cohorts. Alternatively, Callaway and Sant'Anna (2021) present a not-yet-treated group as the second choice: for all t, s, g such that $t \geq g$ and $t \leq s \leq \bar{g}$ we have $\mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = \mathbb{E}[Y_t(0) - Y_{t-1}(0)|X, D_s = 0, G_g = 0]$, where $\bar{g}$ is the last period where some units are treated. The second choice says that the evolution of the baseline potential outcome for treated cohorts is the same for cohorts that are eventually treated. Finally, as a form of pre-test for either of the two parallel trend assumptions above, one can simply test whether the CATTs for the (g, t) cells corresponding to the pre-periods are equal to 0. Because Callaway and Sant'Anna (2021) provide uniform confidence bands, no-multiple hypothesis testing adjustments is necessary.

While Deshpande and Li (2019) do not spell out their parallel trend assumption in the potential outcome framework, it appears to correspond to the not-yet-treated version of the Callaway and Sant'Anna (2021) assumption. Notably, both versions of the Callaway and Sant'Anna (2021) parallel trend assumption is conditional on covariates, which coincides with all the 51 empirical OM staggered DiD studies we surveyed, where all 51 studies include covariates in their models. If covariates play a role in shaping the outcomes and if their distribution differs across cohorts, it is unlikely that the unconditional version of parallel trend assumption can hold, but perhaps the conditional one is still plausible. The conditional version of the parallel trend assumption allows the researcher to use available covariates and institutional knowledge to choose a credible identification strategy, i.e., comparing against never-treated and/or not-yet-treated.

In sum, the choice of how to test the parallel trend assumption is tightly connected and largely reflects the choice of estimator. Researchers need to decide if they wish to believe if the parallel trend

assumption in their data is unconditional or conditional. Then, researchers also need to use their institutional knowledge to select the most suitable comparison group to test the parallel trend assumption. Therefore, we defer readers to the graph in Figure 2 to select which estimator (and identification strategy for testing the parallel trend assumption) is needed on a case by case basis. However, in the next section, we further use simulation to provide more empirical evidence and guidance on how to test the parallel trend assumption during the implementation stage for empirical OM researchers.

7. PUTTING EVERYTHING TOGETHER

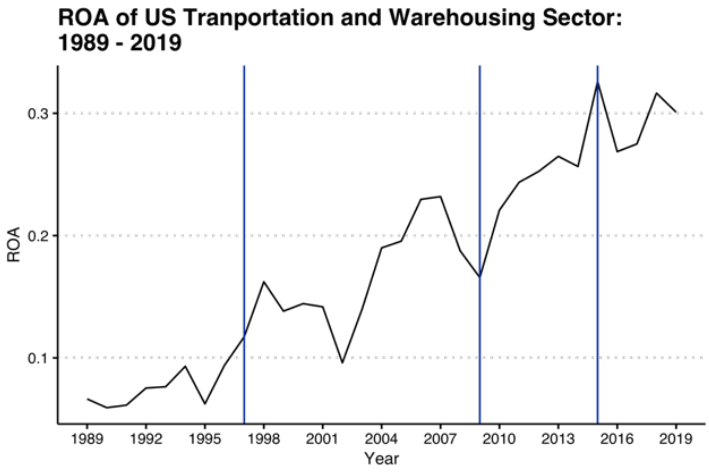
In this section, we use simulation to illustrate the pitfalls of using TWFE to estimate a staggered DiD design and show how alternative estimators can produce correct estimates. In addition, we replicate a recent OM study to further demonstrate when TWFE is used on staggered DiD, misleading results are generated.

7.1 A Simulation of Staggered DiD Design

To demonstrate why using TWFE in a staggered DiD is problematic, we first construct a simulation example by pulling data from Compustat. Compustat is a widely used data source to study operations management topics in recent years (Kim and Henderson 2015, Dong et al. 2020, Barker et al. 2022). Simulation based on Compustat data was also used in other field to illustrate the efficacy of TWFE DiD estimates (Baker et al. 2022). Therefore, we also use Compustat data to simulate a staggered design. We use SQL to pull necessary variables from the *funda* database (Fundamentals Annual) in Compustat from Wharton Research Data Services (WRDS). We then calculate annual ROA for U.S. firms in the Transportation and Warehouse sector (sector code 48-49 in North American Industry Classification System) from 1989 to 2019. We keep firms with at least 20 observations in our data. After data cleaning, we have 64 firms with 1649 observations in our data.

We utilize annual ROA of the U.S. Transportation and Warehousing Sector to simulate three artificial exogeneous shocks to mimic the four distinct stages observed in ROA from 1989 to 2019 (Figure 3). We observe four distinct stages of ROA in Figure 3: less than 0.1 before 1997; 0.1 – 0.2 between 1998 and 2009; 0.2 – 0.3 between 2010 and 2014; and trending at approximately at 0.3 after 2015. Firms were randomly assigned into three groups. The first, second, and the third group receive their respective artificial exogenous shock in 1997, 2009, and 2015 to mimic the four stages in Figure 3.

Figure 3 ROA of US Transportation and Warehousing Sector: 1989-2019



Following recent econometrics literature (Baker et al. 2022) as well to reflect the empirical OM research settings, we design a simulation that forces a flat *pre-trend* for all the three treated groups while allowing a respective annual increase of 8%, 5%, and 4% of the standard deviation of ROA *after* the treatment. In other words, this simulation has no pretrends (parallel and flat) for the treated groups. The data generating process of the first simulation is shown in the following.

$$G_{1997}, G_{2009}, G_{2015} \in \{0,1\}^3$$

$$P(G_{1997} = 1) = P(G_{2009}) = P(G_{2015}) = \frac{1}{3}$$

$$G_{1997} + G_{2009} + G_{2015} = 1$$

$$Y_{it}(G_\infty) \quad set\ from\ dataset$$

$$\delta_{1997} = 0.04 \cdot \hat{s}\big(Y_{it}(G_\infty)\big)$$

$$\delta_{2009} = 0.05 \cdot \hat{s}\big(Y_{it}(G_\infty)\big)$$

$$\delta_{2015} = 0.08 \cdot \hat{s}\big(Y_{it}(G_\infty)\big)$$

$$Y_{it}(G_{1997}) = Y_{it}(G_\infty) + \delta_{1997} \cdot (t + 1 - 1997) \text{ if } t \geq 1997$$

$$Y_{it}(G_{2009}) = Y_{it}(G_\infty) + \delta_{2009} \cdot (t + 1 - 2009) \text{ if } t \geq 2009$$

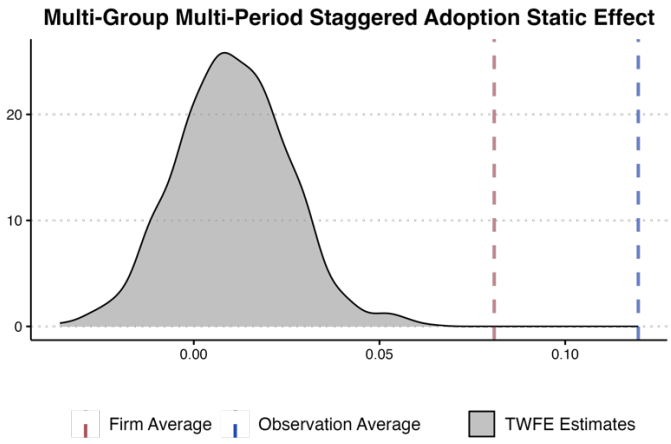
$$Y_{it}(G_{2015}) = Y_{it}(G_\infty) + \delta_{2015} \cdot (t + 1 - 2015) \text{ if } t \geq 2015$$

In this simulation, there are three treatment groups who receive artificial treatment in periods 1997, 2009 and 2015. Treatment is absorbing so the adoption is staggered. The respective treatment effect trajectory for the three groups is linear in the post-treatment periods, but each group has different coefficients. The three groups are randomly assigned and have the same probability mass, equal to 1/3. There is treatment effect heterogeneity across cohorts as the δ is different for each group.

We first use TWFE to estimate static treatment effect. We compute actual firm average treatment effect and actual observation average treatment effect following recent econometrics literature (Baker et al.

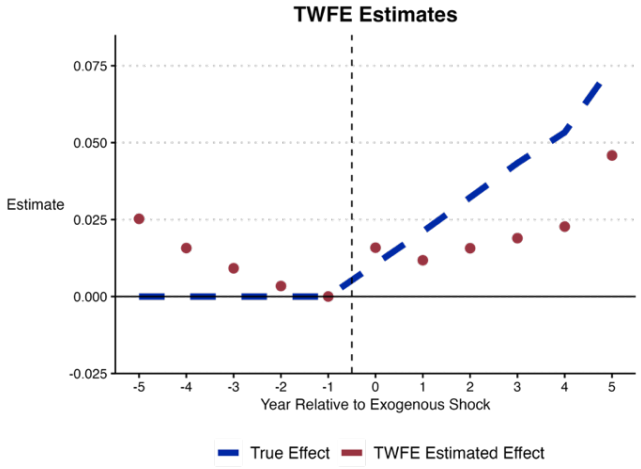
2022). For Firm average treatment effect, we first compute average ATT for each firm then average these ATTs. For observation average treatment effect, we just compute average ATT across all treatment observations. We plot the simulation results in Figure 4. In Figure 4, the red line represents actual firm average treatment effect, the blue line represents actual observation average treatment effect, and the bell-curve represents the distribution of the TWFE estimated treatment effects from 1000 simulations. We see that both actual firm average effect and actual observation average effect are far from the distribution of the estimated TWFE effects from 1000 simulations, indicating that TWFE estimator cannot accurately reproduce the true treatment effect for a staggered DiD design..

Figure 4 Muti-group Multi-Period Staggered Adoption Simulation



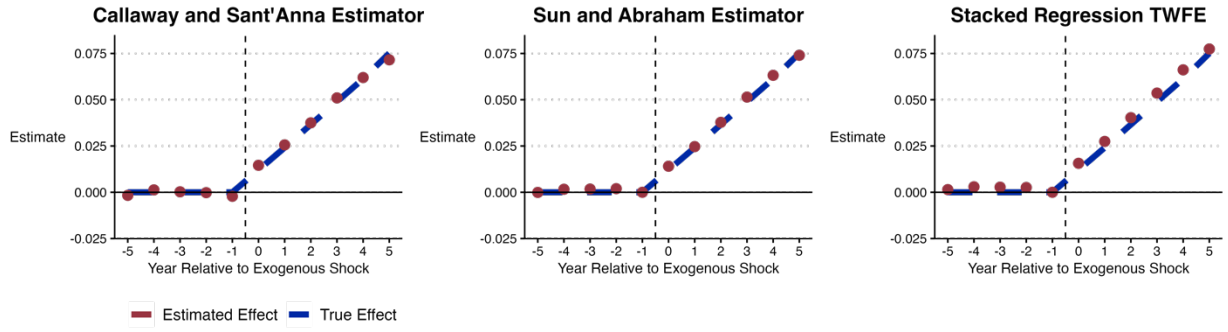
We next use TWFE event study method to estimate the coefficients of relative time indicators (5 years pre and 5 years post the treatment). We plot the result in Figure 5. Blue line represents the simulated trend and the red line represents TWFE estimates. Note that 27 studies we surveyed rely on using coefficients of the leads to test the parallel trend assumption and the coefficients of the lags to test the statistical significance of time indicators (and hence the corresponding research questions). However, we see from our simulation results that this practice is problematic: TWFE estimates a downward going pre-trend when there is actually no pretrend. In addition, for the treatment effect (relative time indicators after time 0), TWFE estimates an incorrect treatment trajectory, underpinning the fact that using TWFE event study to estimate treatment effect is also problematic.

Figure 5 Simulation of a Staggered DiD Design



We now use alternative estimators to run the same simulation. In a staggered DiD design, depending on the specific research setting, data, and the corresponding assumptions, there are three alternatives available (see Figure 2): Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Deshpande and Li (2019). Note that Deshpande and Li (2019) approach is not an estimator per se because Deshpande and Li (2019) also use TWFE estimator. The key step for Deshpande and Li (2019) is to construct relevant event-specific clean 2×2 datasets (i.e., avoid using already-treated units as effective controls). Then all these clean 2×2 datasets were stacked together to be estimated using TWFE. Detailed steps of how to construct the stacked data could be found in our online supplementary code. Figure 6 reports the estimated results. Blue line represents simulated trend and red dots represent estimated effect. We observe that all the three alternatives not only correctly estimate the pretrend but also correctly estimate the treatment effect.

Figure 6 Alternatives to Test for Parallel Trend Assumption



Note that in our simulation, all unit are subject to treatment sooner or later in the three years of 1999, 2007, and 2015. Therefore, there is no never-treated units in our simulation and only not-yet-treated units serve as clean control groups. Alternatively, we keep a portion of observations in our data as untreated (ranging from 10% up to 50%) such that both not-yet-treated and never-treated units can serve as clean

controls. We run the simulation again and our results are the same. In addition, we only use never-treated units as the effective control group and run the simulation again, our results are still the same. Note that our simulation is conducted for illustration purpose without considering which group is the best comparison group. Researchers, in practice, need to consider their data to select the best suitable group to serve as the control group to test the parallel trend assumption and the research question.

In sum, our simulation example shows that using TWFE to estimate the pretrend and the treatment effects (including static and dynamic) on a staggered DiD design leads to biased estimates. The three alternatives (Callaway and Sant'Anna 2021, Sun and Abraham 2021, Deshpande and Li 2019) can all produce accurate estimates. We call for researchers to start adopting any of these three alternatives to test the parallel trend assumption and test the coefficients on the post-treatment time indicators (hence the research question) to increase research rigor and stringency.

7.2 Replication of Wang et al. (2022)

Among the 51 studies we surveyed, 24 studies (47%) adopt a staggered TWFE DiD design and are susceptible to biased and misleading results. Unfortunately, none of the 24 studies provide any replication data and code, therefore, we instead use a recent Management Science publication (Wang et al. 2022) that comes with data to further demonstrate the pitfalls of TWFE in a staggered DiD design in addition to the simulation example. Note that we do not aim to critique the work of Wang et al. (2002). Instead, we hope to use this example to debunk why TWFE produces biased estimates step-by-step to better facilitate the understanding of this issue to help increase research rigor for the whole OM research community. Neither do we aim to produce a comprehensive replication study for all past quasi-experiment staggered DiD designs.

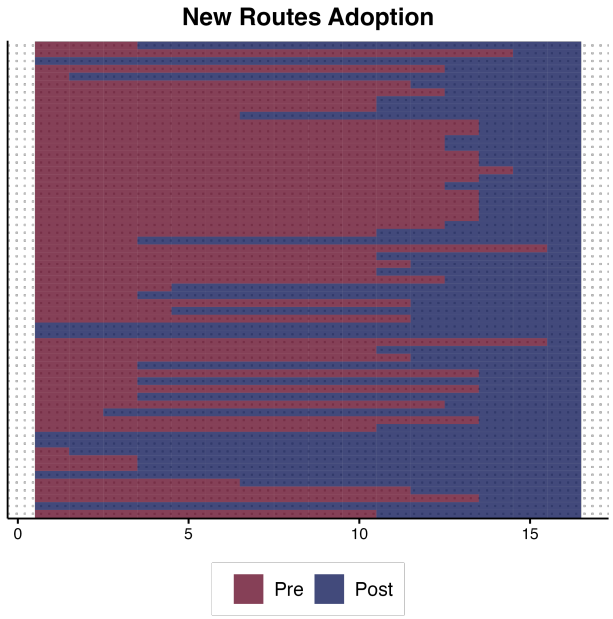
Wang et al. (2022) investigate how the introduction of new airline routes (defined as direct flights between two cities) impact the volume of shared kidney in the U.S. from 2002 to 2017. Adopting a TWFE DiD design, Wang et al. (2022) conclude that after the introduction of new airline routes, there has been a 7.5% increase in shared kidneys – the static effect. Further, using TWFE event study with an event window of 3 years before and 3 years after introducing new routes, Wang et al. (2022) also estimate a statistically significant increase in shared kidney volume in the event year itself (year 0) as well as in year three and beyond (time period after 3 years were binned in Wang et al. 2022). Table 3 summarizes the data in Wang et al. (2022). The data spans 16 years from 2002 to 2017. However, not every year witnesses introduction of new routes. Hence, Table 3 only shows those years when new routes were introduced. For example, in 2005, 128 new routes were introduced. Wang et al. (2022) also compile a control group of 513 existing routes throughout 2002 to 2017.

Table 3 Summary of New Routes in Wang et al. (2022)

Year	# of New Routes Introduced	# of Control Routes
2002	112	513
2003	32	513
2004	16	513
2005	128	513
2006	32	513
2008	32	513
2012	128	513
2013	96	513
2014	128	513
2015	208	513
2016	32	513
2017	32	513

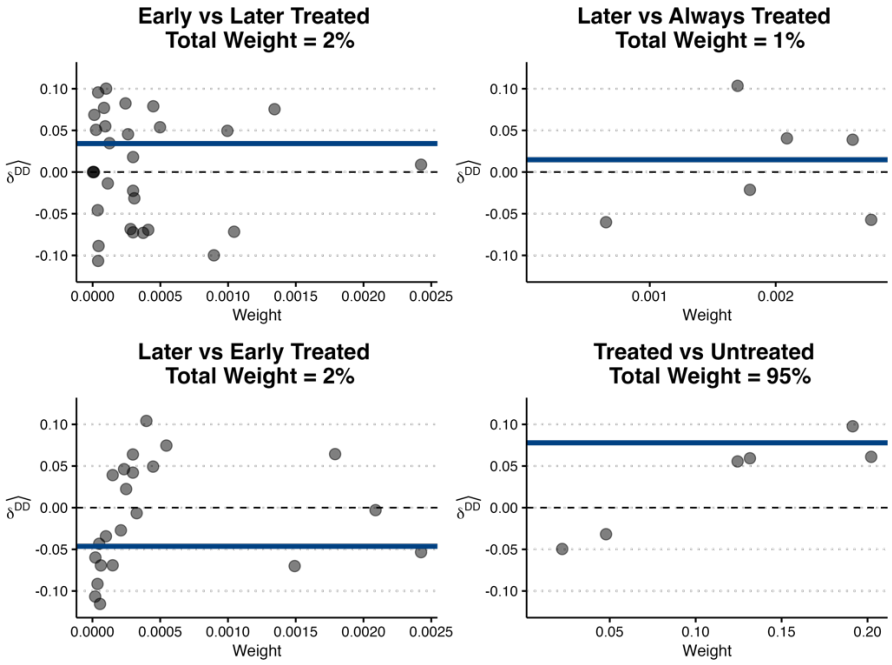
Table 3 shows a classic staggered DiD design as the introduction of new routes differs in time. The variation in treatment timing in a staggered DiD design indicates potential biases associated with TWFE DiD regression, as was discussed in Section 2. We attempt to debunk the potential causes of the biased estimates associated with TWFE DiD regression in a few steps. First, we plot the staggered timing of new routes introduction in Figure 7. Dark red tiles represent the time periods before any new routes were introduced while dark blue tiles represent the time periods after new routes were introduced. From Table 3 and Figure 7, we observe three potential issues when estimating the treatment effect using TWFE. First, for the 112 routes introduced in 2002, these 112 routes become "always-treated" units (as is shown by the all-blue tiles). Second, any new routes, if used as controls for newer routes introduced later, will become "already-treated" units. Always-treated and already-treated units could not serve as clean controls as changes on these routes already reflect the treatment effect. Third, for the 32 routes introduced in 2017, there will be no post-event time periods to estimate the treatment effect for these routes. All the three issues will generate potential biases for TWFE. However, there are also 513 never-treated routes serving as the control group throughout the years, which might mitigate the biases. To quantity the potential biases, we move on to the next analysis.

Figure 7 Bacon Decomposition of Treatment Groups in Wang et al. (2022)



Next, we decompose weights associated with TWFE using Goodman-Bacon (2021) decomposition method. Figure 8 compares all possible 2×2 DiD estimates by TWFE from pooled OLS regression in four categories: early-treated (as treatment) compared to later-treated (as controls); later-treated (as treatment) compared to always-treated (as controls); later-treated (as treatment) compared to early-treated (as controls); and treated (as treatment) compared to un-treated (as controls). Gray dot represents a comparison between treatment-timing cohort, for example, between those new routes introduced in 2004 and those routes introduced in 2015 and so on so forth. Bold blue line is the weighted average of all these comparisons for each group. "Total Weights" is the weights applied by TWFE to each group of comparison. The overall average treatment effect calculated by TWFE is the weighted sum of each weighted average across the four groups. Among the four groups, the potentially problematic 2×2 comparisons are the second group and the third group: those later-treated *vs* early-treated comparisons (1% weight) and later-treated *vs* always-treated comparisons (2% weight). Combined, the problematic 2×2 comparisons only contribute to 3% of TWFE weight. Will this small 3% problematic weight lead to any biased estimates? We answer this question by estimating the static and dynamic effects using Callaway and Sant'Anna (2021) estimator as an example.

Figure 8 Bacon Decomposition of Treatment Groups in Wang et al. (2022)



We first estimate the static effect of introducing new airline routes on the volume of shared kidneys using Callaway and Sant'Anna (2021) estimator. Wang et al. (2022) used an event window of –3 years to 3 years and binned remoted time periods beyond three years to assess treatment effect, i.e., all post-event time periods beyond three years were binned as three years. Unlike Wang et al. (2022), we did not bin remote time periods and only take three years prior and three years post as the event window. The reason is that binning remote time periods also produces biased results (Baker et al. 2022). The original results as well as the new results using Callaway and Sant'Anna (2021) are reported in Table 4. In Table 4, we replicate Model 1 from Wang et al. (2022). We also replicate other models in Wang et al. (2022) and our conclusion does not change. From Table 4, we see that Wang et al. (2022) reported a statistically significant effect, i.e., shared kidney volume increased by 7.3% after new routes were introduced. Callaway and Sant'Anna (2021) estimator also reports a similar effect size (8.4%). However, the estimate becomes statistically non-significant. Therefore, the 3% weight of the problematic comparisons discussed above, albeit small, indeed leads to biased estimate in TWFE, i.e., a statistically non-significant effect was estimated as statistically significant by TWFE, reinforcing the fact that in a staggered DiD design, TWFE produces biased estimates for the static effect. In Wang et al.'s (2022) case, the bias comes from the fact that already-treated and always-treated units were used as effective controls.

Table 4 Static Effect of New Airline Routes on Shared Kidney Volume

Dependent Variable: Volume of Shared Kidney (log-transformed)		
	Wang et al. 2022	Callaway and Sant'Anna 2021 Estimator
Static Effect	0.073** (0.030)	0.084 (0.060)
Year Fixed Effects	Yes	Yes
Airport-pair Fixed Effects	Yes	Yes

***p < 0.01; **p < 0.05; *p < 0.1. Standard Error in parenthesis.

We continue using Wang et al. (2022) to demonstrate the pitfalls of traditional TWFE event study method in a staggered DiD design. Wang et al. (2022) also estimate an event study for the effect of new airline routes on the volume of shared kidney using an event window of –3 to 3 years (years beyond 3 were binned as 3). Different from Wang et al. (2022), we do not bin distant time periods following the advice of Sun and Abraham (2021). We estimate a fully dynamic model but we only report the first 3 post-treatment periods to match the results from Wang et al. (2022). The original results from Wang et al. (2022) and the estimates from using Callaway and Sant'Anna (2021) are reported in Table 5. We see that despite a similar effect size, none of the coefficients is statistically significant when using Callaway and Sant'Anna (2021) estimator, in contrast to the original result of Wang et al. (2022) using TWFE.

Table 5 Dynamic Effects of New Airline Routes on Shared Kidney Volume

Dependent Variable: Volume of Shared Kidney (log-transformed)		
	Wang et al. 2022	Callaway and Sant'Anna 2021 Estimator
-3	0.013 (0.044)	0.013(0.052)
-2	-0.021 (0.041)	-0.033(0.045)
-1	0.018 (0.040)	0.044(0.060)
0	**0.089*** **(**0.053)	0.043(0.063)
1	0.050 (0.041)	0.038(0.053)
2	0.050 (0.046)	0.064(0.044)
3	**0.095***** (0.035)	0.102(0.060)
Year Fixed Effects	Yes	Yes
Airport-pair Fixed Effects	Yes	Yes

***p < 0.01; **p < 0.05; *p < 0.1. Standard Error in parenthesis.

We also present an event study plot in Figure 9 using the event window of –3 to 3 years as used in Wang et al. (2022). Event study plot, as in Figure 9, represents the most common approaches to aggregate treatment effect in OM studies – the average treatment effect at different lengths of exposure to the treatment. We clearly see that based on this popular event study plot and the results in Table 5, none of the treatment effect is statistically significant. However, one of the features of Callaway and Sant'Anna (2021) estimator is that this estimator can report treatment effects at different levels of aggregation (Section 5.1). Therefore, we compute average treatment effects specific to each group. In Wang et al.'s (2022) case, "group" refers to the group of new routes that were introduced in the same year. This effect says for those years when new airline routes were introduced, what is the effect specific to that year/group?

We report group-specific treatment effect in Figure 10. We see that among the nine groups/years when new routes were introduced, group 3 (year 2004) and group 7 (year 2008) show positive and statistically significant estimates. In other words, when using traditional event study to aggregate treatment effects as measured by length of exposure to treatment, none of the estimates is statistically significant. However, when aggregating the treatment effect into group-specific treatment effects, two of the estimates are statistically significant, i.e., in both 2004 and 2008, the shared volume of kidneys indeed increase after introduction of new airline routes. But why only these two years? This could possibly lead to other interesting research topics for the related stream of literature. Therefore, we see that using Callaway and Sant'Anna (2021) estimator not only produces different statistical findings from the original TWFE results in Wang et al. (2022), it also provides more flexible aggregation of treatment effects that could potentially answer more nuanced research questions. We also replicate static and dynamic effect using stacked regression (Deshpande and Li 2019) and the conclusion remains the same. Interested readers can refer to the Online Supplement and its associated code to see how stacked regression was built and tested for this replication study.

Figure 9 Dynamic Effect of New Airline Routes on Volume of Shared Kidney

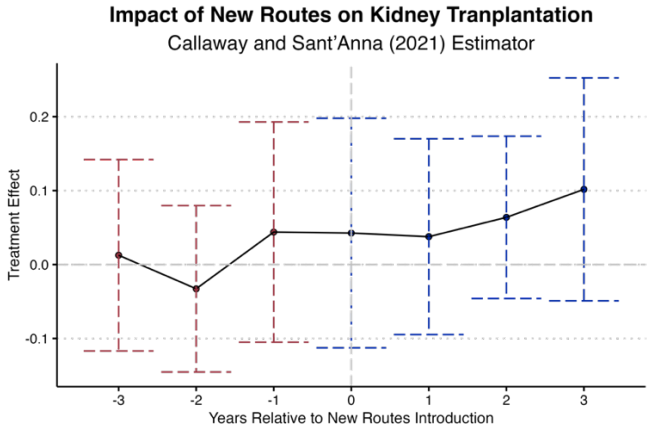
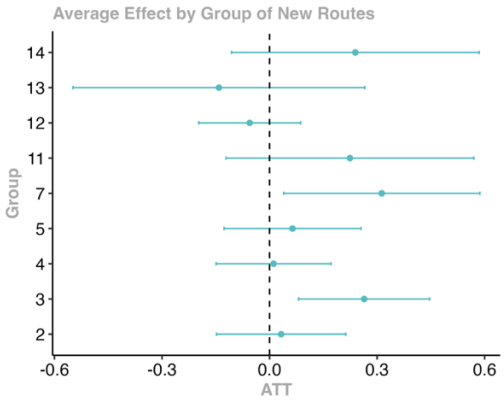


Figure 10 Group-specific Average Treatment Effects of New Airline Routes on Shared Kidney Volume



8. RECOMMENDATION FOR BEST PRACTICES AND FUTURE RESEARCH

8.1 Recommendation for Best Practices

In a staggered DiD design, we recommend researchers to conduct the following. First, researchers can decompose the weights assigned by TWFE to different 2×2 groups using Goodman-Bacon (2021) decomposition method (data needs to be a balanced panel). Then, researchers can detect the problematic weight where the always-treated and already-treated units are used as effective controls. The higher the weight associated with the problematic 2×2 comparisons, the more biased the TWFE estimates. If the data is unbalanced, researcher can create a balanced subset (if not losing too much information of the original data) and run the decomposition test to gauge the severity of the problem. This practice also helps researchers to determine which group should be used as the comparison group to test for the parallel trend assumption as well as estimating treatment effect following the process in our flowchart in Figure 2.

Second, which estimator to use? We highly recommend researchers to follow the process in Figure 2 to select an appropriate estimator to answer the corresponding research question. If researchers are interested to estimate static and/or dynamic effect and if the treatment is an absorbing state, any of the three alternatives (Deshpande and Li 2019, Callaway and Sant'Anna 2021, Sun and Abraham 2021) is available. Researchers, however, need to determine which estimator to use based on their understanding of the data as well as drawing on their domain expertise.

Third, when the treatment effect is not an absorbing state, i.e., treatment effect switches on and off, only De Chaisemartin and d'Haultfoeuille (2020) estimator is available. De Chaisemartin and d'Haultfoeuille (2020) estimator also stands out for its ability to accommodate a fuzzy DiD design whereas none of the other three estimators is able to do this. In addition, if researcher are interest to report instantaneous effect, all the four estimators are available but the choices again differs largely on if the treatment is in an absorbing state or not.

Fourth, for the parallel trend assumption test, researchers need to answer three questions based on their understanding of the data: 1) is the treatment in an absorbing state or not? 2) is the parallel trend assumption unconditional or conditional? 3) which group (never-treated, last-treated, not-yet-treated/eventually treated) serves as the best comparison group? After these three questions are answered, researchers can use either of the four alternatives (De Chaisemartin and d'Haultfoeuille 2020, Deshpande and Li 2019, Callaway and Sant'Anna 2021, Sun and Abraham 2021) to test this assumption. However, as repeatedly discussed in this paper, De Chaisemartin and d'Haultfoeuille (2020) estimator is more geared towards instantaneous treatment effect where treatment switches on and off. De Chaisemartin and d'Haultfoeuille (2020) indeed provide implementation code that researchers can easily modify to test the parallel trend assumption and the treatment effect. For a strict staggered design, the other three alternatives perform equally well in testing the parallel trend assumption (Section 7.1). However, Callaway and

Sant'Anna (2021) and Sun and Abraham (2021) only require researchers to create an additional variable of treatment identifier (the first occasion when different groups were treated) without the need of reconstructing the original data. In contrast, Deshpande and Li (2019) requires researchers to create event-specific clean 2×2 datasets (i.e., already-treated units should not be used as effective controls). These clean 2×2 datasets need to be stacked together to be estimated using TWFE. Therefore, stacked regression necessitates extra amount of data reconstruction work for researchers.

In addition to the biased estimates associated with TWFE, we also notice another common issue with DiD analysis in operations management: when presenting DiD results, researchers do not report the event window for the treatment effect. Only 5 out of the 51 studies clearly report event window when presenting results. For example, Chun et al. (2022) studied how former kidney-transplant patients' mentoring on current kidney-transplant patients impacts the anxiety of the current patients. Chun et al. (2022) concluded that treatment group experienced on average 3.42 points decrease in anxiety score in a 30 day period after former patients started to mentor current patients, which provides a clear picture of the event window for the treatment effect. Differently, Ren et al. (2023) studied the impact of sharing retail store product availability on online sales. Ren et al. (2023) concluded that after implementing the policy of sharing retail store product availability to consumers, the online sales increased 13.1% within a 50 miles radius of retail stores. This conclusion is one of the key findings but is ambiguous: does the 13.1% increase of online sales happen immediately after the policy implementation or does the 13.1% increase happen over time? If the increase happened over time, does it happen within a month, 12 months, or 24 months, and etc.? Without a comprehensive reading of the entire section of data and analysis, readers will not be able to tell that the 13.1% increase is the averaged increase in 27 months after policy implementation. Unclear reporting of event window when presenting results creates challenges for both academia and practitioners. To avoid confusion as well as to increase stringency when presenting research findings, we encourage researchers to clearly state the event window for the treatment effects. This also matters greatly for practitioners and policymakers as merely presenting a mere number of treatment effect without mentioning the event window does not provide much guidance for policy making.

Lastly, we want to remind researchers that propensity score matching (PSM), coarsened exact matching (CEM), or synthetic control method, even if well executed in a TWFE context, still cannot solve the biases associated with TWFE in a staggered DiD design as these matching methods are used to address the issue of sample selection (Ho et al. 2017, Shang and Rönkkö 2022), not the negative weight issue and contamination issue associated with TWFE estimator itself as highlighted by Sun and Abraham (2021).

8.2 Other Causes of Biases and Future Research

The econometrics literature keeps evolving and recently it has made progress that have enriched the TWFE DiD in multiple directions. For example, as studied in Callaway et al. (2024), when treatment is continuous,

the parallel trend assumption now has to hold for all treatment groups, which is considerably a more stringent requirement. If the value of the covariates is influenced by the treatment itself, as studied in Caetano et al. (2022), then TWFE may be again biased. Caetano et al. (2022) propose a set procedure based on regression adjustment, imputation and double robust procedure to circumvent the issue. Deaner and Ku (2024) consider a form of the parallel trend assumption where the outcome is a duration and highlight scenarios when it is likely to be violated. Along these lines, an overall discussion of the conditions under which the parallel trend assumption is robust to changes in the functional form appears in Roth and Sant'Anna (2023). Lastly, Wooldridge (2023) derives simple yet flexible strategies to accommodate the nonlinear structure in a DiD design. We hope interested researchers can continue synthesizing the continuous development from econometrics field on this topic to further enhance research rigor in the OM research community.

REFERENCES

Abadie A (2005) Semiparametric difference-in-differences estimators. *Rev. Econ. Studies*. *72*(1):1-19.

Agarwal S, Mani D, Telang R (2023) The impact of ride-hailing services on congestion: Evidence from Indian cities. *Manufacturing Service Oper. Management* 25(3):862-883.

Akturk MS, Ketzenberg M (2022) Impact of competitor store closures on a major retailer. *Production Oper. Management* 31(2):715-730.

Baker AC, Larcker DF, Wang CC (2022) How much should we trust staggered difference-in-differences estimates? *J. Financial Econom.* 144(2):370-395.

Barker JM, Hofer C, Hoberg K, Eroglu C (2022) Supplier inventory leanness and financial performance. *J. Oper. Management 68*(4):385-407.

Barrios JM, Hochberg YV, Yi H (2022) The cost of convenience: Ridehailing and traffic fatalities. *J. Oper. Management* 69(5):1–33.

Borusyak K, Jaravel X, Spiess, J (2024) Revisiting event-study designs: robust and efficient estimation. *Rev. Econ. Studies*. Print in Advance.

Caetano C, Callaway B, Payne S, Rodrigues HAS (2022) Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903*.

Callaway B, Goodman-Bacon A, Sant'Anna PH (2024) *Difference-in-differences with a continuous treatment* (No. w32117). National Bureau of Economic Research.

Callaway B, Sant'Anna PH (2021) Difference-in-differences with multiple time periods. *J. Econom.*. 225(2):200-230.

Calvo E, Cui R, Wagner L (2023) Disclosing product availability in online retail. *Manufacturing Service Oper. Management 25*(2):427-447.

Chan TH, Bharadwaj A, Varadarajan D (2023) Business Method Innovation in US Manufacturing and Trade. *Manufacturing Service Oper. Management* 25(1):50-69.

Chen J, Xu Y, Yu P, Zhang J (2023) A reinforcement learning approach for hotel revenue management with evidence from field experiments. *J. Oper. Management*: 1-26.

Chun Y, Harris SL., Chandrasekaran A, Hill K (2022) Improving care transitions with standardized peer mentoring: Evidence from intervention based research using randomized control trial. *J. Oper. Management* 68(2):185-214.

Cui R, Ding H, Zhu F (2022) Gender inequality in research productivity during the COVID-19 pandemic. *Manufacturing Service Oper. Management* 24(2):707-726.

Cui R, Zhang DJ, Bassamboo A (2019) Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Sci.* 65(3):1216-1235.

Davis AM, Flicker B, Hyndman KB, Katok E, Keppler S, Leider S, Long X, Tong J (2023) A Replication Study of Operations Management Experiments in Management Science. *Management Sci.* Published online July 11, 2023.

De Chaisemartin C, D'Haultfoeuille X (2020) Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* 110(9):2964-96.

Deaner B, Ku H (2024) Causal Duration Analysis with Diff-in-Diff. *arXiv preprint arXiv:2405.05220*.

Deshpande, M. and Li, Y., 2019. Who is screened out? Application costs and the targeting of disability programs. *Am. Econ. J.: Econ. Policy* 11(4), pp.213-248.

Dong Y, Chung M, Zhou C, Venkataraman S (2019) Banking on "Mobile Money": The Implications of Mobile Money Services on the Value Chain. *Manufacturing Service Oper. Management* 21(2):290-307.

Dong Y, Skowronski K, Song S, Venkataraman S, Zou F (2020) Supply base innovation and firm financial performance. *J. Oper. Management* 66(7-8):768-796.

Ergin E, Gümüş M, Yang N (2022) An Empirical Analysis of Intra-Firm Product Substitutability in Fashion Retailing. *Production Oper. Management* 31(2):607-621.

Fan D, Zhou Y, Yeung AC, Lo CK and Tang C (2022) Impact of the US–China trade war on the operating performance of US firms: The role of outsourcing and supply base complexity. *J. Oper. Management* 68(8).928-962.

Gao Y, Li M, Sun S (2023) Field experiments in operations management. *J. Oper. Management* 69(4).676-701.

Ge C, Huang H, Wang Z, Jiang J, Liu C (2023) Working from home and firm resilience to the COVID-19 pandemic. *J. Oper. Management* 69(3):450-476.

Gong J, Greenwood BN, Song Y (2023) An empirical investigation of ridesharing and new vehicle purchase. *Manufacturing Service Oper. Management* 25(3):884-902.

Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. *J. Econ.* 225(2):254-277.

Han, BR, Sun T, Chu LY, Wu, L (2022) COVID-19 and E-commerce Operations: Evidence from Alibaba. *Manufacturing Service Oper. Management* 24(3):1388-1405.

Ho TH, Lim N, Reza S, Xia X (2017). OM forum—Causal inference models in operations management. *Manufacturing & Service Operations Management* 19:509–525.

Jacobs BW, Singhal VR, Zhan X (2022) Stock market reaction to global supply chain disruptions from the 2018 US government ban on ZTE. *J. Oper. Management* 68(8):903-927.

KC D, Kim T (2022) Impact of universal healthcare on patient choice and quality of care. *Production Oper. Management* 31(5):2167-2184.

Klöckner M, Schmidt CG, Wagner SM (2022) When blockchain creates shareholder value: empirical evidence from international firm announcements. *Production Oper. Management* 31(1):46-64.

Kokkodis M, Lappas T, Kane GC (2022) Optional purchase verification in e-commerce platforms: More representative product ratings and higher quality reviews. *Production Oper. Management* 31(7):2943-2961.

Lam HK, Ding L, Dong Z (2022) The impact of foreign competition on domestic firms' product quality: Evidence from a quasi-natural experiment in the United States. *J. Oper. Management* 68(8):881-902.

Lee HS, Kesavan S, Kuhnen C (2022) When do group incentives for retail store managers work? *Production Oper. Management* 31(8):3077-3095.

Li J, Wu D (2020) Do corporate social responsibility engagements lead to real environmental, social, and governance impact? *Management Sci.* 66(6):2564-2588.

Li Y, Lu LX., Lu SF, Chen J (2022) The value of health information technology interoperability: Evidence from interhospital transfer of heart attack patients. *Manufacturing Service Oper. Management* 24(2):827-845.

Li Z, Liang C, Hong Y, Zhang Z (2022) How do on-demand ridesharing services affect traffic congestion? The moderating role of urban compactness. *Production Oper. Management* 31(1):239-258.

Lo CK, Tang CS, Zhou Y (2022) Do polluting firms suffer long term? Can government use data-driven inspection policies to catch polluters? *Production Oper. Management* 31(12):4351-4363.

Miao W, Deng Y, Wang W, Liu Y, Tang CS (2022) The effects of surge pricing on driver behavior in the ride-sharing market: Evidence from a quasi-experiment. *J. Oper. Management* 69(5):1-29.

Mithas S, Chen Y, Lin Y, De Oliveira Silveira A (2022) On the causality and plausibility of treatment effects in operations management research. *Production Oper. Management* 31(12):4558-4571.

Pan Y, Qiu L (2022) How ride-sharing is shaping public transit system: A counterfactual estimator approach. *Production Oper. Management* 31(3):906-927.

Qiu L, Kumar S, Sen A, Sinha AP (2022) Impact of the Hospital Readmission Reduction Program on hospital readmission and mortality: An economic analysis. *Production Oper. Management* 31(5):2341-2360.

Rambachan A, Roth J (2023) A more credible approach to parallel trends. *Rev. Econo Studies 90*(5):2555-2591.

Ren X, Windle RJ, Evers PT (2023) Channel transparency and omnichannel retailing: The impact of sharing retail store product availability information. *J. Oper. Management* 69(2):217-245.

Roth J (2022) Pretest with caution: Event-study estimates after testing for parallel trends. American Econ. Rev.: Insights, 4(3):305-322.

Sant'Anna PH, Zhao J (2020) Doubly robust difference-in-differences estimators. *J. Eonom*. 219(1):101-122.

Scott A, Li M, Cantor DE, Corsi TM (2023) Do voluntary environmental programs matter? Evidence from the EPA SmartWay program. *J. Oper. Management* 69(2):284-304.

Shang G, Rönkkö (2022) Empirical research methods department: Mission, learnings, and future plans. *J. Oper. Management* 68(2):114-129.

Simchi-Levi D. (2022) From the Editor. *Management Sci.* 68(1):1-6.

Song S, Dong Y, Kull T, Carter C, Xu K (2023) Supply chain leakage of greenhouse gas emissions and supplier innovation. *Production Oper. Management* 32(3):882-903.

Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.* 225(2):175-199.

Wang G (2022) Stay at home to stay safe: Effectiveness of stay-at-home orders in containing the COVID-19 pandemic. *Production Oper. Management* 31(5):2289-2305.

Wang G, Zheng R, Dai T (2022) Does transportation mean transplantation? Impact of new airline routes on sharing of cadaveric kidneys. *Management Sci. 68*(5):3660-3679.

Wang H, Overby EM (2022) How Does Online Lending Influence Bankruptcy Filings? *Management Sci.* 68(5):3309-3329.

Wang L, Rabinovich E, Guda H (2023) An analysis of operating efficiency and policy implications in last-mile transportation following Amazon's integration. *J. Oper. Management* 69(1):9-35.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Wooldridge JM (2023) Simple approaches to nonlinear difference-in-differences with panel data. *Econom. J. 26*(3):C31-C66.

Zhou Z, Wan X (2022) Does the sharing economy technology disrupt incumbents? Exploring the influences of mobile digital freight matching platforms on road freight logistics firms. *Production Oper. Management* 31(1):117-137.

APPENDICES

Appendix 1 Studies in Top OM Journals using TWFE DiD Regressions: January 2022 to May 2023 – Breakdown

No.	Journal	Author	Susceptible to Biased Estimates	Static Effect	Dynamic Effect	Variation in Treatment Timing	Include Never-Treated Units	Test for Parallel
1	JOM	Barrios et al. 2022	Y	Y	Y	Y	Y	Coefficient of leads and plots
2	JOM	Chen et al. 2023		Y		N	Y	
3	JOM	Chen L. et al. 2023		Y	Y	N	N	
4	JOM	Chun et al. 2022	Y	Y		Y	Y	
5	JOM	Fan et al. 2022		Y		N	Y	Coefficient of leads and plots
6	JOM	Ge et al. 2023	Y	Y		Y	N	Coefficient of leads and plots
7	JOM	Jacobs et al. 2022			Y	N	N	
8	JOM	Lam et al. 2022	Y	Y		Y	Y	Coefficient of leads and plots
9	JOM	Li et al. 2022	Y	Y		Y	Y	
10	JOM	Miao et al. 2022		Y		N	N	Coefficient of leads
11	JOM	Ren et al. 2023		Y		N	Y	Coefficient treat*trend
12	JOM	Scott et al. 2023		Y		N	Y	Coefficient of leads and plots
13	JOM	Wang et al. 2023	Y	Y	Y	Y	N	Coefficient of leads
14	POM	Akturk and Ketzenberg 2022		Y		N	Y	Coefficient treat*trend
15	POM	Ba et al. 2022	Y	Y		Y	Y	
16	POM	Cheng et al. 2023	Y	Y		Y	N	Coefficient of leads
17	POM	Ergin et al. 2022	Y		Y	Y	Y	Coefficient of leads
18	POM	Hu et al. 2023	Y	Y		Y	Y	N (post hoc analysis)
19	POM	KC and Kim 2022		Y		N	Y	Coefficient of leads
20	POM	KC et al. 2022	Y	Y		Y	Y	Coefficient of leads
21	POM	Klöckner et al. 2022			Y			
22	POM	Kokkodis et al. 2022	Y	Y		Y	N	checking plots
23	POM	Lee et al. 2022		Y		N	Y	Coefficient treat*trend
24	POM	Li et al. 2022	Y	Y		Y	N	Coefficient of leads
25	POM	Lo et al. 2022			Y			
26	POM	Pan and Qiu 2022						Coefficient of leads
27	POM	Pu 2022		Y		N	Y	Coefficient of leads
28	POM	Qiu et al. 2022a	Y		Y	Y	Y	Coefficient of leads
29	POM	Qiu et al. 2022b		Y		N	N	
30	POM	Song et al. 2023	Y	Y		Y	Y	
31	POM	Wang 2022	Y	Y	Y	Y	Y	Coefficient of leads and plots
32	POM	Wang et al. 2023		Y		N	Y	checking plots
33	POM	Zhou and Wan 2022		Y	Y	N	Y	Coefficient of leads and plots
34	MSOM	Agarwal et al. 2023		Y	Y	N	Y	
35	MSOM	Calvo et al. 2023	Y	Y	Y	Y	Y	Coefficient of leads
36	MSOM	Cao et al. 2022		Y		N	Y	
37	MSOM	Caro et al. 2023		Y		N	Y	

38	MSOM	Chan et al. 2023		Y		N	Y	Coefficient of leads
39	MSOM	Cui et al. 2022		Y		N	N	Coefficient of leads
40	MSOM	Gong et al. 2023	Y		Y	Y	Y	Coefficient of leads and plots
41	MSOM	Gopalakrishnan et al. 2023		Y		N	Y	Coefficient of leads
42	MSOM	Han et al. 2022		Y	Y	N	Y	checking plots
43	MSOM	Hwang et al. 2022	Y	Y	Y	Y	Y	Coefficient of leads and plots
44	MSOM	Jain and Tan 2022		Y	Y	N	N	Coefficient of leads and plots
45	MSOM	Jeong and Lee 2022	Y	Y	Y	Y	Y	
46	MSOM	Li et al. 2022	Y	Y		Y	Y	
47	MSOM	Li et al. 2023	Y	Y		Y	Y	Coefficient treat*trend
48	MSOM	Schmidt and Raman 2022		Y	Y	N	Y	Coefficient of leads and plots
49	MSOM	Wan 2022		Y		N	Y	Coefficient of leads and plots
50	MSOM	Wang 2022	Y	Y	Y	Y	Y	Coefficient of leads and plots
51	MSOM	Wang et al. 2023	Y	Y		Y	Y	Checking plots

Appendix 2 Studies in Top OM Journals using TWFE DiD Regressions: January 2022 to May 2023 – Summary

	2022 (Jan - Dec)	2023 (Jan -May)	Total
JOM			
Total Empirical Articles	19	30	49
DiD Articles	5	8	13
% DiD Articles	26%	27%	27%
POM			
Total Empirical Articles	16	67	83
DiD Articles	4	16	20
% DiD Articles	25%	24%	24%
MSOM			
Total Empirical Articles	18	31	49
DiD Articles	7	11	18
% DiD Articles	39%	35%	37%
Grand Total			
Grand Total Empirical Articles	53	128	181
DiD Articles	16	35	51
% DiD Articles	30%	27%	28%