JOURNAL OF OPERATIONS MANAGEMENT
ASCM ASSOCIATION FOR SUPPLY CHAIN MANAGEMENT

# Assessing Treatment Effect In Quasi-Experiment Design In Operations Management

| | |
|---|---|
| Journal: | *Journal of Operations Management* |
| Manuscript ID | Draft |
| Wiley - Manuscript type: | Research Article |
| Topics: | Research methodology, Best practices |
| Methods: | Econometrics, Event Study, Experiment - quasi- |
| Additional Keywords: | simulation |
| Abstract: | Two-way-fixed-effects (TWFE) regression using Difference-in-Difference (DiD) is a workhorse to assess treatment effect across research fields for quasi-experiment design, including operations management. However, latest advances in econometrics field prove that this approach is prone to yield biased estimates and misleading inferences introduced by treatment effect heterogeneity when there are multiple time periods and when the treatment timing is different. Given 47% of DiD studies we surveyed in the top three operations management journals are susceptible to biased results, we systematically review five different DiD designs, explain the statistical foundations of each design, theoretically highlight the validity and/or pitfalls for each design, empirically illustrate why each design is valid and/or invalid, and further demonstrate how to use the latest econometrics estimators to avoid the bias associated with TWFE. In addition, we show why the common practice of testing the parallel trend assumption using event study is problematic and accordingly recommend alternatives that can correctly test this assumption. Our study serves as a reference for researchers to increase the validity of quasi-experiment DiD design. |
| | |

# Assessing Treatment Effect In 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 across research fields for quasi-experiment design, including operations management. However, latest advances in econometrics field prove that this approach is prone to yield biased estimates and misleading inferences introduced by treatment effect heterogeneity when there are multiple time periods and when the treatment timing is different. Given 47% of DiD studies we surveyed in the top three operations management journals are susceptible to biased results, we systematically review five different DiD designs, explain the statistical foundations of each design, theoretically highlight the validity and/or pitfalls for each design, empirically illustrate why each design is valid and/or invalid, and further demonstrate how to use the latest econometrics estimators to avoid the bias associated with TWFE. In addition, we show why the common practice of testing the parallel trend assumption using event study is problematic and accordingly recommend alternatives that can correctly test this assumption. Our study serves as a reference for researchers to increase the validity of quasi-experiment DiD design.



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

# ASSESSING TREATMENT EFFECT IN 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 incredible amount of attention from researchers in various operations management (OM) field, such as buyer supplier relationship (Fan et al. 2022), Covid-19 impact (Cui et al. 2023, 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 literature, the exogenous shock itself 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 research on exogenous shocks is mostly post hoc analysis after a police change, a quasi-experiment rather than a truly randomized experiment is utilized in 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 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 are *susceptible* to biased and misleading results (Appendix 1). In the recent wake of conducting responsible operations management research (MSOM 2020 Issue 6, POM upcoming special issue), we accordingly call for responsible research in drawing causal inference for treatment effect in the OM field. Despite scattered emphasis on the importance of correctly assessing treatment effect (Shang and Rönkkö 2022, Mithas et al. 2022, Barrios et al. 2022), there is no systematic review and in-depth analysis in the OM field on why some DiD design suffer from biased estimates and how to adopt the appropriate estimator to avoid the bias. Current study aims to fill this void for the OM research community – hence our first and main research motivation. 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 broadcast and educate practitioners with misleading research findings might result in catastrophic business consequences, harming both the stringency of academic research and the practicality of business practices.

The majority of treatment effect research in the OM filed investigates the average treatment effect on the treated (ATT) during a certain period of time after the treatment while some research 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 (Callaway and Sant'Anna 2021, 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). However, the instantaneous effect of the treatment has been largely ignored. Instantaneous effect refers to the immediate effect in a time window following the treatment, such as the impact of Covid-19 during the first day/week/month of its outbreak. To better combat future pandemics by drawing learnings from Covid-19, merely presenting an average effect from the past (such as the total death tolls in the two years after the breakout of Covid-19) will not serve as an effective policy guide as most countries suffered from the collapse of public health systems and high death tolls right after the first wave (first treatment) and immediately after the second wave (second treatment) (WHO 2023), the times of which required maximum healthcare resources. Neither static effect nor dynamic effect can provide the required insight for future healthcare policymakers to design immediate action plans as static effect averages the treatment effect while dynamic effect normally omits the event time window to avoid multicollinearity in analysis. Therefore, our second research motivation is to remind researchers the importance of instantaneous effect of exogenous shocks and introduce the latest econometric method to correctly estimate the instantaneous effect.

To achieve our two research motivations, we first categorize the different DiD designs adopted in the OM field 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 identified a total of 181 empirical studies, 51 studies (28%) of which utilized TWFE DiD and/or event study to draw causal inference (Appendix 2) – approximately 1 out of 3 empirical articles uses TWFE DiD estimator. Based on the 51 articles surveyed as well as drawing from econometrics literature, we classify DiD design into five different categories. The first design is the 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 multiple group and multiple time period setting where the treatment timing is the *same* for the treatment groups. The last two designs also access static and dynamic effect respectively in a multiple group and multiple time period setting. However, the treatment timing is *different* for the treatment groups – also known as staggered DiD design in the extant literature (Wooldridge 2010).

We then review the econometrics basics for each design, theoretically explain the validity and/or pitfalls of TWFE of each design, empirically demonstrate the validity and/or pitfalls of TWFE using simulations and published studies, and propose new estimators to correctly estimate the treatment effect. TWFE

estimator is used in 49 out of the 51 studies to access the treatment effect (Appendix 1). Recent development in econometrics reveals that TWFE estimator produces biased and misleading results when the treatment timing is staggered (Sun and Abraham 2021, Callaway and Sant'Anna 2021). A scrutiny of the 51 DiD articles reveals that 24 articles (47%) adopted a staggered design and used TWFE to estimate the treatment effect. Therefore, in the top three OM journals, approximately 1 out of 2 DiD articles is *susceptible* to biased and misleading results. We then replicate two pertinent published studies, debunk why these studies are prone to biased results, and accordingly show that new estimator yields fundamentally different results from TWFE results published in these studies. Publishing biased TWFE results not only deviates from the highest research standard in these top journals but also leads to potential damage to the related industries if policymakers rely on these published results to design their corresponding policies. To this end, we reiterate the importance of correctly assessing treatment effect to draw statistical inference in the OM field.

The validity of all the five DiD designs rely on 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 event study methodology to estimate the coefficients on the pre-treatment periods. In the 51 studies we surveyed, 35 studies (69%) tested for the parallel trend assumption and 27 of them checked for the pre-treatment coefficients to test for the parallel pretrend. However, using pre-treatment coefficients to test for the parallel pretrend can lead to significant distortions in causal inference in a staggered design (Roth 2020, Callaway and Sant'Anna 2021, de Chaisemartin and D'Haultfoeuille 2020). We simulate a staggered design using Compustat data and show that event study methodology fails to estimate any pretrend when there is actually a true pretrend in the data; while there is no true pretend in the data, event study estimates a pretend, resulting in both Type-I and Type-II errors. We accordingly propose several alternatives to correctly test the parallel trend assumption and empirically show that the alternatives can correctly estimate the pretrends in our simulated data.

We make four 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 part of quasi-experiment and remind researchers of the importance to correctly draw statistical inference from quasi-experiment to increase research validity. Second, despite DiD being 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 correct DiD estimator remain unanswered. Our study fills this void by systematically reviewing the five different DiD designs and provide a framework for researchers to increase research stringency for DiD designs. Third, using simulations and published DiD studies as examples, coupled by the new theoretical development in DiD estimator, we demonstrate that TWFE estimator produces biased results

when there is variation in treatment timing. We call for researchers to use the latest estimators to access treatment effect in a staggered DiD design as 47% DiD research we surveyed in the OM field is *susceptible* to biased and misleading results. In addition, we empirically show that the widespread practice of using event study methodology to test the parallel trend assumption can result in both Type-I and Type-II errors as the coefficient on the leads (i.e., coefficient of time indicators prior to the treatment) is contaminated by the estimates from other time periods and by the treatment heterogeneity (Sun and Abraham 2021). We accordingly propose alternative estimators that can correctly estimate the pretrend. Researchers can adopt the alternatives to test the parallel trend assumption as this assumption is paramount to the validity of all DiD designs. Fourth, we highlight the importance of assessing instantaneous effect, as is evidenced by the recent Covid-19, and encourage researchers to assess instantaneous effect in addition to static and dynamic effect to make more meaningful and useful policy recommendations. In sum, our study contributes to conducting more stringent research in operations management when drawing causal inferences for treatment effect, echoing the recent call for responsible research in the OM field.

There are two important notes in our study we want to share. First, the current study does not aim to critique or negate any of the 51 studies surveyed as methodology always evolves. Instead, our sole purpose is to remind researchers of the deficiencies (at the time of writing) associated with TWFE in a quasi-experiment 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 two publications to facilitate the understanding of the "why" and "how". However, we do not aim to conduct an exhaustive replication study of all quasi-experiment publications in the past to document the status of bias, such as how many studies are positively/negatively biased. We hope to use the replication examples to help researchers better understand the rationale of why the biases occurred and how to handle them.

In the following sections, we review the statistical fundamentals of the five DiD designs, explain which DiD design results in biased TWFE estimates and why, and demonstrate how to use alternative estimators to avoid the bias. We first explain two group two time period design, followed by multiple group and multiple time period design. We then demonstrate why using event study methodology to test the parallel trend assumption is problematic. We conclude the study by proposing best practices to conduct DiD research in the OM field.

## 2. TWO GROUP TWO TIME PERIOD

Throughout this article, we use $i = 1 \ldots I$ to represent cross-sectional units, $t = 1 \ldots T$ to represent time periods, $g = 1 \ldots G$ to represent different groups. For example, different individuals $i$ coming from different income groups $g$ have improved their earnings over a time period $t$.

### 2.1 Two Group Two Time Period Design

DiD in its simplest form consists of two groups (G = 2) and two time periods (T = 2). Groups can be individuals living in different states, firms operating in different countries and regions, and etc. Time period can be a quarter, a month, a week, and etc. Neither group was exposed to an exogenous shock, such as a policy intervention or a rare event, in the first time period. However, one group was exposed to the exogenous shock at the start of the second time period, therefore, this group was assumed to be experiencing the potential impact of the exogenous shock during the entirety of the second time period – hence the terminology "treatment" and this group was accordingly called "the treatment group". The other group that did not experience the exogenous shock is referred to as "the control group". The outcome Y was observed for each group over these two time periods and the two outcomes for the two groups were compared to each other to estimate the so-called "treatment effect".

Statistically, let A be the control group, B the treatment group, $t$ a time indicator for the two time periods, $t_2$ a time period dummy variable equal to 1 for any units $i$ in the second time period. Altogether we observe four groups: A before, A after, B before, B after. We write Equation 1 in below where Y is the outcome variable (Wooldridge 2010):

$$Y = \beta_0 + \beta_1 tB + \delta_0 t_2 + \delta_1 t_2 * tB + \varepsilon \qquad \text{Equation 1}$$

A breakdown of the different effects is summarized in Table 1 (Wooldridge 2010). $\beta_1$ captures the potential differences between the control and treatment groups before the policy intervention while $\delta_1$ captures the differences between the control and treatment groups (the first difference) before and after the policy change (the second difference) (Equation 2), hence the term "difference-in-difference". $\delta_1$ is commonly known as the treatment effect on the treated. Throughout this article, we use $\delta$ or $\delta^{DD}$ to indicate the treatment effect.

Table 1 Two Groups Two Time Periods

| | Before Treatment | After Treatment | After – Before |
|---|---|---|---|
| Control Group A | $\beta_0$ | $\beta_0 + \delta_0$ | $\delta_0$ |
| Treatment Group B | $\beta_0 + \beta_1$ | $\beta_0 + \delta_0 + \beta_1 + \delta_1$ | $\delta_0 + \delta_1$ |
| B – A | $\beta_1$ | $\beta_1 + \delta_1$ | $\delta_1$ |

$$\delta^{DD} = \hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}) \qquad \text{Equation 2}$$

Ordinary least squares (OLS) is the most commonly used estimator to estimate $\delta_1$ the treatment effect, where researchers normally estimate heteroskedasticity-robust standard errors, allowing different group variances and different time period variances in the regression (Wooldridge 2010). OLS also provides straightforward causal inferences for the treatment effect.

### 2.2 Pitfalls of Two Group Two Time Period Design

Although the basic 2×2 design is powerful and straightforward to implement using OLS, this approach suffers from some pitfalls. First, during the two time periods, the control group and treatment group may be trending at different rates on their outcomes that have nothing to do with the intervention, hence violating the "parallel trend" assumption (more in Section 6). One simple solution to this problem is to obtain more control groups (G > 2) and/or more time periods (T > 2), which leads to our next topic – multiple group and multiple time period design. Second, there may exist some compositional effects in the design. For example, the number of individuals belonging to the same income group might change during a two year 2×2 design as some individuals may have changed jobs and accordingly their income have changed during the two years. To solve this issues, researchers normally attempt to control for changes in composition by including additional control variables in the regression.

Due to these two major pitfalls, the basic 2×2 design is most commonly applied in a lab-controlled randomized experiment environment where the problems associated with simple 2×2 design can be alleviated. A recent example is Peinkofer and Jin (2022) who ran 6 randomized experiments through Amazon Mechanical Turk to investigate how disclosing order fulfillment information (Shipped by Amazon for example) impacts participant's reactions to deceptive counterfeit products.

## 3. MULTIPLE GROUP MULTIPLE TIME PERIOD – SINGLE TREATMENT

Since most operations management policy interventions/changes are not randomized experiments, quasi-experiment design is a universal approach used to estimate the counterfactual effects and draw causal inference in recent OM DiD studies. To alleviate the potential problems associated with the basic 2×2 design in a quasi-experiment setting, researchers normally include multiple groups (G > 2) and multiple time periods (T > 2) in the design. In a multiple group multiple time period design, the treatment timing may be the same for all groups and all units or the treatment timing may be different for different groups and different units. The former is an extension of the basic 2×2 design and we use "single treatment" to define the design. The latter is commonly known as staggered DiD design (Barrios et al. 2022, Li et al. 2022, Mithas et al. 2022, Cheng et al. 2023, Gong et al. 2023) or variation in treatment timing (Callaway and Sant'Anna 2021). Regardless of a single treatment or staggered treatment, researchers tend to estimate either a static effect or a 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 changes over time (Sun and Abraham 2021, Baker et al. 2022).

In a multiple group multiple time period design, therefore, there are altogether four different designs: single treatment with static effect, single treatment with dynamic effect, staggered treatment with static effect, and staggered treatment with dynamic effect. We next discuss each type of design, demonstrate the validity and/or potential pitfalls of TWFE in each design, and propose relevant approaches to handle these pitfalls.

### 3.1 Single Treatment Static Effect

#### *3.1.1 Single Treatment Static Effect Design*

Among the 51 studies we surveyed in the three top OM journals, 23 studies (45%) examined static effect with a single treatment consisting of either multiple groups, or multiple time periods, or a combination of both. A standard static effect specification, in its simplest form, is illustrated in Equation 3.

$$Y_{it} = \alpha_i + \lambda_t + \delta^{DD} D_{it} + \varepsilon_{it} \quad \text{Equation 3}$$

$\alpha_i$ is the unit fixed effect, $\lambda_t$ is the time fixed effect, and $\varepsilon_{it}$ is the error term. $D_{it}$ is a binary variable taking the value of 1 if unit $i$ is treated in time period $t$ and taking the value of 0 otherwise. The parameter of interest in the static specification is again $\delta^{DD}$, which is typically known as the average treatment effects on the treated (ATT), i.e., the overall treatment effect for all units that have participated in the treatment up to a certain time period T.

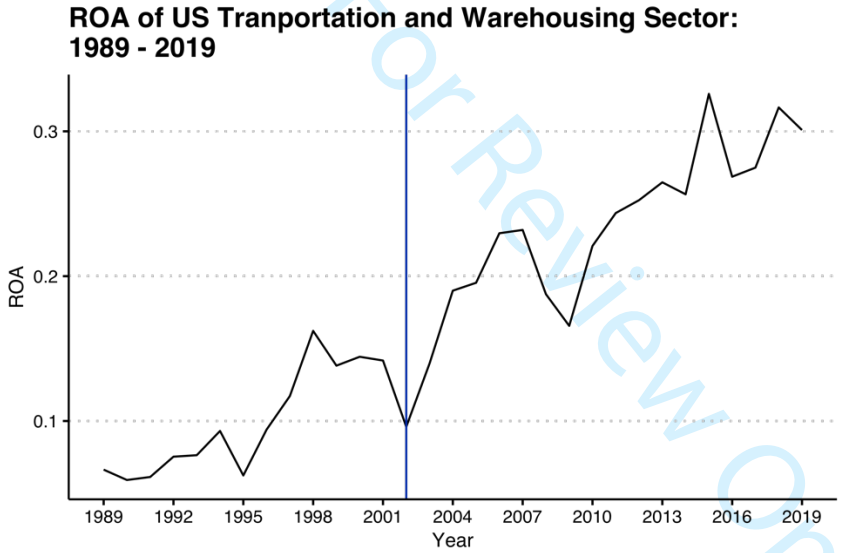
#### *3.1.2 Validity of TWFE in Single Treatment Static Effect Design – A Simulation*

Adding multiple groups and/or multiple time periods attempts to increase the validity of a basic 2×2 design where there is no variation in treatment timing (Wooldridge 2010). We empirically test this practice in this section. 23 out of the 51 studies we reviewed estimated static effect with a single treatment (Appendix 1). But we were not able to locate any electronic companions of data or code to replicate any of these 23 studies. To this end, we echo Pagell's (2020) advocate that operations and supply chain management field is going through a replication crisis. We call for researchers to share their data (if not proprietary) and code for replication purposes to facilitate knowledge dissemination and accumulation in the OM field (Frohlich and Dixon 2006, Davis et al. 2023), as is evidenced by the recent reproducibility project conducted by Davis et al. (2023).

We elect to use simulation to demonstrate the validity of estimating static effect using TWFE in a multiple group and multiple time period setting when there is no variation in treatment timing. 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 the finance field to illustrate the efficacy of TWFE DiD estimates (Baker et al. 2022). Therefore, we also use Compustat data to simulate a single treatment static effect 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. Figure 1 shows the trend of ROA for the Transportation and Warehouse sector from 1989 to 2019 in the U.S.

Figure 1 ROA of US Transportation and Warehousing Sector: 1989 – 2019




To simulate a single treatment effect, we artificially create an exogenous shock in 2002 (blue vertical line in Figure 1) by randomly assigning half of the firms as treatment group and half of the firms as control group. For the treated firms, we artificially assign a 5% increase of the standard deviation of ROA each year after the treatment (5% is the actual average year-over-year change in ROA from 2003 to 2019 in our data). Figure 2 below represents the visualization of the two random groups (note that the graph shows a parallel trend for the two groups before 2002. More in Section 6). We then run our simulation 1000 times and use Equation 3 to estimate the treatment effect. We compute actual firm average treatment effect (average of the average treatment effect of all firms) and actual observation average treatment effect (average of the treatment effect of all treated observations) following recent econometrics literature (Callaway and Sant'Anna 2021, Baker et al. 2022). We plot the simulation results in Figure 3. In Figure 3, 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 closely align to the center of the distribution of the estimated TWFE effects from 1000 simulations, indicating that TWFE estimate can reproduce the true treatment effect when there is only a single treatment in a multiple group and multiple time period setting.

Figure 2 Comparison between Control Group and Treatment Group

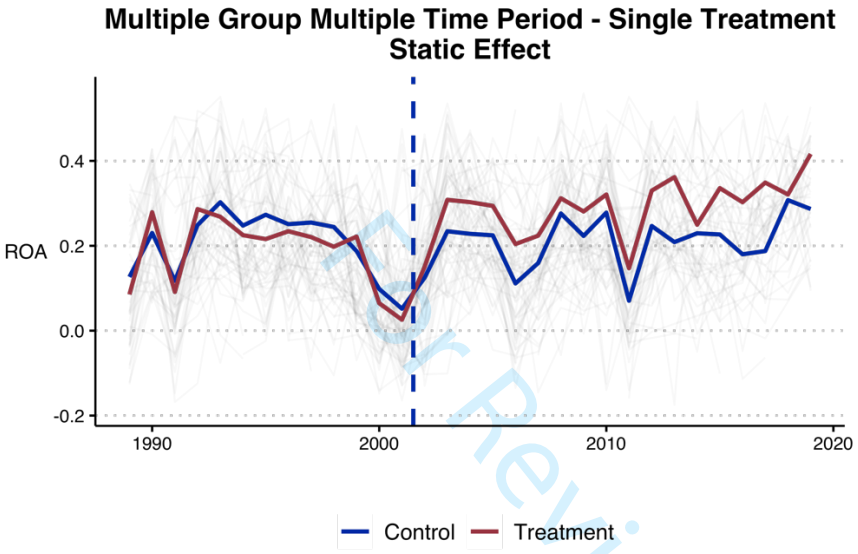
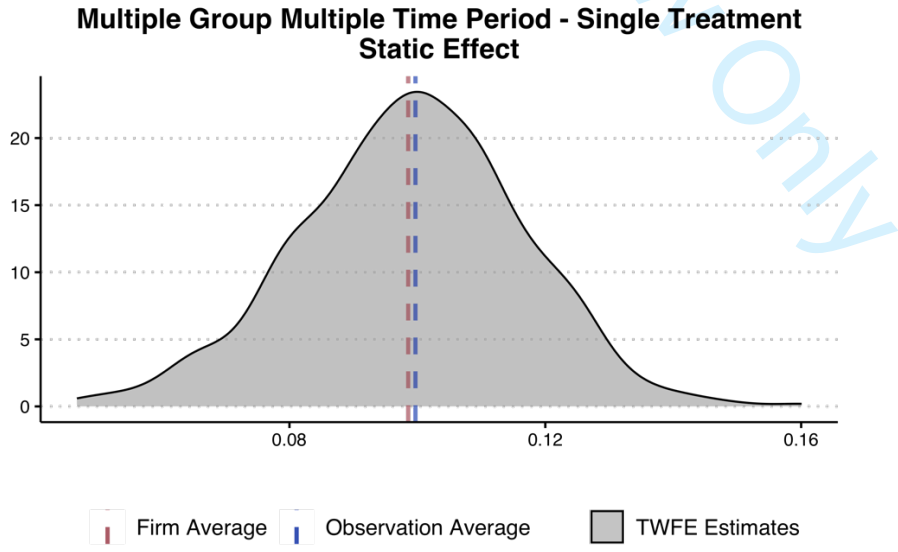



Figure 3 Simulation Result – Single Treatment Static Effect




### 3.2 Single Treatment Dynamic Effect

#### *3.2.1 Single Treatment Dynamic Effect Design*

In a multiple group multiple time period design where the treatment timing is the same for all groups, researchers are also interested in estimating the coefficients of relative time indicators after the treatment,

i.e., how the treatment effect evolves over time. These coefficients are interpreted as the average treatment effect at different lengths of exposure to the treatment. The baseline model of a standard dynamic effect is illustrated in Equation 4:

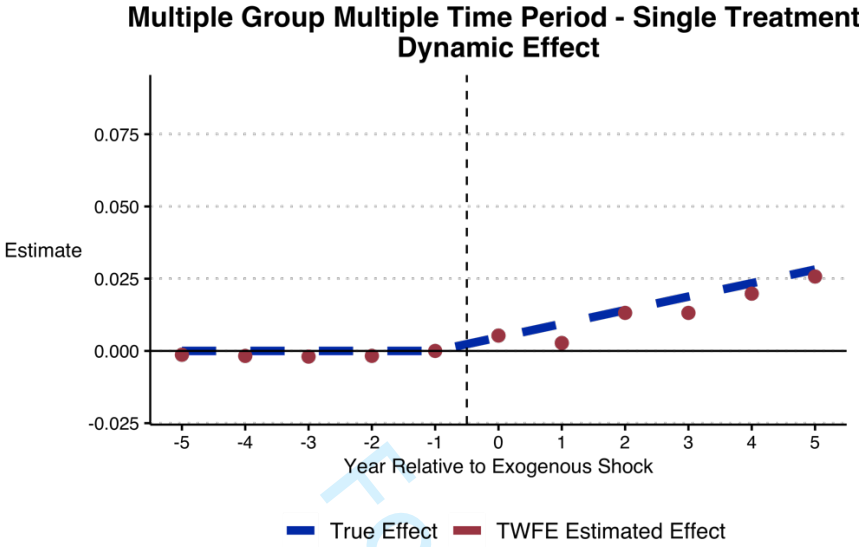
$$Y_{it} = \alpha_i + \lambda t + \sum_{l=-K}^{-2} \mu_l \boldsymbol{D_{it}^l} + \sum_{l=0}^{L} \mu_l \boldsymbol{D_{it}^l} + \beta \boldsymbol{X}_{it} + \varepsilon_{it} \quad \text{Equation 4}$$

The interpretation of $\alpha_i$ and $\lambda t$ remains the same: they are vectors of individual fixed effects and time fixed effects respectively. $\boldsymbol{X}_{it}$ is the vector of time-varying control variables and $\varepsilon_{it}$ is the error term. Instead of using a single binary indicator of $D_{it}$ in Equation 3, dynamic effect or event study methodology adopts a set of relative time indicators $D_{it}^l$. $l$ = (–K, ... , L) represents the length of time periods relative to the time period when the treatment started, such as (–2, –1, 0, 1, 2, 3) where 0 is when the treatment started, –2 is two time periods before the treatment, and 2 is two time periods after the treatment. In an event study specification, it is necessary to exclude some relative time periods to avoid multi-collinearity. The most common practice is to exclude relative periods close to the initial treatment (Sun and Abraham 2021). When there are no never-treated units in a panel balanced data, at least two relative time periods need to be excluded (Borusyak et al. 2021, Sun and Abraham 2021, Baker et al. 2022). In Equation 4, the time period of $t - 1$ was omitted to avoid multicollinearity. Hence the –2 in $\sum_{l=-K}^{-2} \mu_l D_{it}^l$ which captures the time periods up to the second time period before the policy change. $\sum_{l=0}^{L} \mu_l D_{it}^l$ represents the time periods after the policy change (hence the $l = 0$). The main parameters of interest in Equation 4 are the $\mu_l$s which captures the differences in the outcome $Y_{it}$ between treated and untreated units $l$ time periods apart from the treatment (i.e., the time period when the treatment started).

***3.2.2 Validity of TWFE in Single Treatment Dynamic Effect Design***

Among the 51 studies reviewed, only 5 studies fall in this category. Due to data availability issue, we continue to use the simulation setup from the previous section to demonstrate the validity of TWFE estimates of this design. The simulation setup was exactly the same from the previous section where half of the random firms receive an artificial treatment in 2002. We use Equation 4 to estimate a classic event study with 5 years before and 5 years after the treatment (with $t - 1$ omitted) using TWFE estimator. Figure 4 plots the results of TWFE event study estimates. Blue line represents the true treatment effect (calculated from the data) while the red dots represent the TWFE estimated effect for each relative year dummy. All coefficients estimated by TWFE are statistically significant for both the pre-periods and the post-periods, reflected in the close alignment between the blue actual line and the red estimated line, indicating the validity of TWFE estimator when assessing dynamic treatment effect, or event study, in a multiple group and multiple time period setting when there is no variation in treatment timing.

Figure 4 Simulation Result – Single Treatment Dynamic Effect




## 4. MULTIPLE GROUP MULTIPLE TIME PERIOD – STAGGERED DESIGN

In this section, we explain the design, potential pitfalls of TWFE, and the proposed approaches to handle the pitfalls for the other two types of designs where there are multiple groups and multiple time periods but the treatment timing is different for different groups: staggered treatment with static effect, and staggered treatment with dynamic effect.

### 4.1 STAGGERED TREATMENT STATIC EFFECT

#### *4.1.1 Staggered Treatment Static Effect Design*

The basic form of this design shares the same equation (Equation 3) with the single treatment static effect design. The only difference is that in a staggered design when researchers still estimate the average treatment effect on the treated after the intervention, the treatment timing is different for different units in different groups. TWFE regression again is a popular method to test static treatment effect across many fields. However, recent development in econometrics shows that TWFE regression produces biased estimates under this design, as is explained below.

#### *4.1.2 Pitfalls of TWFE in Staggered Treatment Static Effect Design*

Among the 51 studies we surveyed, only 1 study briefly touched upon the issues associated with estimating static effect in a staggered DiD design (Barrios et al. 2022). Albeit some researchers claim that a staggered DiD design increases the validity of the design, we show how TWFE suffers from biased estimates both theoretically and practically in this section. Our analysis reinforces the call for adopting the appropriate estimator in scientific research in the OM field (Shang and Rönkkö 2022).

We first explain the theoretical underpinning of why TWFE DiD estimator is biased when there is variation in treatment timing. When estimating static effect from staggered treatments, OLS effectively conducts four different comparisons: early-treated groups VS never-treated groups, later-treated groups VS

never-treated groups, early-treated groups VS later-treated groups, and later-treated groups VS early-treated groups (the latter groups are used as controls). If there are no never-treated control groups, 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 comparison is the *later-treated groups VS early-treated groups* where the early-treated groups are used as controls. The reason is that when used as controls, early-treated groups have received treatments already, so the changes in the early-treated groups over time may result from the treatment itself. Therefore, the comparison is not valid. In other words, *early-treated groups* are not "clean controls" and the *later-treated groups VS early-treated groups* itself is not a clean 2×2 design.

Further, when OLS subtracts the changes of 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) showed that even ATT for every treated group is positive, the average treatment effect across all groups is negative and statistically significant. This is because $\delta^{DD}$ 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 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).

#### *4.1.3 How to Solve the Problem of TWFE in Staggered Treatment Static Effect Design*

Recent development in econometrics has proposed several alternative estimators (Callaway and Sant'Anna 2021, Sun and Abraham 2021) to address the issues associated with TWFE in staggered DiD design. The commonality of these alternative estimators is to modify those units that can serve as control units, i.e., avoid using already-treated units as effective controls. We use Callaway and Sant'Anna (2021) (Equation 5) as an example to illustrate how the alternative estimator avoids the contamination issue in TWFE DiD regression.

$$\boldsymbol{\theta^O = \sum_{g\,\epsilon\,\mathcal{G}} \theta(g) P(G = g | G \leq T)} \qquad \text{Equation 5}$$

To differentiate from the treatment effect of $\delta$ or $\delta^{DD}$ in TWFE regression, $\theta^O$ is used for the new estimators to represent the average treatment effect for all units that have participated in the treatment in group $g$. $\theta^O$ first computes the average effect of each group $g$ across all time periods $t$. Then, $\theta^O$ averages the different group effects across all groups into a single aggregated treatment effect. To this end, $\theta^O$ is the equivalent to

ATT in a canonical 2×2 DiD design (i.e., 2 groups and 2 time periods), thus avoiding the contamination issue associated with TWFE. $\theta^O$ performs well in research settings where there are variations in treatment timing across multiple time periods. In sum, the new estimator first computes the average ATT for each treatment group across all post-periods, then aggregate all the ATTs based on each group's sample share to avoid the problems associated with TWFE.

#### *4.1.4 An Illustrative Example using a Recent Study*

Among the 51 studies we surveyed, 21 studies (41%) fall into this category and are susceptible to biased and misleading research results produced by TWFE regression. Due to data availability issue of these 21 studies, we instead use a recent Management Science publication (Wang and Overby 2022) that comes with data to explain in detail why TWFE produces biased result. Note that we do not aim to critique the work of Wang and Overby (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. Neither do we aim to produce a comprehensive replication study for all past quasi-experiment designs.

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 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 5$^{th}$ occasion, hence "Availability of Lending Club" was coded as 0 before the 5$^{th}$ occasion and 1 after the 5$^{th}$ 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. Other control variables are population, number of employed individuals, average monthly earnings per individual, size of the labor force, and median household income. In a similar specification to our Equation 3 , Wang and Overby (2022) utilized TWFE DiD (Equation 6) to estimate the static effect. $LC_{it}$ is the binary variable of Lending Club availability used to test the static effect. $T_t$ are quarter fixed effects, $C_i$ are county fixed effects, $X_i$ are control variables, and $\varepsilon_{it}$ is the error term.

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 |

$$Y_{it} = \alpha + \beta LC_{it} + T_t + C_i + \gamma X_{it} + \varepsilon_{it} \quad \text{Equation 6}$$

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; and 9 states dataset where Lending Club service was approved after 2010. We use the same two datasets to replicate. More specifically, we replicate the static effect of Lending Club availability on per capita bankruptcy filing as was reported in Table 4 in Wang and Overby (2022).

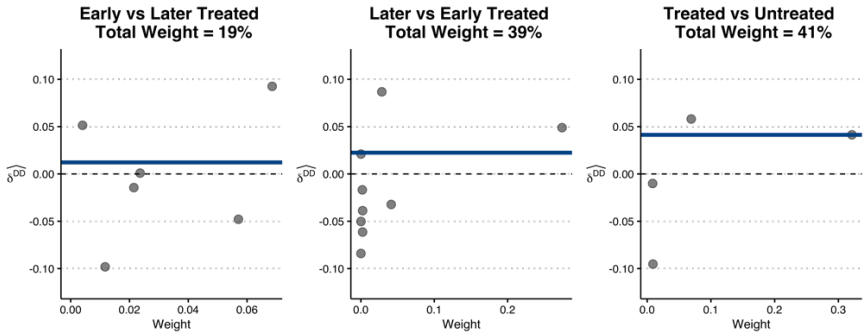
As was explained in Section 4.1.2, TWFE DiD regression has its pitfalls that lead to biased estimates in a staggered DiD design. We attempt to debunk the potential causes of the biased estimates associated with TWFE DiD regression. First, we plot the staggered time of approval of Lending Club services by different states in Figure 5. Dark red tiles represent pre-approval observations while dark blue tiles represent post-approval observations. We observe two things from Figure 5: First, there is variation in the timing of approval of Lending Club services. In the all-states data, the approval happened in Occasion 5, 13, 21, 22, 27 while in the nine-states data, the approval happened in Occasion 13, 21, 22. The variation in treatment timing indicates potential biases associated with TWFE DiD regression. Second, in the all-states data, the number/percentage of never-treated units (all-blue tiles) is small while in the nine-states data, the number of never-treated units (all-blue tiles) is large. The smaller the percentage of never-treated units, the more problematic the TWFE DiD regression might be (Callaway and Sant'Anna 2021).

Figure 5 Time of Approval of Lending Club Services


Note: Figure 10 plots time of approval of Lending Club services in all states and in 9 states. Dark red tile represent pre-approval periods and dark blue tiles represent post-approval periods.

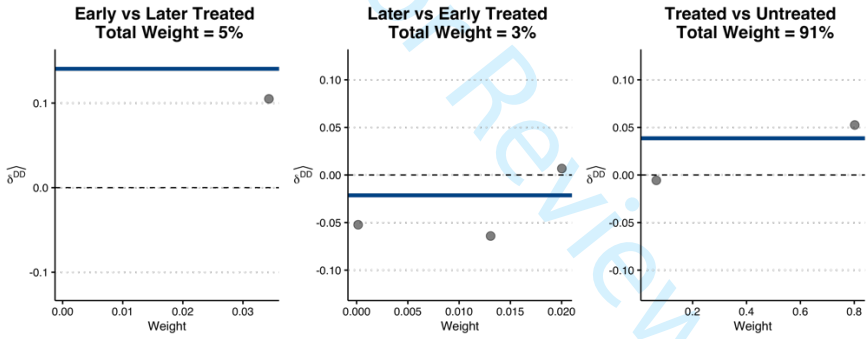
Next, we explain why the smaller the percentage of never-treated units, the more problematic the TWFE DiD regression. To do so, we decompose weights associated with TWFE using Goodman-Bacon (2021) decomposition method. Figure 6 and Figure 7 compare all possible 2×2 DiD estimates from pooled OLS regression in three categories: Early treated (as treatment) compared to later 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 counties who approved Lending Club service in Occasion 5 and those counties who approved Lending Club service in Occasion 13. 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. The overall average treatment effect calculated by TWFE is the weighted sum of each weighted average across the three groups. Among the three groups, the potentially problematic 2×2 comparisons are the second group: those Later Treated VS Early Treated comparisons. We see that in the all-states data (Figure 6) where there are less never-treated units, the weight assigned to Later VS Early Treated (problematic group) is 39% while in the nine-states data (Figure 7) where there is a large amount of never-treated units, the weight assigned to Later VS Early Treated (problematic group) is only 3%. So, theoretically, the static effect estimated by TWFE DiD regression should demonstrate larger bias for the all-states data and less bias for the nine-states data as the weight of the problematic 2×2 is 39% for the former and 3% for the latter.

Figure 6 Goodman-Bacon (2021) Decomposition of TWFE Weights – All States




Note: Figure 11 reports the decomposition the weights of TWFE static regression for all states. TWFE computes all possible 2×2 DiD estimates from pooled OLS regression in three categories: Early treated (as treatment) compared to later treated (as controls), later treated (as treatment) compared to early treated (as controls) and treated VS never-treated (as controls). Each grey dot represents a comparison between treatment-timing cohorts. The blue thick line is the weighted average of all comparisons in each group. The total weights applied by TWFE to each group is labeled as "Total Weights" in the chart title. The ATT, overall average treatment effect across groups, is the weighted sum of each weighted average.

Figure 7 Goodman-Bacon (2021) Decomposition of TWFE Weights – Nine States




Note: Figure 12 reports the decomposition the weights of TWFE static regression for nine states. TWFE computes all possible 2×2 DiD estimates from pooled OLS regression in three categories: Early treated (as treatment) compared to later treated (as controls), later treated (as treatment) compared to early treated (as controls) and treated VS never-treated (as controls). Each grey dot represents a comparison between treatment-timing cohorts.

Next, we use the estimator of Callaway and Sant'Anna (2021) to estimate the static effect of Lending Club availability on per capita bankruptcy filing. The estimates were reported in Table 3. For the static effect, we see that both TWFE and Callaway and Sant'Anna (2021) estimator report a significant positive effect, meaning per capita bankruptcy filing did increase in the given data period after Lending Club service was approved. However, compared with TWFE estimates (Wang and Overby 2022), new estimator (Callaway and Sant'Anna 2022) yield a 124% *larger effect* for all-states data and *74% larger effect* for nine-states data, corroborating that the bias is bigger when there are fewer never-treated units as is in the all-states data.

Table 3 Static and Dynamic Effects of Lending Club on Bankruptcy Filing

| | All-state matched sample | | Nine-state matched sample | |
|---|---|---|---|---|
| | Dependent Variable: Bankruptcy Filing Per Capita | | | |
| | Wang and Overby 2022 | Callaway and Sant'Anna 2021 Estimator | Wang and Overby 2022 | Callaway and Sant'Anna 2021 Estimator |
| **Static Effect** | **0.034(0.008)** | **0.076(0.013)** | **0.050(0.017)** | **0.087(0.032)** |
| Controls | Yes | Yes | Yes | Yes |

In sum, using Callaway and Sant'Anna (2021) estimator on Wang and Overby (2022) data reveals a very different findings: the static treatment effect size is significantly larger. When there is variation in treatment timing, TWFE produces biased static estimates. The bias comes from the fact that already-treated units were used as effective controls. Increasing the proportion of never-treated units can help alleviate TWFE biased estimates.

## 4.2 STAGGERED TREATMENT DYNAMIC EFFECT

### *4.2.1 Staggered Treatment Dynamic Effect Design*

Equation 4 from Section 3.2.1 also applies to dynamic effect estimation in a staggered DiD design. Event study DiD is a popular method adopted in the OM field to study time-varying policy impact (Dong et al. 2019, Cui et al. 2019, Li and Wu 2020, Cui et al. 2022). Among the 51 studies surveyed, 8 (16%) fall into this category. Popular as it is, recent advancement in econometrics again reveals major pitfalls of using event study to estimate time varying treatment effect when there is variation in treatment timing and treatment heterogeneity (Sun and Abraham 2021, Callaway and Sant'Anna 2021, Baker et al. 2022).

### *4.2.2 Pitfalls of TWFE in Staggered Treatment Dynamic Effect Design*

Sun and Abraham (2021) decomposed the regression coefficient $\sum_{l=0}^{L} \mu_l D_{it}^l$ on the relative time indicators from Equation 4 and revealed that these coefficients are a linear combination of the average treatment effect from its own time period as well as from other relative time periods. In addition, the weights associated with these coefficients are non-linear functions of the distribution of the groups, and these weights are still prone to the issue of negative weights as discussed in the previous sections. These two factors together contaminate the estimation of $\sum_{l=0}^{L} \mu_l D_{it}^l$. More statistical proof can be found in Sun and Abraham (2021).

### *4.2.3 How to Solve the Problem of TWFE in Staggered Treatment Dynamic Effect Design*

To cope with the contamination, both Sun and Abraham (2021) and Callaway and Sant'Anna (2021) developed very similar estimators to assess dynamic treatment effect. Due to the ease of incorporating covariates, we focus on Callaway and Sant'Anna (2021) estimator to illustrate. Callaway and Sant'Anna (2021) proposed to estimate the following average treatment effect in Equation 7. We again use $\theta$ to represent the treatment effect in the new estimators as opposed to $\delta$ or $\delta^{DD}$ in TWFE. $e$ represents length of time periods, equivalent to $l$ in Equation 4.

$$\theta(\boldsymbol{e}) = \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}} \mathbf{1}\{\boldsymbol{g}+\boldsymbol{e} \leq \boldsymbol{T}\}\boldsymbol{P}(\boldsymbol{G}=\boldsymbol{g}|\boldsymbol{G}+\boldsymbol{e} \leq \boldsymbol{T})\boldsymbol{ATT}(\boldsymbol{g},\boldsymbol{g}+\boldsymbol{e}) \quad \text{Equation 7}$$

$\theta(e)$ is "the average effect of participating in the treatment $e$ time periods after the treatment was adopted across all groups that are ever observed to have participated in the treatment for exactly $e$ time periods" (Callaway and Sant'Anna 2021, p. 209). $\theta(e)$ is equivalent to $\sum_{l=0}^{L} \mu_l D_{it}^l$ in Equation 4 in standard event study design while completely avoids the contamination associated with these coefficients (c.f. Section 3

in Callaway and Sant'Anna 2021). In sum, new estimator first estimates the group-specific treatment effects for each time period, then aggregate these effects to produce an overall effect for each time indicator.

***4.2.4 An Illustrative Example using a Recent Study***

We continue using Wang and Overby (2022) to demonstrate the pitfalls of traditional TWFE event study method. Wang and Overby (2022) also estimated an event study of the effect of availability of Lending Club using Equation 8. $LC_{it+\tau}$ is the relative time indicator of Lending Club availability used to test the dynamic effect. Wang and Overby omitted $t-1$ time period to avoid multicollinearity and binned distant periods that are beyond −8 and +8 time periods.

$$Y_{it} = \alpha + \sum_{\tau=-8}^{-2} \boldsymbol{\rho_\tau LC_{it+\tau}} + \sum_{\tau=0}^{8} \boldsymbol{\rho_\tau LC_{it+\tau}} + Tt + Ci + \gamma X_{it} + \varepsilon_{it} \quad \text{Equation 8}$$

We again use the estimator of Callaway and Sant'Anna (2021) to estimate the dynamic effects of Lending Club availability on per capita bankruptcy filing. Different from Wang and Overby (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 8 post-periods to match the results of Wang and Overby (2022). The estimates were reported in Table 4.

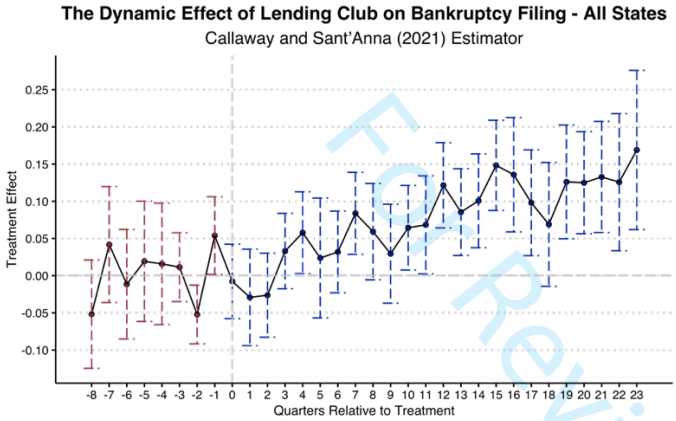
Table 4 Static and Dynamic Effects of Lending Club on Bankruptcy Filing

| | All-state matched sample | | Nine-state matched sample | |
|---|---|---|---|---|
| | Dependent Variable: Bankruptcy Filing Per Capita | | | |
| | Wang and Overby 2022 | Callaway and Sant'Anna 2021 Estimator | Wang and Overby 2022 | Callaway and Sant'Anna 2021 Estimator |
| -8 | 0.008(0.013) | -0.052(0.024) | 0.037(0.032) | -0.040(0.046) |
| -7 | 0.002(0.018) | 0.042(0.026) | 0.017(0.037) | 0.009(0.048) |
| -6 | 0.013(0.021) | -0.012(0.025) | 0.053(0.040) | -0.002(0.047) |
| -5 | 0.005(0.018) | 0.019(0.027) | 0.010(0.032) | 0.022(0.043) |
| -4 | 0.004(0.012) | 0.016(0.027) | -0.015(0.032) | 0.017(0.047) |
| -3 | 0.002(0.012) | 0.011(0.015) | -0.024(0.037) | -0.021(0.039) |
| -2 | -0.015(0.012) | **-0.052(0.013)** | -0.011(0.012) | -0.010(0.036) |
| -1 | - | **0.054(0.017)** | - | 0.029(0.036) |
| 0 | **0.024(0.013)** | -0.008(0.017) | 0.027(0.033) | 0.038(0.041) |
| 1 | 0.023(0.016) | -0.029(0.022) | 0.067(0.044) | 0.027(0.043) |
| 2 | 0.021(0.016) | -0.026(0.019) | **0.070(0.034)** | 0.042(0.036) |
| 3 | **0.030(0.014)** | 0.033(0.017) | 0.046(0.037) | 0.076(0.040) |
| 4 | **0.054(0.016)** | **0.058(0.018)** | 0.022(0.035) | 0.030(0.034) |
| 5 | **0.030(0.018)** | 0.024(0.027) | 0.046(0.039) | 0.050(0.040) |
| 6 | 0.024(0.017) | 0.032(0.018) | 0.063(0.047) | 0.131(0.046) |
| 7 | **0.047(0.014)** | **0.084(0.018)** | **0.068(0.038)** | 0.109(0.039) |
| 8 | **0.061(0.014)** | 0.059(0.022) | **0.085(0.033)** | 0.125(0.045) |
| Controls | Yes | Yes | Yes | Yes |

To unravel the dynamic effect, we plot the full dynamic effect in Figure 8 and Figure 9. We see that the maximum post-treatment effect estimated is Occasion 23 for all-states data and Occasion 15 for nine-states data, consistent with the robustness test of Wang and Overby (2022). In contrast to Wang and Overby (2022) who found the treatment effect stabilizes over time in a fully dynamic specification, we see a clear

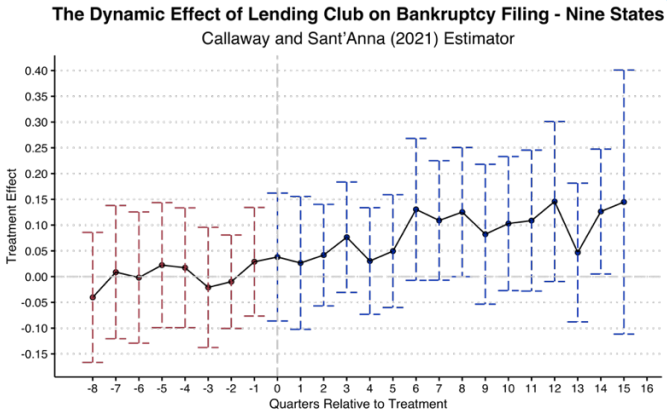
upgoing trend of the treatment effect, especially for the all-states data in Figure 8. This tells us that the effect of Lending Club on bankruptcy filing is a long-run driven effect: the longer the time of holding the loan from Lending Club, the more likely consumers will file for bankruptcy. This in turn corroborates Wang and Overby's (2022) reasonings that 1) issuing loans to unqualified borrowers might increase bankruptcy filing as these borrowers are unlikely or unable to pay back; and 2) some consumers may deliberately take advantage of the marketing lender to pay off their secured loans (which may take some time) and then deliberately file for bankruptcy once the secured loans are paid off.

Figure 8 Dynamic Effect of Lending Club on Bankruptcy Filing – All States




Note: Figure 13 reports estimates and confidence intervals for the dynamic effects of bank merges on per capita bankruptcy filing in all states. Data was from Wang and Overby (2022). Standard errors are clustered at the county level. A fully dynamic model was estimated and up to 23 post-treatment estimates were estimated.

Figure 9 Dynamic Effect of Lending Club on Bankruptcy Filing – Nine States




Note: Figure 14 reports estimates and confidence intervals for the dynamic effects of bank merges on per capita bankruptcy filing in nine states. Data was from Wang and Overby (2022). Standard errors are clustered at the county level. A fully dynamic model was estimated and up to 15 post-treatment estimates were estimated.

In sum, using Callaway and Sant'Anna (2021) estimator on Wang and Overby (2022) data reveals a fundamentally different finding: the dynamic effect increases over time, rather than stabilizes over time. The upgoing trend we found actually explains Wang and Overby's (2022) reasonings and suspicions. Our finding, therefore, provides a very different insight to banks and market lenders for proper decision making,

such as aggressively stop lending to unqualified borrowers and place more stringent background checks on those borrowers with secured loans to avoid being taken advantage of by these borrowers.

### 4.3 WHEN DOES TWFE DID PRODUCE BIASED ESTIMATES

After reviewing the 5 different DiD designs, we briefly summarize the key findings in recent econometrics literature using an empirical-focused OM journal JOM to reinforce the current status of the application of DiD design in the OM field. First, a simple two group two time period 2×2 design is not prone to biased estimates. Among the 13 DiD articles we surveyed in JOM, only one article loosely falls into this category (Chen et al. 2023). Second, when there are multiple group and multiple time periods but with only one single treatment, i.e., the treatment timing is the same for all groups, TWFE estimates, either static or dynamic, are not biased. Five of the 13 articles we surveyed fall within this category (Ren et al. 2023, Scott et al, 2023, Fan et al. 2022, Miao et al. 2022, Chun et al. 2022). Third, in a multiple group and multiple time period setting with variation in treatment timing, TWFE estimates of the static effect is biased (see Wang and Overby 2022 example). In addition, TWFE estimates of event study/dynamic effect is also biased (see Wang and Overby 2022 example). Staggered design is a more commonly used design in OM research. Four out of the 13 articles in JOM DiD studies adopted such a staggered design to estimated static (Wang et al. 2023, Lam et al. 2022, Barrios et al. 2022, Ge et al. 2022) and/or dynamic effects (Wang et cl. 2023, Barrios et al. 2022). This category is where the TWFE estimator is most susceptible to biased results.

## 5. INSTANTANEOUS TREATMENT EFFECT

So far, we have focused on the static and dynamic effect of DiD design. However, the instantaneous effect of exogenous shocks is also important for policymakers. Take Covid-19 for example. One of the bottlenecks causing high death tolls during the immediate outbreak of the first wave was the unprepared healthcare resources when we knew too little about the pandemic (Webb et al. 2022). To better battle similar future pandemics by learning from Covid-19, the impact of its immediate outbreak, i.e., the instantaneous effect of Covid-19, is crucial for public health policymakers to refer to so that policymakers can fully mobilize necessary resources to save lives. Merely presenting a static average effect or a dynamic effect from previous pandemics will not serve the right purpose to cope with the immediate outbreak of future pandemics. Similar arguments can be extended to operations management policy changes and interventions when the instantaneous effect could be more pronounced due to the abrupt exogenous shock. However, the instantaneous effect is largely ignored in extant OM research. In this section, we review the TWFE estimator of instantaneous effect, explain the pitfalls of TWFE, propose relevant estimator to alleviate the pitfalls, and use one recent publication to illustrate the difference between TWFE estimator and the proposed method.

### 5.1 Instantaneous Treatment Effect Design using TWFE

To test the instantaneous treatment effect, researchers usually rely on the following TWFE DiD regression where a general form is expressed in Equation 9 (adapted from Bliese and Lang 2016):

$$Y_{it} = \alpha_i + \lambda_t + \delta^{DD} D_{it} + \beta_1 \boldsymbol{T_\ell} + \beta_2 \boldsymbol{\mathcal{T}_t} + \beta_3 X_{it} + \varepsilon_{it} \quad \text{Equation 9}$$

The essence of using TWFE DiD regression to estimate instantaneous effect is still to estimate the treatment effect by comparing groups experiencing different exposures to treatment. $Y_{it}$ measures the outcome of unit $i$ in time period $t$. The interpretation of $\alpha_i$, $\lambda_t$, and $\varepsilon_{it\ is}$ is the same with previous equations: $\alpha_i$ is the units fixed effect, $\lambda_t$ is the time fixed effect, and $\varepsilon_{it}$ is the error term. $X_{it}$ is a vector of time-varying control variables. $D_{it}$ is the variable of interest and takes the value of 0 in the time period before the treatment and 1 in the time period of the treatment and thereafter. $\mathcal{T}_t$ is a relative time indicator where the relative time $\ell$ starts from the first time period after the treatment (1, 2, 3, and etc.). $T_\ell$ is an overall time indicator where $t$ starting from the first observed time period for each unit. The first observed period of $T_\ell$ is coded as 0 and the remaining coded in sequential order (1, 2, 3, and etc.) However, $T_\ell$ stops one period before the treatment and remains constant thereafter. The three time indicators are coded in this way such that the coefficient $\delta^{DD}$ measures the instantaneous effect of treatment while $\beta_2$ captures the post-treatment slope of the outcome (Bliese and Lang 2016). A statistically positive (negative) $\delta^{DD}$ means that outcome improved (deteriorated) immediately upon treatment. Table 5 below illustrates the data setup for a TWFE instantaneous effect where the treatment takes place in measurement occasion 25.

Table 5 An Example of TWFE Instantaneous Effect Design

| Year | Quarter | Measurement Occasion | $\boldsymbol{T_\ell}$ | $\boldsymbol{D_{it}}$ | $\boldsymbol{\mathcal{T}_t}$ |
|---|---|---|---|---|---|
| 2004 | 1 | 1 | 0 | 0 | 0 |
| 2004 | 2 | 2 | 1 | 0 | 0 |
| 2004 | 3 | 3 | 2 | 0 | 0 |
| 2004 | 4 | 4 | 3 | 0 | 0 |
| … | … | … | … | … | … |
| 2009 | 1 | 21 | 20 | 0 | 0 |
| 2009 | 2 | 22 | 21 | 0 | 0 |
| 2009 | 3 | 23 | 22 | 0 | 0 |
| 2009 | 4 | 24 | **23** | 0 | 0 |
| 2010 | 1 | 25 | **23** | **1** | 0 |
| 2010 | 2 | 26 | **23** | **1** | **1** |
| 2010 | 3 | 27 | **23** | **1** | **2** |
| 2010 | 4 | 28 | **23** | **1** | **3** |
| … | … | … | **…** | **…** | **…** |
| 2019 | 1 | 61 | **23** | **1** | **36** |
| 2019 | 2 | 62 | **23** | **1** | **37** |

### 5.2 Pitfalls of Using TWFE to Estimate Instantaneous Effect

The specification in Table 5 is one of the popular methods to estimate instantaneous treatment effect in social sciences (Bliese and Lang 2016). Similar to static and dynamic effect, recent development in

econometrics reveal that using TWFE regression to estimate instantaneous effect is also problematic (de Chaisemartin and D'Haultfoeuille 2020). TWFE estimate was found to suffer from contamination issues, especially when treatment effects involve multiple time periods and variation in treatment timing. To illustrate the potential contamination issues in TWFE, $\delta^{DD}$ from Equation 9 can be re-defined in the following Equation 10 (adapted from de Chaisemartin and D'Haultfoeuille 2020):

$$\delta^{DD} = E(\sum_{(g,t):D_{g,,t=1}} \boldsymbol{W_{g,t}} \Delta_{g,t}) \quad \text{Equation 10}$$

$\delta^{DD}$ is a weighted sum of the average treatment effects on the treated (ATT) for group *g* in each time period *t*. However, the weights ($\boldsymbol{W_{g,t}}$) may be negative. The reason is that $\delta^{DD}$ is the weighted sum of "several difference-in-differences … which compare the evolution of the outcome ($\Delta_{g,t}$) between consecutive time periods" (de Chaisemartin and D'Haultfoeuille 2020, p. 2905) across observation units. If any of the control group used in these comparisons are treated at both time periods, their treatment effect will be differenced out at the second period, resulting in negative weights. These negative weights will cause a biased and misleading estimation of the coefficient $\delta^{DD}$, especially "if the treatment effect is heterogeneous across groups and time periods" (de Chaisemartin and D'Haultfoeuille 2020).

**5.3 How to Correctly Estimate Instantaneous Effect**

To cope with the negative weights issue, de Chaisemartin and D'Haultfoeuille (2020) proposed to estimate the following coefficient $\delta^{S}$ (Equation 11) using a different estimator $DID_M$ (cf. p. 2978 for more detailed interpretations). de Chaisemartin and D'Haultfoeuille (2020) showed that the new estimator $DID_M$ is an unbiased and consistent estimator that is robust to treatment effect heterogeneity across groups and time periods in both a single treatment design and a staggered design.

$$\delta^{S} = E\left[\frac{\mathbf{1}}{\boldsymbol{N_S}} \sum_{(i,g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}} \left[\boldsymbol{Y_{i,g,t}(1) - Y_{i,gt}(0)}\right]\right] \quad \text{Equation 11}$$

In a staggered design where there are variations in treatment timing across multiple time periods, $\delta^{S}$ is the average of treatment effects at the time when all units start receiving treatmment by comparing units whose treatment status has changed with those units whose treatment status is unchanged between *t – 1* and *t* ($\left[\boldsymbol{Y_{i,g,t}(1) - Y_{i,gt}(0)}\right]$). Therefore, $\delta^{S}$ captures the average treatment effect among switchers right at the time period when they switch and avoids the negative weights issues and contamination issues discussed above.

**5.4 An Illustrative Example using a Recent Study**

We use a recent publication to illustrate the difference between TWFE and the estimator proposed by de Chaisemartin and D'Haultfoeuille (2020). The number of studies estimating instantaneous effect is extremely limited and we again elect to analyze one study from Management Science that comes with data.

Avramidis et al. (2022) investigated the impact of bank mergers on the number of bank branches in the U.S. using a nine-year panel data. The authors controlled for year fixed effect, demand driven changes

(measured by predicted changes in the number of bank branches), and market competition (measured by changes in the number of loans issued by the biggest market lender Lending Club). Table 6 is a mini version of the original data. "Group" is a truncated three-digit zip code. Each unique three-digit zip code represents a local market. "Number of bank branches" shows the number of bank branches in each local market in each year. Dependent variable is "Change in Number of Bank Branches" from years $t-1$ to $t$. "Merger Event" is an indicator variable "that takes the value of 1, 0 otherwise, if local market z is affected by such a merger from years $t-1$ to $t$" (p. 3101). The last two columns of Table 6 are the two control variables capturing demand-driven changes and market competition. We see that there were two merger events in the local market 010: one in 2011 (change of status from 2010 to 2011) and one in 2014 (change of status from 2013 to 2014). The dependent variable calculates the changes in the number of bank branches from $t-1$ to $t$. Avramidis et al. (2022) coded the data in such a way that the regression will capture the instantaneous effect among local markets who experienced bank mergers right at the time period when the merger occurred.

Table 6 Example Data of Avramidis et al. 2022

| Year | Group | Number of Bank Branches | Change in Number of Bank Branches | Merger Event | Predicted Change in Number of Bank Branches | Year Lagged Annual Proportional Changes in Marketing Lending |
|---|---|---|---|---|---|---|
| 2008 | 010 | 149 | | 0 | | 0.00 |
| 2009 | 010 | 152 | 3 | 0 | 0.77 | 1.33 |
| 2010 | 010 | 151 | -1 | 0 | (0.54) | 1.41 |
| 2011 | 010 | 151 | 0 | 1 | 0.15 | 0.67 |
| 2012 | 010 | 149 | -2 | 0 | (0.37) | 0.84 |
| 2013 | 010 | 150 | 1 | 0 | (0.52) | 1.11 |
| 2014 | 010 | 149 | -1 | 1 | (0.83) | 0.91 |
| 2015 | 010 | 144 | -5 | 0 | (0.72) | 0.51 |
| 2016 | 010 | 144 | 0 | 0 | (0.30) | 0.55 |

We specifically replicate column 2 and column 3 of the original Table 4 in Avramidis et al. (2022) and report our findings in Table 7. Panel A tests the impact of bank merges on the number of bank branches by controlling for year fixed effects and demand-driven changes. Panel B further adds market competition as an additional control. $\widehat{\beta_{fe}}$ is the estimates reported by Avramidis et al. (2022). We see that in both panels, the effect of $\widehat{\beta_{fe}}$ is statistically significant at the 5% confidence level, indicating that bank mergers reduced the number of local bank branches by 0.20 in the merger year.

Table 7 Instantaneous Effect of Bank Mergers on Bank Branch Changes

| | Panel A (Year fixed effects and demand driven changes) | | Panel B (further adds market competition) | | |
|---|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error | *N* |
| $\widehat{\beta_{fe}}$ | -0.200* | 0.098 | -0.200* | 0.098 | 7056 |
| $\mathrm{DID_M}$ | -0.125** | 0.019 | -0.137 | 0.384 | 6174 |
| $\mathrm{DID_M^{pl}}$ | -0.255 | 0.818 | -0.268** | 0.040 | 3833 |
| $\mathrm{DID_M^{pl,2}}$ | -0.202 | 0.448 | -0.202* | 0.068 | 2723 |
| $\mathrm{DID_M^{pl,3}}$ | 0.404 | 0.481 | 0.406 | 0.202 | 1904 |
| $\mathrm{DID_M^{pl,4}}$ | 0.247 | 0.534 | 0.249 | 0.780 | 1225 |

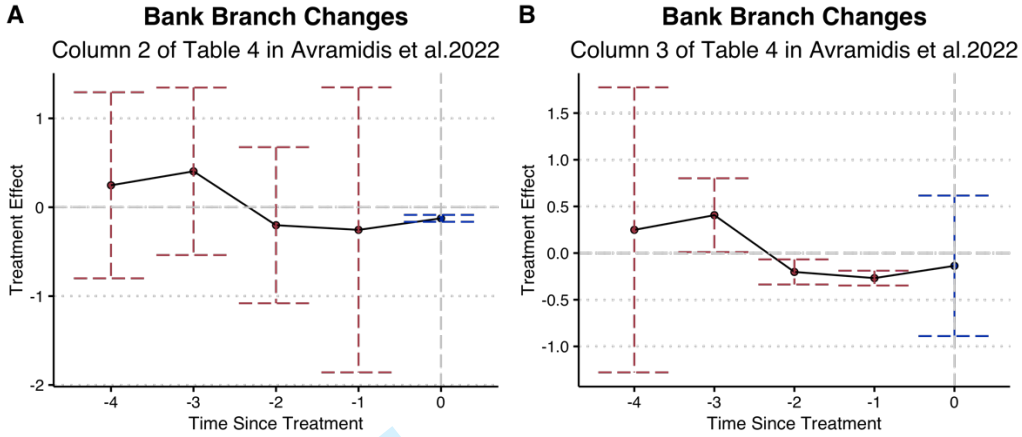
P value: *** 0.001, ** 0.01, * 0.05, °0.1

Note: This table reports estimates of the effect of under the common trend assumption. Estimators are computed using the data of Avramidis et al. 2022. Standard errors are clustered at the zip code level.

Since the data measures an instantaneous treatment effect, we use the estimator of de Chaisemartin and D'Haultfoeuille (2020) to re-compute the effect. The estimate for the instantaneous effect is reported as $\mathrm{DID_M}$ in Table 6. When only controlling for year fixed effect and demand-driven changes in Panel A, we find a similar conclusion with Avramidis et al. (2022) but with a slightly smaller effect size ($\mathrm{DID_M}$=−0.125, $t$=6.58), indicating that bank mergers reduced the number of local bank branches by 0.125 in the merger year. However, when further adding market competition as an additional control in Panel B, the treatment effect is still at the expected sign but becomes statistically insignificant ($\mathrm{DID_M}$=−0.137, $t$=0.36).

To further delineate the difference in findings, we compute placebo estimators to assess what happened before the mergers. $\mathrm{DID_M^{pl}}$ compares the change of bank branches between those experiencing mergers and those not experiencing mergers one year before the merger. We also compute three other placebo estimators $\mathrm{DID_M^{pl,2}}$, $\mathrm{DID_M^{pl,3}}$, and $\mathrm{DID_M^{pl,4}}$ that conduct the same comparison two, three, and four years before the merger. As shown in Table 7 and in Figure 10, in Panel A, none of the placebo estimators is statistically significant. However, in Panel B when market competition is added as an additional control, both $\mathrm{DID_M^{pl}}$and $\mathrm{DID_M^{pl,2}}$ are negative and significant. This indicates that local market who experienced mergers started experiencing a differential negative pretrend one and two years before eventually merged.

Figure 10 Instantaneous Effect of Bank Mergers on Bank Branch Changes




Note: Figure 9 reports estimates and confidence intervals of $DID_M$ and four placebos using data of Avramidis et al. 2022. Standard errors are clustered at the zip code level. Panel A controls for year fixed effect and demand-driven changes. Panel B controls for year fixed effect, demand-driven changes, and market competition.

The different pre-trend and treatment effect between Panel A and Panel B trigger us to propose that (without embarking on a rigorous causality test) bank merger was mainly triggered by market competition, i.e., when the market competition is high in local markets where Lending Club aggressively lends to consumers, local banks in these local markets suffer from the competition of Lending Club and started to close their branches as a result of the competition, which eventually leads to the final merger. In sum, new estimator not only finds a smaller effect size but also reveals a different story when market competition is taken into consideration.

## 6. THE PARALLEL TREND ASSUMPTION IN DID

### 6.1 Review of the Parallel Trend Assumption in Operations Management

The validity of all the above mentioned DiD design, including the simple 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 4) to test for the differences between the trend of the treatment groups and the trend of the control groups for the time periods before the treatment. 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%) tested for the parallel trend assumption, among which 4 studies only relied on visualization tools to test for the parallel trend assumption. 14 studies used event study methodology (Equation 4) to test the statistical difference on the pre-trend between treatment groups and control groups. 13 studies used both visualization tools and event study to conduct the parallel trend assumption test. 4 studies tested on the coefficient of the interaction term (treatment*pretrend time indicator). In sum, 31 (88.5%) out of the 35 studies used TWFE DiD regression to test the parallel trend assumption.

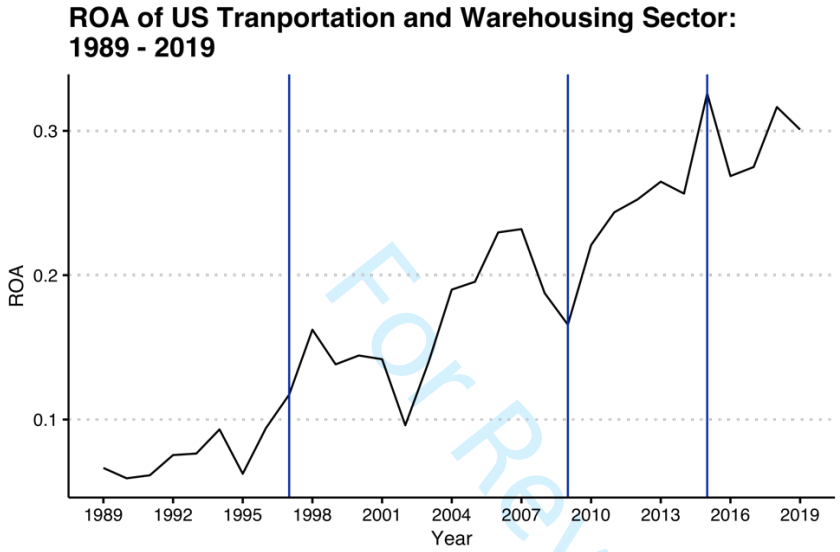
In other research fields, Roth (2020) did a survey of the applied literature and found that checking the coefficients of the pre-treatment periods is also a widespread practice to test for parallel trends. A popular practice as it is, using TWFE DiD regression to test the parallel trend assumption has also been proved to be problematic. Roth (2020) showed that using pre-treatment coefficients 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 this practice (i.e., checking the coefficients on the leads) 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. 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 Using TWFE to Test Parallel Trend Assumption is Problematic – An Example**

To demonstrate why using event study methodology (i.e., TWFE regression) to test for the parallel trend assumption is problematic, we again utilize annual ROA of the U.S. Transportation and Warehousing Sector to simulate three artificial exogeneous shocks to mimic the four distinct stages in Figure 11. From Figure 11, we observe four distinct stages of ROA: 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 11. Following recent econometrics literature (Baker et al. 2022), we design two opposite simulations. The first simulation 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, the first simulation has no pretrends (flat) for the treated groups. The second simulation forces a flat *post-trend* for all the three groups while allowing a respective annual increase of 1%, 2%, and 3% of the standard deviation of ROA *before*

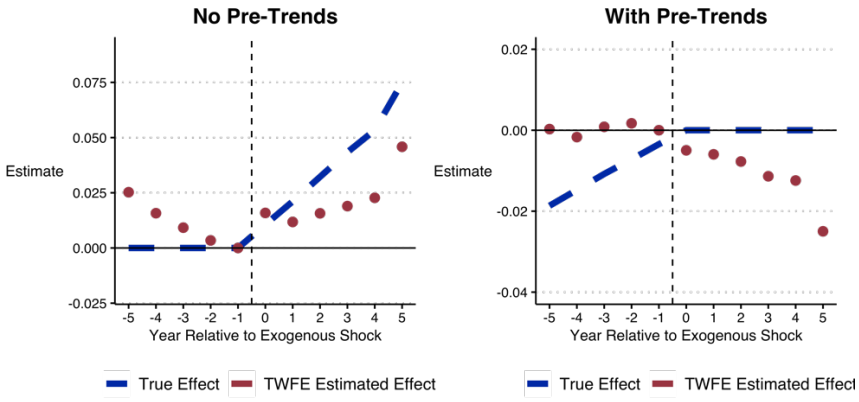
the treatment. In other words, the second simulation witnesses an upward-going pretrend for the treated groups. The choice of the annual increase in ROA in terms of number of $\sigma_{ROA}$ is random. We also randomly tested different combinations of random $\sigma_{ROA}$ and our simulation results do not change.

Figure 11 ROA of US Transportation and Warehousing Sector: 1989-2019




We then use TWFE event study (Equation 4) to estimate the coefficients of relative year time indicators (5 years pre and 5 years post the treatment). We plot the result in Figure 12. Blue line represents the actual simulated trend and the red line represents TWFE estimates. The left graph represents simulation 1 where there is no pretend (flat blue line before time 0) and the right graph represents simulation 2 where there is an actual pretend (upward trending blue line before time 0). 88.5% of the 51 studies we surveyed rely on testing coefficients of the leads to test the parallel trend assumption. However, we see from our simulation that this practice is problematic: TWFE estimated a downward going pre-trend when there is actually no pretrend (left graph) and TWFE estimated no pretrend when there is actually an upward-going pretrend (right graph). In addition, for the treatment effect (relative year indicators after time 0), TWFE estimated a downward-going effect when there is no actual treatment effect (right graph), underpinning the theoretical explanation of why using TWFE to estimate an event study is problematic discussed in the previous section.

Figure 12 Simulation to Test the Parallel Trend Assumption

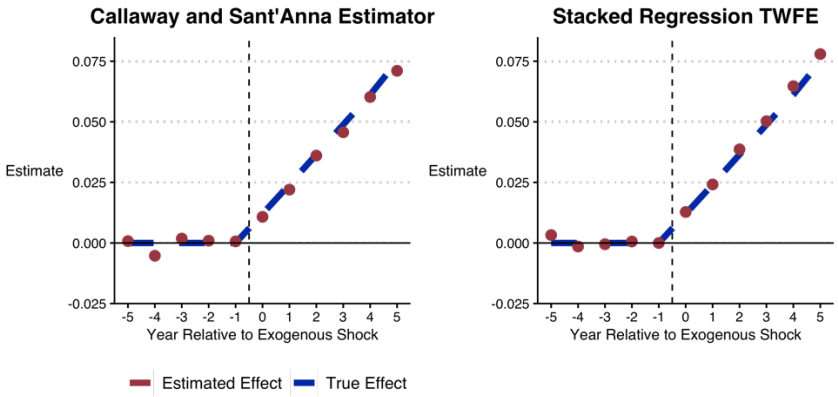



### 6.3 How to Test the Parallel Trend Assumption

We show that the common practice of checking pre-treatment coefficients using TWFE event study is problematic for testing the parallel trend assumption, we now propose two alternatives to test for the parallel trend assumption. The first alternative is either of the two recently developed estimators from the econometrics field (Callaway and Sant'Anna 2021, Sun and Abraham 2021). We use Callaway and Sant'Anna (2021) for demonstration in this study as this estimator may be what most empirical researchers choose due to the ease of incorporating covariates. The second alternative is stacked regression (Deshpande and Li 2019, Cengiz et al. 2019). Stacked regression is not an estimator per se because stacked regression also uses TWFE estimator. The trick here is to create 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 (c.f. Deshpande and Li 2019 replication package for step-by-step details). Hence, the name "stacked" regression.

We continue using simulation to demonstrate the validity of the two alternatives. We use the same simulation data that created the graph on the left in Figure 12, i.e., there is no actual pretrend for all the three groups that received exogenous shocks in 1997, 2009, and 2015. We then estimate the data using Callaway and Sant'Anna (2021) estimator and TWFE (TWFE was estimated on a reconstructed stacked data). Figure 13 reports the estimated results. Blue line represents true effect and red dots represent estimated effect. We observe that the two alternatives not only correctly estimated the pretrend but also correctly estimated the treatment effect with all coefficients statistically significant. Given the paramount role of the parallel trend assumption in the validity of DiD design, we call for researchers to adopt these two alternatives to test for parallel trend assumption to increase research stringency.

Figure 13 Alternatives to Test for Parallel Trend Assumption




In sum, to test for the parallel trend assumption, we propose a two-step approach. The first step is a preliminary visualization approach where researchers can plot the mean outcomes by group and time period to visually detect if the trends for different groups are approximately parallel. However, when the data is noisy, visualization of trends is less convincing as it might be challenging to distinguish between statistical noise and true deviations from the common trend. Therefore, the second step of using statistical analysis to verify the parallel trend is essential. During the second step, either of the two statistical alternatives (i.e., Callaway and Sant'Anna 2021 and stacked regression) explained above can be used to test for the parallel trend assumption.

## 7. CONTRIBTION AND RECOMMENDATION FOR BEST PRACTICES

Being a workhorse in OM empirical research, TWFE DiD design has been widely adopted. However, recent advancement in econometrics has proven that some TWFE DiD designs are prone to produce biased estimates. We use JOM studies to highlight this issue again. In writing this article (May 2023), we surveyed all research articles published in JOM for all issues in 2022 (9 issues) and 2023 (3 issues). The percentage of research articles adopting DiD or event study methodology among all published research articles in JOM is 26.7% (8 out of 30) and 26.3% (5 out of 19) for 2022 and 2023 respectively (Appendix 2). In other words, for every 4 research articles published in JOM, 1 of them uses DiD or event study methodology. For all the 13 DiD articles in JOM, 6 (46%) studies with staggered design are susceptible to biased and misleading research results (Appendix 1). Given the high percentage of DiD adoption and given the even higher percentage of DiD studies that are susceptible to biased results, out study provides a useful guidance for OM researchers to correctly assess treatment effect and draw statistical inference.

Our study makes several contributions to the OM literature. First, our study echoes the recent field experiment review (Gao et al. 2023) and the recent replication project of lab-controlled experiment (Davis et al. 2023). While Gao et al. (2023) reviewed field experiment design and Davis et al (2023) tested the

research reliability by replicating laboratory experiments, we examine research validity of quasi-experiment design. Therefore, our research is complementary to Gao et al. (2023) and Davis et al. (2023), contributing to experiment design as a whole in the area of operations management. Second, we systematically review different DiD designs, provide fundamental statistical interpretation of each design, theoretically discuss TWFE validity/pitfalls of each design, and empirically provide an example to illustrate TWFE validity/pitfalls for each design. Despite the popularity of TWFE DiD design in the OM field and despite the scattered discussion of the potential pitfalls of TWFE DiD designs (Shang and Rönkkö 2022, Mithas et al. 2022, Barrios et al. 2022), a systematic review of TWFE DiD design and its associated problems in the OM field is missing. Our study fills this void by answering the two fundamental questions of why some TWFE DiD designs are problematic and how to solve the problems using both simulation and published studies as examples. Therefore, researchers can use our study as a reference to increase research stringency for their DiD designs. To this end, our study also echoes the recent trend of conducting responsible operations management research advocated in the OM field (MSOM 2020 Issue 6, POM upcoming special issue). We call for researchers to conduct responsible research when drawing statistical inference from assessing treatment effect, especially when there is variation in treatment timing. Third, all the 51 studies we surveyed investigate either static and/or dynamic effect. Instantaneous effect of policy change has been largely ignored. However, instantaneous effect may provide a more meaningful insight for policymakers to make necessary decisions, such as using learnings from the instantaneous effect of Covid-19 to battle the immediate outbreak of future similar pandemics. Our study explains the statistical foundations of instantaneous design and proposes a relevant estimator to correctly estimate the instantaneous effect. We encourage researchers to estimate instantaneous effect using the latest estimator from the econometrics field (de Chaisemartin and D'Haultfoeuille 2020) as the extant practice to estimate the instantaneous effect is contaminated and produces biases estimates. Fourth, we use simulation to show that the popular method of using event study to test the parallel trend assumption is problematic in a staggered design. The reason is that when estimating leads (and lags) in a staggered design, the estimates of the coefficients on the leads (and lags) are contaminated by the ATTs from all relative-time periods and all treatment cohorts (Sun and Abraham 2021). Therefore, the treatment effects estimated do not reflect the actual ATTs for each relative time periods. We recommend researchers to adopt alternatives recommended to test the parallel trend assumption to increase research validity.

We next recommend some best practices to overcome these pitfalls. First, in a staggered design where the estimate for both the static effect and the dynamic effect is biased, we recommend researchers to conduct the following. 1) If the data is a balanced panel, researchers can decompose the weights assigned by TWFE to different $2\times2$ groups using Goodman-Bacon (2021) decomposition method (Section 4.1.4). Then, researchers can detect the problematic weight, i.e., the weight for Later treated VS Early treated groups

where the already-treated units were used as effective controls. The higher the weight associated with this problematic 2×2 group, the more biased the TWFE estimates. If the data is unbalanced, researcher can try to create a balanced version (if not losing too much information in the original data) and run this decomposition test. 2) As was explained in the example of Section 4.1.4, increasing the size of never-treated units can help reduce the bias associated with TWFE estimates. We encourage researchers to increase the size of never-treated units in the research design whenever possible. 3) To avoid the biased estimates associated with TWFE, we highly recommend researchers to use the latest estimators, such as Callaway and Sant'Anna (2021), Sun and Abraham (2021), or stacked regression when their DiD design is staggered. If researchers continue to report TWFE results from a traditional event study, we encourage researchers to at least use any of the recommended alternatives as a robustness test. Second, for the parallel trend assumption test, researchers can use either Callaway and Sant'Anna (2021) or Sun and Abraham (2021) to estimate the coefficients on the leads without significantly changing the data structure. The two estimators differ in the ease of incorporating covariates with Callaway and Sant'Anna (2021) being easier to implement. Both estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021) require the creation of a new treatment identifier (the first occasion different groups were treated) without making changes to the original data. Another alternative to test the parallel trend assumption is stacked regression. However, stacked regression requires researchers to create event-specific clean 2×2 datasets including outcome variable and covariates for both the treated units and other clean control units (i.e., already-treated units should not be used as effective controls). These clean 2×2 datasets are then stacked together to be estimated using TWFE. Therefore, stacked regression necessitates extra data reconstruction work for researchers. Third, we call for researchers to share data and code wherever possible to increase research transparency and reliability as none of the 51 studies in the three top OM journals provide any data or code.

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 clearly mention the time frame for the treatment effect. Only 5 out of the 51 studies clearly mentioned the time frame during which the effect takes place 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 effect time window. 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 of 27 months after policy implementation. To avoid confusion as well as to increase stringency when presenting research findings, we encourage researchers to clearly state the time window for the treatment effects. This also matters greatly for practitioners and policymakers as merely presenting a single percentage of decrease or increase of treatment effect without mentioning the time window does not provide much guidance for policy making.

Lastly, we want to reiterate that propensity score matching (PSM), coarsened exact matching (CEM), or synthetic control method does not solve the problem of biased estimates associated with TWFE regression in assessing treatment effect 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.

In sum, DiD is the workhorse adopted by empirical OM researchers to assess treatment effect. Despite some research has highlighted the potential drawbacks of using TWFE regression to estimate the treatment effect (Barrios et al. 2022), a tremendous amount of research still publish using TWFE regression regardless of which DiD design was applied, potentially providing misleading or biased research results, especially when there is variation in treatment timing. In the wake of conducting responsible research in operations management field, we accordingly call for responsible research in drawing causal inference from DiD design in operations management.

## REFERENCES


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.

Avramidis P, Mylonopoulos N, Pennacchi GG (2022) The role of marketplace lending in credit markets: Evidence from bank mergers. *Management Sci.* 68(4):3090-3111.

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.

Bliese PD, Lang JW (2016) Understanding relative and absolute change in discontinuous growth models: Coding alternatives and implications for hypothesis testing. *Organ. Res. Methods* 19(4):562–592.

Borusyak K, Jaravel X, Spiess J (2021) Revisiting event study designs: Robust and efficient estimation. Working Paper 1-85.

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.

Cengiz D, Dube A, Lindner A, Zipperer B (2019) The effect of minimum wages on low-wage jobs. *Q. J. Econ.* 134(3):1405-1454.

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.

Cheng S, Lin P, Tan Y, Zhang Y (2023) "High" innovators? Marijuana legalization and regional innovation. *Production Oper. Management* 32(3):685-703.

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.

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.

Frohlich MT, Robb Dixon J (2006) Reflections on replication in OM research and this special issue. *J. Oper. Management* 24(6):865-867.

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.

Kim YH, Henderson D (2015) Financial benefits and risks of dependency in triadic supply chain relationships. *J. Oper. Management* 36(1):115-129.

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.

Pagell M (2021) Replication without repeating ourselves: Addressing the replication crisis in operations and supply chain management research. *J. Oper. Management* 67(1):105-115.

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

Peinkofer ST, Jin YH (2023) The impact of order fulfillment information disclosure on consequences of deceptive counterfeits. *Production Oper. Management* 32(1):237-260.

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.

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 (2020) Pre-test with caution: Event-study estimates after testing for parallel trends. Working Paper 1–84.

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 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.

Webb E, Hernández-Quevedo C, Williams G, Scarpetti G, Reed S, Panteli D (2022) Providing health services effectively during the first wave of COVID-19: A cross-country comparison on planning services, managing cases, and maintaining essential services. *Health Policy* 126(5):382-390.

WHO 2023: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

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

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% |

**RESPONSE LETTER**

"Reject and Resubmit" of JOM Manuscript ID **JOOM-22-0501**

Dear Editorial Board,

Thank you for providing us with the opportunity to resubmit our manuscript (JOOM-22-0501). In the original manuscript, we aim to use U.S. airline mergers as an example to introduce the latest econometrics estimators to assess the treatment effect in quasi-experiment design into the OM field.

After carefully studying the feedback from reviewers and the editorial board, we reckon that mixing airline mergers with a methodological review of quasi-experiment design is not a perfect way to materialize our research objective. Therefore, in the resubmitted manuscript, we removed the content of U.S. airline mergers and purely focused on a methodological review of quasi-experiment design in the OM field by explaining the "why" (why the biases exist in the design) and "how" (how to fix the biases). As we removed the airline merger content from the original manuscript, some of the reviewer's comments about U.S. airline mergers will not be reflected in the current re-submission. However, we took all the relevant comments about airline mergers and relevant changes were accordingly made in the other manuscript where we continue examining U.S. airline mergers.

Overall, we consider our re-submitted manuscript is a better fit for the "review" category of JOM empirical method department. To make this response document easier to read, we highlighted the concerns from each reviewer in green, pasted them in this document, and replied in blue to those green highlighted concerns. Hopefully, this helps to pinpoint the changes/revisions faster.

# REVIEWER 1

(Titled: *Review Report for "The Effect of Mergers on Airline On-Time Performance: Assessing Instantaneous, Static, and Dynamic Effects with Variation in Treatment Timing*", **two pages**)

**MAJOR CONCERNS:**

1. When I read the abstract saying the advanced new methodology was applied, I was initially very excited since this would be a nice contribution. I expected the authors to propose a new type of modeling or at least a variant of the current econometric approach. I was then surprised that the paper just applied the latest approach in the econometrics field and compared it against the existing approach.

Response: We thank the reviewer for this extremely useful comment. Our resubmitted version restructures the manuscript to a systematic review of TWFE DID in quasi-experiment design in the OM field to provide a reference for researchers to correctly assess the treatment effect, similar to Gao et al. (2023). As we are doing a "review" and recommendation, we do not focus on developing new DID estimators.

2. The authors need to add clear explanations and introductions to the latest econometric approach. For example, what are the key assumptions that hold for each approach? How are these assumptions addressed in your dataset? Since this paper highlights introducing the latest approach, better clear explanations are needed. It would be nice if the authors provided the table, including when/which methods are used in a certain case, with necessary assumptions.

Response: In the revised version, clear explanation and introduction of each DID design were added in the beginning of each section (Section 2 to Section 5). We also explained which design is used in what kind of situations, what is the current practice to estimate that design, what are the potential pitfalls, and how to solve these pitfalls. We also summarized all the five designs in Section 4.3 and made recommendations in Section 7 on how to move forward in quasi-experiment design.

In Section 6, we specifically elaborate on the parallel trend assumption that is paramount to all DID designs, explaining why the current practice to test this assumption is problematic and how to fix it.

3. On page 13, the authors adopted the new estimator $DID_M$ for their analysis since deChaisemartin and D'Haultfoeuille (2020) showed that it is an unbiased and consistent estimator that is robust to treatment effect heterogeneity across groups and time periods in both a sharp design and a staggered design. I am not fully persuaded that its unbiasedness and consistency still hold in your data.

Response: Not relevant anymore as we did not use U.S. airline merger data in the current resubmission. However, we added 20 never-treated carriers in the other manuscript where we continue studying U.S. airline mergers to address this issue.

4. On page 20, No anticipation assumptions: "reviewing the relevant news bulletins related to each merger, we did not find…" This is not informative enough. What have you reviewed specifically? What if the relevant news bulletins might not display that carriers have prior private knowledge about the future treatment path?

Response: Not relevant anymore as we did not use U.S. airline merger data in the current resubmission. However, we did take this advice and added an Appendix listing out all sources of news bulletins in the other manuscript where we continue examining U.S. airline mergers.

5. Based on the latest approach, Hypothesis 1 is supported, while Hypothesis 2 and Hypothesis 3 are not supported. Would you provide any underlying reasons for these results? Especially it would be interesting for me to know why the U-shaped curve for OTP following mergers is not theoretically supported in Hypothesis 3.

Response: Not relevant anymore as we did not use U.S. airline merger data in the current resubmission. However, we did take this advice and explained theoretically about the U-shape performance curve in the other manuscript.

6. Some of the equations and notations could be confusing to a reader. I was frequently lost when trying to read sections 4, 5, and 6 so the exposition of this material needs to be improved. For example,

Response: In the current resubmission, we made sure that each notation, when it first appeared in the article, was clearly explained. We also explained all other relevant equations in as many details as possible. Since we did not use U.S. airline merger data, the bullet points under this category are not relevant anymore as those bullet points are about the analysis of the merger effect. But we sincerely thank the reviewer for pointing it out. We also took this advice and revised in as many details as possible in the other manuscript where the focus is on U.S. airline mergers.

# REVIEWER 2

(Titled: *A Review for "The Effect of Mergers on Airline On-Time Performance: Assessing Instantaneous, Static, and Dynamic Effects with Variation in Treatment Timing*", **six pages**)

**1. Theoretical Contribution**

1.1 The authors claimed that the study is the first to simultaneously examine three treatment effects (i.e., instantaneous, static, and dynamic effects). I'm confused if this could be a theoretical contribution since the authors do not specify what theory has been contributed to the study. To my understanding, the "first-to-examine" itself should not be a theoretical contribution. It will be helpful if the authors elaborate on why these first examined treatment effects matter theoretically (e.g., what O.M. theory it contributes to and how it adds to the previous studies regarding such theory by using the new method). Further, the empirical analyses of the dynamic treatment suffer from some potential flaws (more on this later), which further undermines the potential contribution can be made by the study.

Response: As we have removed U.S. airline mergers from our resubmission, this concern is not relevant anymore. However, this is a very constructive feedback and we took the advice and made sure we highlight the contribution of the current manuscript, which is reflected in Section 7. In addition, we also took the advice and strengthened the contribution section in the other manuscript where U.S. airline merger is the focus.

1.2 A minor point, I'm not sure if we can call instantaneous, static, and dynamic effects as "three treatment effects", since the treatment (merger events) are consistent throughout the analyses, while just measured in different specifications/estimators. In my view, it seems more of "three measures of the treatment effects."

Additionally, the authors also stressed that the study included a greater number of merger events than previous studies (p.5). Again, what does this better data coverage lead to in terms of theoretical contribution? Using a dataset with better quality/coverage itself should not be a contribution. I think the authors should focus more on how it will provide deeper or contrasting insights compared with previous studies.

Response: We agree with the reviewer on the "three measures of treatment effects". We accordingly revised and reflected the changes in our resubmission. Since we do not use U.S. airline data anymore, "deeper and contrasting insights" is not relevant for the current study. However, we made sure that the current resubmission contributes to the OM field by providing deeper understanding of quasi-experiment design (Section 2 – 5, Section 7). In addition, we also took this advice and made sure deeper and contrasting insights were reported in the other manuscript where we continue to investigate U.S. airline mergers.

1.3 Since the study aimed to contribute to the O.M. field, it should be grounded and positioned against O.M. literature. However, O.M. literature is rarely cited in the study: only two management science papers are cited, which are not even from the O.M. department. Instead, I find the majority of the references are from Organization Sciences and Strategic Management Journal. The stream of O.M. literature, particularly those focusing on the airline industry (e.g., Deshpande and Arıkan 2012, and Atkinson et al. 2016), has been completely ignored by the study, which I believe will considerably hinder the potential theoretical contribution of this study.

Response: These advices were reflected in the other manuscript where we continue to examine U.S. airline mergers, i.e., we grounded our work in the airline merger literature. However, we also took the advice and applied it in the current version where we ground our study in three top OM journals by surveying all recent quasi-experiment designs in the OM field.

**2. Methodological Contribution**

2.1 First, many previous studies in the O.M. field deal with the empirical setting where treatments did not occur at the same time period (e.g., Dong et al. 2018, Cui et al. 2019, Dhanorkar 2019, and Calvo et al. 2020). Though the previous study used different methods, such an empirical setting is nothing new. The "dynamic effects" examined in this study (Equation 6 and Figure 3) have a long track record of application in the O.M. field (e.g., Dong et al. 2018, Cui et al., 2019, Dhanorkar 2019, Li and Wu 2020, Cui et al. 2022), though it may be named differently and may be used for different purposes (e.g., parallel trends assumption checking). Therefore, a potential contribution that this study could make can only be in the new method (Callaway and Sant'Anna 2021). However, the method is not appropriately applied in the study, which will lead to the next point.

Response: We totally agree that event study/dynamic effect has a long track record in OM field. However, our main purpose is to remind researchers of the potential pitfalls associated with event study when it is

used to estimate treatment effects and/or to test the parallel trend assumption. Detailed explanations are in Section 3.2, 4.2, and Section 6 in the current resubmission.

In the original submission, we aim to examine time effects, not group effects of mergers. However, since we did not use U.S. airline merger data anymore, this concern is not relevant for the current resubmission. But we did take this advice and accordingly made changes in the other manuscript where we continue examining U.S. Airline mergers.

2.2 Second, in the staggered DID setting (variation in the treatment timing), there are types of heterogeneity in the treatment effects: 1) heterogeneous treatment effects across events (given the timeline, how the treatment effects differ depending on the subgroups that the event occurs) and 2) heterogeneous treatment effects across time periods (given all the events, how the treatment effects vary as longer exposure to the treatment). The two heterogeneity in the treatment effects has been mixed up with each other in the study. The contribution claim by the authors, as well as the strength of Callaway and Sant'Anna (2021), is the former one, which is unique to the varying-treatment-timing settings. The analysis in the paper, however, is really the latter one, where the authors estimated how the treatment effect on OTP varies as each time period after the merge events. This really damages the contribution of the study because the latter is not unique in the varying-treatment-timing setting but can also be estimated with a homogeneous-treatment-timg setting. Further, the estimation of treatment effects based on the relative time periods to the event also has been widely used in the O.M. field (e.g., Dhanorkar 2019). Therefore, I'm not sure if the current analyses make a contribution methodologically. Potentially, the authors can leverage the true strength of the Callaway and Sant'Anna (2021) estimator and explore the heterogeneity across events (how the treatment effects differ based on the different timing of merger events).

Response:

1) In the original study, we were interested to estimate time effects, not group effects. However, we find this advice very useful and accordingly conducted a group effect analysis in the other manuscript where U.S. airline mergers is the focus.

2) It is true that event study can be estimated in both a staggered design (variation in treatment timing ) and a sharp design (no variation in treatment timing). However, the former is prone to biased estimates while the latter is not. We expand on this difference in Section 3.2 and Section 4.2 in our current resubmission.

2.3 Third, though the methodological contribution is positioned against the application of the Callaway and Sant'Anna (2021) estimator, the application of the estimator in the analysis shows no significant results and thus does not offer any insight (for all 20 periods in the post-treatment periods, none of the estimates is statistically significant). By the authors' design, Callaway and Sant'Anna (2021) is used "to illustrate the potential differences between TWFE and the new estimator." (p.23). The insignificant results fail to demonstrate the difference. The authors try to claim that the insignificant results are the new insights offered by the new estimator. I'm not sure it is the case because "not finding significant effects" is completely different from "there is no significant effects," where the study needed the latter to support the contribution claimed but what is suggested by the estimation results is the other one.

Response: 1) We did not study U.S. airline merger in the current resubmission. Therefore, "does not offer any insight" for the airline industry is not irrelevant for the current version anymore. However, we took this advice and examined financial performance instead of operational performance in the other study to offer more meaningful insights for the airline industry. 2) The potential difference between TWFE and newer estimator is reflected in Section 4.1.4, 4.2.4, and Section 6 in the current resubmission.

2.4 Fourth, the Callaway and Sant'Anna (2021) estimator results are not comparable with the TWFE results. In the TWFE event study estimation on p.21, the authors included the dummies for each period to capture the difference between the groups at each time period, where t=-1 and t=-2 are left out to avoid multicollinearity (thus will serve as the reference group). Thus, the point estimates ($\widehat{\mu_l}$) is interpreted as how the difference between the treatment and control groups at

Response: Not relevant anymore as we did not use U.S. airline merger data in the current version. However, we did a similar comparison between TWFE, Callaway and Sant'Anna (2021), and Stacked regression in Section 6.2 and 6.3 in the current resubmission.

**3. Managerial Contribution**

The managerial implications offered in the study are disconnected from the results from the empirical analyses and seem too general in wording, such as "strength their effort" and "work harder," and thus lack insight.

Response: Although not relevant anymore, we find this advice very pertinent and have accordingly made changes in the other manuscript where we come forward with detailed managerial recommendations regarding U.S. airline mergers.

# REVIEWER 3

(no title)

**Major Comments**

1. We understand that the authors believe that the commonly used DID methods have certain biases, but we believe that the authors do not fully explain the shortcomings of the old methods and the details and advantages of the new methods, and even lack the necessary clarity in many places. For example, the authors do not explain clearly how they constructed the treatment and control groups on a dataset of 7 samples. As the authors said, they believe that there is a fundamental difference between a carrier that has not merged and a carrier that has undergone merge, so they do not include any untreated sample. On such a fully treated sample, user stated "the average of treatment effects at the time when all carriers start receiving treatments (i.e., started to merge with another carrier) by comparing carriers whose treatment status has changed with those carriers whose treatment status is unchanged between t–1 and t". Suppose the "unchanged" here is direct to carriers that have not been merged, then who is the control group for the last sample to experience the merge? Or does "unchanged" mean that the state has not changed (that is, if the merged state remains unchanged, it can also become unchanged)?

Response: 1) In the current version, we break down quasi-experiment design into 5 different categories to fully explain the shortcomings of each DID design, how the new estimators help to overcome these shortcomings, and use simulation and/or replication study to showcase the validity of each design. 2) How the treatment and control group were constructed is not relevant anymore as we did not use U.S. airline mergers data in the current resubmission. However, we find this advice very useful and accordingly added 20 carriers who have not experienced mergers as the control group in the other manuscript where we continue to examine airline mergers to address this concern. We also explained in details about how data was structured in the other manuscript.

2. de Chaisemartin and D'Haultfoeuille (2020) proposed that "*The first point of the stable groups assumption requires that between each pair of consecutive time periods, if there is a group that switches from being untreated to treated, then there should be another group that remains untreated at both dates. The second point requires that between each pair of consecutive time periods, if there is a group that switches from being treated to untreated, then there should be another group that remains treated at both dates.*" Given the context of airline merger, the carrier's status become treated in the quarter when the merger event took place and in all the quarters thereafter. Therefore, there is no carrier can switch from treated to untreated, stable groups should only be those who keep untreated between t–1 and t. If carriers which keep treated in t and t-1 are considered, there will be errors in the estimation results. The authors should clearly state how they design the groups and give examples if possible.

Response: This concern is not relevant to the current resubmission. However, in the other manuscript, we added never-treated units (20 carriers that have not experienced mergers) so this concern was addressed.

3. As shown in table 1, the Merger Announcement Date is usually half a year before the Completion Date, but the author did not consider the potential impact of the Merger Announcement in this paper. Such announcements mean that the impact of the merger may be advanced. Therefore, I suggest that the author should consider the impact of announcement when designing the empirical model, to analyze the impact of merger more accurately.

Response: This comment is not relevant to the current resubmission as we did not use U.S. airline data in the current version. However, we find this advice helpful and did run a placebo test by using merger announcement date to make sure all our hypotheses still stand in the other manuscript.

4. Most of the empirical part of this study is the introduction to some latest econometric methods, and the author has hardly improved and innovated the methods. In addition, the authors do not explicitly describe how the advanced econometric methods improve the estimation errors in traditional studies, only a brief description of the method. Therefore, from the perspective of method, this paper does not meet the requirements of this journal

Response: Our initial research objective was to conduct a methodological review of DID in quasi-experiment design using U.S. airliner mergers as an example. Therefore, we do not intend to improve or innovate the current methods. However, this advice is very useful in helping us to reorganize the writing of the current version by elaborating all the 5 different DID designs, their validity, and recommendations on how to use each design. We sincerely thank this reviewer for the helpful comment.

**Minor Comments**

1. In "Data" section, the authors should provide some descriptive statistics include maximum value, minimum value, number of observations and so on. Although the relevant variables are described in detail in this section, there are still some unclear description. A table with summary statistics is a good way to solve this problem. If possible, it would be helpful for the authors to report some Example Data, like Table 10 and Table 12.

Response: We added example data table in the current study (Table 5, Table 6) wherever possible. We also took this advice and added descriptive statistics table and example data table in the other manuscript.

2. "4.1 Instantaneous Effect of Mergers on OTP" section propose the staggered design by reviewing advanced econometric models. I suggest that the authors can use some diagrams to show the differences between groups.

Response: We took this advice and used appropriate diagrams in the current study (Figure 1, 2, 3 for example).

3. When reporting the analysis results, the author should include more information in the table. Specifically, a single table can contain the results of multiple analyses, which makes it more concise and facilitates comparison between different methods. The table should contain information representing the differences between different methods.

Response: We took this advice and the changes were reflected in Table 4.

4. Authors also need to keep consistency across tables. For example, Table 4 contains N value and t-stat, but some of the other tables have this information.

Response: Changes were reflected in all the Tables in the current resubmission as well as in the other manuscript where we continue studying U.S. airline mergers.