ONLINE SUPPLEMENT

ONLINE SUPPLEMENT 1: TWO GROUP TWO TIME PERIOD

Throughout this supplement, 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.

1.1 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):

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

A breakdown of the different effects is summarized in Table 1 (Wooldridge 2010). β_1 captures the potential differences between the control and treatment groups before the policy intervention while δ_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". δ_1 is commonly known as the treatment effect on the treated. Throughout this article, we use δ or δ^{DD} to indicate the treatment effect.

Table 1 Two Groups Two Time Periods

	Before Treatment	After Treatment	After – Before
Control Group A	β_0	$\beta_0 + \delta_0$	δ_0
Treatment Group B	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
B – A	β_1	$\beta_1 + \delta_1$	δ_1

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

Ordinary least squares (OLS) is the most commonly used estimator to estimate δ_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.

1.2 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.

ONLINE SUPPLEMENT 2: 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.

2.1 Single Treatment Static Effect

2.1.1 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.

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

α_i is the unit fixed effect, λ_t is the time fixed effect, and ε_{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 δ^{DD}, which is typically known as the average treatment effects on

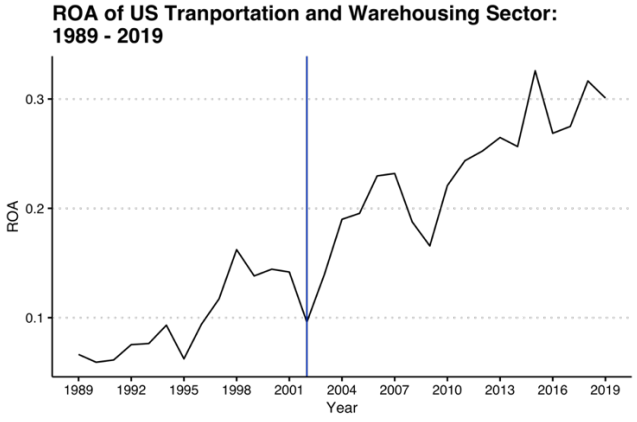
the treated (ATT), i.e., the overall treatment effect for all units that have participated in the treatment up to a certain time period T.

2.1.2 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). Simulation is written as:

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

$$\delta = 0.05 \cdot \hat{s}(Y_{it}(G_\infty))$$

$$D_{it} = \left\{ \begin{array}{ll} 0, & \text{for } t < 2002 \\ 0, & \text{with probability 0.5, for } t \geq 2002 \\ 1, & \text{with probability 0.5, for } t \geq 2002 \end{array} \right\}$$

$$Y_{it}(G_{2002}) = Y_{it}(G_\infty) + D_{it} \cdot \delta \cdot (t + 1 - 2002)$$

Here the group timing is the same for everyone and coincides with year 2002. There are only 2 groups, G_{2002} treated at time t = 2002 and G_∞, never treated. Group participation is randomized with a probability equal to 50%. $CATT(G_{2002}, t) = \delta \cdot (t + 1 - 2002)$, which corresponds to an ATT growing linearly with the number of periods of exposure. Of course since G_∞ is never treated, its ATT is not identified.

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). We then run our simulation 1000 times and use Equation 3 to estimate the treatment effect. We compute actual firm average treatment effect and actual observation average treatment effect following recent econometrics literature (Callaway and Sant'Anna 2021, Baker et al. 2022). For Firm average treatment effect, we first compute average ATT for each firm then average these ATTs. For observation average treatment effect, we just compute average ATT across all treatment observations. We plot the simulation results in Figure 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

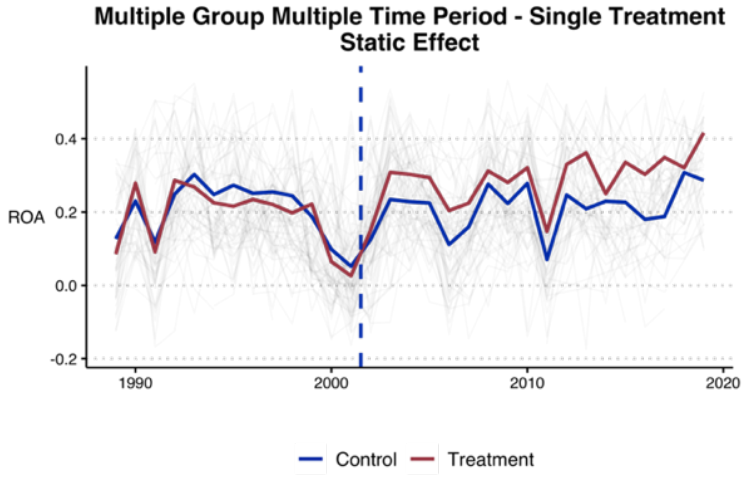
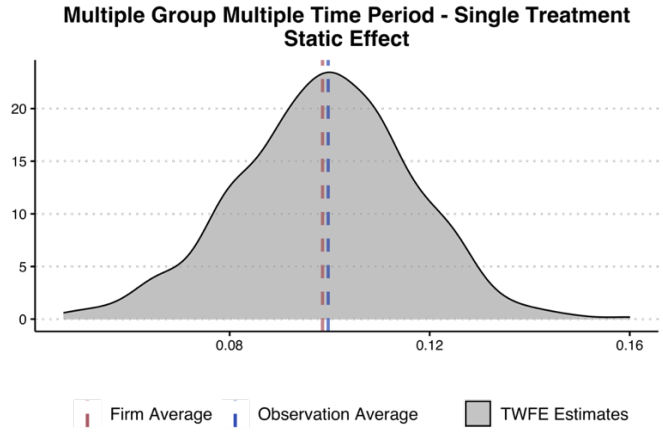


Figure 3 Simulation Result – Single Treatment Static Effect



2.2 Single Treatment Dynamic Effect

2.2.1 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:

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

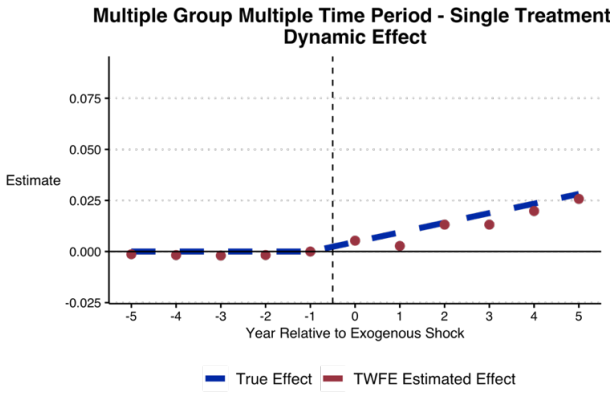
The interpretation of α_i and λt remains the same: they are vectors of individual fixed effects and time fixed effects respectively. $\boldsymbol{X}_{it}$ is the vector of time-varying control variables and ε_{it} is the error term. Instead of using a single binary indicator of D_{it} in Equation 3, 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 $t - l$ 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 μ_ls 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).

2.2.2 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 $t - l$ 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. We observe that for both the pre-periods and the post-periods, TWFE estimates (red line) are closely aligned with true effect (blue 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



ONLINE SUPPLEMENT 3: SATURATED VS NON-SATURATED REGRESSION

A regression is saturated or fully saturated if there are as many independent variable terms included in the regression as there are potential values for the regressors. In the most basic example, suppose that there are two regressors $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1\}$. There are 4 possible values of the covariates since $(X_1, X_2) \in (0, 0), (0, 1), (1, 0), (1, 1)$. A fully saturated regression of an outcome variable Y on X_1 and X_2 is a regression of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} \cdot X_{2i} + \epsilon_i \quad (1)$$

Conversely a non-saturated regression may omit some of these terms, say omitting the interaction term.

$$Y_i = \widetilde{\beta_0} + \widetilde{\beta_1} X_{1i} + \widetilde{\beta_2} X_{2i} + \widetilde{\epsilon_i} \quad (2)$$

A fully saturated OLS regression like in Equation (1) with finitely supported covariates always estimates the conditional expectation function nonparametrically (in this case $\mathbb{E}[Y|X_1,X_2]$). To see this, note that in the population regression:

$$\mathbb{E}[Y|X_1 = 0, X_2 = 0] = \beta_0$$

$$\mathbb{E}[Y|X_1 = 1, X_2 = 0] = \beta_0 + \beta_1$$

$$\mathbb{E}[Y|X_1 = 0, X_2 = 1] = \beta_0 + \beta_2$$

$$\mathbb{E}[Y|X_1 = 1, X_2 = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

Because $\mathbb{E}[Y|X_1, X_2]$ can only take 4 values and there are 4 parameters, the functional form of the model cannot be misspecified. This is exactly what happens in the 2 periods, 2 groups DiD, for which the TWFE is always valid because it's not imposing any functional form restriction. On the other hand the regression in Equation (2) is not fully saturated and the $\tilde{\beta}$ estimate a weighted average of the underlying conditional expectation.

In the multi-period DiD with staggered adoption, the fully saturated regression allows each groups and each time period since treatment to have potentially heterogeneous dynamic treatment effects (hence $ATT(g,t)$). The static and dynamic specification are non-fully saturated regressions, (they are missing some of the potential interactions) hence the weighted average problem detailed out, for example, by Sun and Abraham (2021).

ONLINE SUPPLEMENT 4: Rationale of Staggered Regression

We use Wang et al. (2022) to illustrate how to construct a stacked regression. This online supplement explains the concepts while R code gives the step-by-step details about how to construct a stacked regression. Table 2 is taken from the manuscript Section 7.2 using Wang et al. (2022) as an example.

Wang et al. (2022) compiles data from 2002 to 2017 to investigate how the introduction of new airline routes impact volume of shared kidneys. Not every year sees new routes so Table 2 only summarizes those years when new routes were introduced. The problem with TWFE is that, say, if the 32 new routes in 2003 were used to compared to the 96 new routes in 2013, it will be problematic; however, if the 32 new routes in 2003 were used to compared to the 513 existing routes, it will be a valid design, i.e., already-treated units cannot serve as clean controls.

Table 2 Summary of New Routes in Wang et al. (2022)

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

The key idea is that when we create comparison groups for each year of new routes introduced, already-treated units cannot be used as the comparison group. Never-treated groups (the 513 routes in Wang et al. 2022) and/or not-yet-treated groups (there are no not-yet-treated groups, unfortunately, in Wang et al. 2022) are the eligible comparisons. So, constructing stacked regression is to literally create a separate dataset for each single group in Table 2 ("year" in Wang et al.'s 2022 case) by only using the 513 never-treated units and exclude any already-treated units. In other cases, researchers can include both never-treated and/or not-yet-treated groups (Figure 2 in our manuscript).

Following this logic, we will have 12 separate datasets (as we have 12 years when new routes were introduced from Table 2). Each dataset will have its unique dataset identifier. Then, these 12 datasets will be stacked together as one single dataset where TWFE estimator is applied. Attached R code explains the step-by-step construction process. We recommend readers to run the code to see how the data was created in practice.