THE IMPACT OF MERGERS ON ON-TIME PERFORMANCE

A Re-examination in the U.S. Airline Industry

ABSTRACT

We re-examine merger impact on on-time performance (OTP) in the U.S. airline industry as related research has demonstrated two deficiencies. First, a staggering 71.9% of non-carrier-induced delays, such as weather delays, were used to assess merger impact on OTP, which is highly likely to produce biased causal inferences. Second, all related research adopted two-way-fixed-effects (TWFE) difference-in-difference (DiD) regression to assess merger impact on OTP using multiple time periods of data (i.e., longitudinal data with more than two time periods) with variation in treatment timing (i.e., merger event happened at different times for different airlines), under the situation of which TWFE has been proved to produce biased estimates. By narrowing down OTP to carrier-induced delays only and by adopting the latest advancement in econometrics, we find that mergers worsen carrier's OTP up to the first quarter in year five post-merger, different from findings in the extant literature. In a simulation analysis using Compustat data, we show that our methodological approach accurately estimates both the pretend and the true effect while TWFE fails to estimate both. To increase research rigor, we call for researchers to focus on carrier-induced delays to evaluate policy impact on airlines' operational performance as well as using the latest econometrics advancement to assess treatment effect in operations studies when there are multiple time periods with variation in treatment timing.

Keywords: airline merger, on-time performance, staggered difference-in-difference, stacked regression

THE IMPACT OF MERGERS ON ON-TIME PERFORMANCE

A Re-examination in the U.S. Airline Industry

1. INTRODUCTION

Since the deregulation in 1978, the U.S. airline industry has experienced numerous mergers (Singal 1996a;1996b, Department of Transportation 2023). Accordingly, researchers have also conducted extensive research on the effect of U.S. airline mergers. However, the focus of U.S. airline mergers has largely been concerned with market competition such as fare and flight frequency or stock market response (see Appendix 1 for a detailed summary). Only recently have researchers started to examine the effect of mergers on service quality, such as on-time performance (OTP) (Steven et al. 2016, Prince and Simon 2017, Rupp and Tan 2019). Being one of the critical dimensions of air travel service quality, OTP provides practical guidance for travelers to make booking decisions and travel plans (McCartney 2010). Vaze et al. (2017) also called for more future research to continue exploring more service quality issues, such as OTP. Therefore, the overarching objective of the current study is to continue knowledge accumulation in the airline merger literature by re-examining the effect of mergers on OTP.

Among the few research that has investigated merger impact on OTP, different OTP measures have been compiled, such as on-time arrivals (Steven et al. 2016, Rupp and Tan 2019) and elapsed travel time (Prince and Simon 2017). Department of Transportation (DOT) reports a flight on time if it arrives within 15 minutes of its scheduled arrival time regardless of what has caused the delay – and this is exactly where the extant research demonstrated deficiencies. DOT reports five different causes of arrival delays, i.e., carrier delay, weather delay, national aviation system delay, security delay, and late arriving aircraft delay. When assessing policy impact on OTP, we reckon that only carrier delay, which is within carrier's control (DOT 2023), should be attributed to potential policy impact whilst extant airline merger research has included all five different arrival delays to compute OTP, which is mostly likely to yield biased causal inferences, especially given that carrier delay only accounts for 28.1% of all delays in our data. Nicolae et al. (2017) also reported that carrier delay accounts for 27.7% of total delays in their data. Therefore, including non-carrier-induced delay to assess policy impact on OTP seems biased. Our first research objective, accordingly, is to narrow down OTP to carrier-induced delay only to re-examine merger impact on OTP.

Additionally, the few research that has investigated merger effect on OTP (Steven et al. 2016, Prince and Simon 2017, Rupp and Tan 2019) only adopted binary measures, i.e., a dummy variable 1 was assigned to post-merger periods and 0 otherwise. A binary measure only captures the overall average effect of mergers for the entire post-merger periods and this average effect is accordingly named "static effect" in the literature (Sun and Abraham 2021). However, researchers and decision makers may be more interested to know how the effect of mergers unravels over time (Sun and Abraham 2021, Callaway and Sant'Anna

2021). The effect of mergers over time is referred to as "dynamic effect" or "event study" in the literature. Related airline merger research has not investigated the dynamic effect of mergers yet. Accordingly, the second objective of our study is to explore the dynamic effect of merges on OTP to unravel the effect of mergers over time, thus, providing a more nuanced picture of merger impact for airline managers.

Lastly, among the 31 studies we surveyed in airline merger literature (Appendix 1), 15 papers adopted two-way-fixed-effects (TWFE) difference-in-difference (DiD) regression and four papers used event study methodology (TWFE regression with leads and lags) to assess merger effects. TWFE DiD regression and event study have been the prevalent estimation method to draw causal inference on treatment effects in operations management (OM) field. However, the latest advancement in econometrics has shown that TWFE DiD regression and event study are prone to yield biased estimates and misleading inferences when there are multiple time periods with variation in treatment timing, also known as staggered DiD design (Goodman-Bacon 2021, Callaway and Sant'Anna 2021, Sun and Abraham 2021). All extant airline merger research falls within the category of multiple time periods (i.e., longitudinal data with more than two time periods) and variation in treatment timing (i.e., merger event happened at different times for different carriers), indicating potential biased estimates. Moreover, related research on merger impact on OTP has produced mixed findings. To this end, it is both necessary and interesting to adopt the latest econometrics development to revisit the effect of mergers on OTP to see if using the latest econometrics development can provide new findings. Therefore, our last and most important objective is to introduce the latest econometrics development to estimate unbiased treatment effects in a staggered DiD design, hoping to contribute to research rigor for the whole OM research community.

To examine our research objectives, we collect data from DOT and study seven recent U.S. airline mergers. Applying the latest development from econometrics field to build a stacked regression, we find different results regarding merger impact on OTP. First, contrary to Rupp and Tan (2019) who found that OTP improved in the immediate four quarters following mergers, our results indicate that OTP worsened in the first four quarters following mergers. Second, contrary to Prince and Simon (2017) whose findings suggest travel time was not impacted in 1-2 years post-merger and even improved in 3-5 years post-merger, we show that after a merger, OTP keeps deteriorating from year one up to the first quarter in year five post-merger. Lastly, our results extend Steven et al. (2016)'s findings in that Steven et al. (2016) found that OTP deteriorated in the first three years following mergers while our results indicate that the deterioration continues up to the first quarter of the fifth year following a merger. Additionally, we utilize the grace period allowed by DOT for the acquirer and the target to still report as two individual carriers to examine merger impact for both the acquirer and the target. Our results reveal that only the acquirer suffered from worsened OTP following a merger while the target airline did not see statistically significant impact on OTP post-merger.

Given the already mixed findings on the impact of airline mergers on OTP in the extant literature, we conduct two robustness tests and an additional simulation using Compustat data to make sure our research does not add another layer of "noise" to the literature. We first test the parallel trend assumption using the latest advancement in econometrics field, rather than using the conventional event study method to test the coefficients on the leads. Our results indicate that the parallel trend assumption holds in our data. Next, we adopt merger announcement date as a placebo test to see if merger announcement has any significant impact on OTP. Our results show that merger announcement does not have significant impact on OTP following mergers. Lastly, we retrieve data from Compustat to simulate a staggered DiD design and show that when TWFE regression was applied on the original data, TWFE failed to estimate the true effect. In addition, TWFE also incorrectly estimates a pretrend when there is no actual pretend. However, using stacked regression on a restructured stacked data, both the true pretrend and the true treatment effect were correctly estimated.

By examining the four research objectives, we contribute to knowledge accumulation theoretically, methodologically, and practically. Theoretically, by narrowing down OTP to carrier-induced delays only, we offer new and different findings compared with related airline merger literature, thus, extending the understanding of merger impact on OTP for both researchers and airline mangers. Our research also responds to calls to explore more recent U.S. airline mergers (Hüschelrath and Müllera 2014) and to investigate service quality issues post-merger (Vaze et al. 2017).

Methodologically, we introduce the latest econometrics development to assess treatment effects in a staggered DiD design to the OM field. Similar to other recent studies (Callaway and Sant'Anna 2021, Baker et al. 2022), our different findings, compared with related literature, show that traditional estimators (TWFE DiD regression and event study) may produce biased causal interpretations on treatment effects in a staggered DiD design. In addition, our simulation analysis also highlights the pitfalls of using event study to estimates the coefficients on the leads to test the parallel trend assumption. Given the widespread adoption of TWFE estimators in a staggered DiD design in OM field, our research echoes the recent advocate from Baker et al. (2022) that researchers should pay special attention to traditional TWFE estimators when drawing causal inference on treatment effects. We call for OM researchers to consider adopting the latest econometrics estimators to draw causal inferences from staggered DiD design as well as using latest advancement to test the parallel trend assumption.

Managerially, our research provides more accurate insights for both airline managers and airline policy makers as we narrow down OTP to carrier-induced delay only and we utilize the latest advancement of econometrics to estimate the treatment effects. For airline managers, our analysis shows that the deterioration of OTP continues up to the first quarter of year five post-merger, indicating that airline managers should treat post-merger operations integration as a 4-5 years of planning and scheduling, rather

than a short-term planning and scheduling. For airline policy makers, our result suggests that the Department of Justice may need to consider future airline mergers more carefully before granting approvals. One of the key reasons often cited by the Department of Justice (2019) to reject any airline merger is the potential service deterioration post-merger. Our results do suggest this concern is not unfounded, as we observe that OTP deteriorates into the fifth year following mergers.

2. LITERATURE REVIEW

2.1 Mergers in the U.S. Airline Industry – Background Information

Although airline mergers predate industry deregulation in 1978 (Lichtenberg and Kim 1989), a significant amount of U.S airline mergers occurred post-deregulation, which can be classified into two waves: an initial wave in the 1980s immediately following the deregulation, and a second wave beginning in the late 1990s. The first wave was characterized by a sharp increase in merger activity, with 27 mergers recorded from 1985 to 1988 (Singal 1996a; 1996b). This period featured two main types of mergers: those between small carriers, such as the Braniff–Florida Express merger in 1988; and those between mega and small carriers, such as the American Airlines and Air Cal merger in 1987. Additionally, this phase was also known for repeated mergers involving the same carrier within a short period of time, as seen in the case of Piedmont Airlines, which merged with Empire in 1986 and subsequently with US Air in 1988.

The second merger wave, beginning in the late 1990s, witnessed a decreased frequency of mergers with only 20 instances occurring between 1999 and 2019. Different from the first wave, the second wave started to see mergers among mega carriers, notably Delta–Northwest in 2008, United–Continental in 2010, and American–US Airways in 2013, resulting in the consolidation of the industry into three legacy mega-carriers: Delta, United, and American (DOT 2023).

2.1 Operational Performance Implication of Mergers in the U.S. Airline Industry

Responding to the call of examining more recent U.S. airline mergers (Hüschelrath and Müllera 2014) to provide more current managerial insights, we focus on seven recent U.S. airline mergers (more in Section 3.2). While research on U.S. airline mergers has examined various operational performance implications, such as OTP (Steven et al. 2016, Rupp and Tan 2019) and travel time (Prince and Simon 2017), there is still a need to re-explore this topic.

First, when examining merger impact on airline OTP, previous studies have unfortunately mixed merger-induced operational impact with other impacts that are outside the control of the merged carriers. Table 1 below illustrates this in detail. In Table 1, a United Airline aircraft with the tail number N343UA performed six flights on January 5, 2004. DOT measures delayed/on-time flights in minutes in two categories: departure delay and arrival delay. For departure delay, DOT does not break down the specific reasons (Nicolae et al. 2017). For example, on January 5 2004, N343UA flew from DEN to RNO and incurred a 24 minutes departure delay. The only information known is that on the route of DEN→RNO

operated by United Airlines on January 5 2004, there was a 24 minutes departure delay at the origin DEN. Was this 24 minutes departure delay caused by airlines or by airport or by weather? DOT unfortunately does not provide any further information (Nicolae et al. 2017). Therefore, using departure delay, even spill-adjusted departure delay (Nicolae et al. 2017), to measure merger impact on on-time performance cannot delineate merger-induced departure delays from other various delays, such as weather delays, which should not have anything to do with the merger event itself.

Table 1 Schedule of UA Aircraft with Tail Number N339UA on January 05, 2004

Flight Date	Carrier	Tail Number	Origin	Destination	Departure Delay (Min)	Arrival Delay (Min)	Carrier Delay (Min)	Weather Delay (Min)	NAS Delay (Min)	Security Delay (Min)	Late Arriving Aircraft Delay (Min)
1/5/04	UA	N343UA	BOS	ORD	-3	27	0	0	27	0	0
1/5/04	UA	N343UA	ORD	DTW	0	19	0	0	19	0	0
1/5/04	UA	N343UA	DTW	DEN	-6	17	0	0	17	0	0
1/5/04	UA	N343UA	DEN	RNO	24	33	17	0	9	0	7
1/5/04	UA	N343UA	RNO	SFO	26	20	0	0	0	0	20
1/5/04	UA	N343UA	SFO	PHX	5	15	0	0	15	0	0

However, for arrival delay which was frequently used to measure merger impact on OTP (Steven et al. 2016, Rupp and Tan 2019), DOT indeed breaks down arrival delay into five different categories: Arrival Delay = Carrier Delay + Weather Delay + National Aviation System (NAS) Delay + Security Delay + Late Arriving Aircraft Delay. For example, on the same DEN→RNO route on January 5 2024, there was a 33 minutes arrival delay at destination RNO: 33 minutes Arrival Delay = 17 minutes Carrier Delay + 9 minutes NAS Delay + 7 minutes Late Arriving Aircraft Delay. DOT reports a flight on-time if it arrives within 15 minutes of its scheduled arrival time. Therefore, the flight on the route DEN→RNO on January 5th 2024 would be reported as late, which is acceptable when included in the data to draw causal inference as carrier-induced delay itself is already 17 minutes on this route, more than the 15 minutes threshold. However, it would become questionable if another route BOS→ORD is included to assess merger impact on OTP. The reason is that the 27 minutes arrival delay on the route of BOS→ORD is purely caused by NAS delay, such as air traffic control, which should not have much to do with the merger event itself. An extreme case is weather delay. Nicolae et al. (2017) reported that weather delay accounts for 45.5% of all delays in their data while there are 42% weather delays in ours. Including the 40%+ weather delay to assess merger impact on OTP would most likely yield biased causal inferences as weather should have nothing to do with airline mergers. Similar argument could also be made with using travel time (elapsed time between scheduled departure time and actual arrival time, Prince and Simon 2017) as weather delay, NAS delay, and security delay, which have nothing to do with the merger event, also impact travel time.

Due to the above-discussed concerns, we elect to only use carrier delay as our on-time performance measure. DOT defines carrier delay as "the cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, etc.)". Among all categories of arrival delays reported by DOT, carrier delay should most accurately reflect potential merger impact that is within

carrier's control as crew problems, maintenance, and other operations integration issues have been widely reported to plague merged carriers for many years post-merger (Mouawad 2012, Josephs 2018).

Second, all previous research studying the effects of mergers on OTP adopts TWFE DiD regressions and presents mixed findings (Steven et al. 2016, Prince and Simon 2017, Rupp and Tan 2019), making further exploration important, especially given the recent findings from econometrics that TWFE DiD regressions yield biased estimates and misleading inference in a staggered DiD design (Callaway and Sant'Anna 2021, Sun and Abraham 2021). Steven et al. (2016) studied three U.S. domestic airline mergers, finding that mergers worsened on-time arrivals in a three-year post-merger window. Prince and Simon (2017) examined five U.S. domestic airline mergers and found that in the short run (i.e., 1-2 years post-merger), mergers do not impact consumer travel time and in the long run (i.e., 3-5 years post-merger), consumer even benefited from shorter travel time. Rupp and Tan (2019) investigated four U.S. domestic airline mergers, finding that OTP improved immediately in the four quarters following mergers. Given all these studies adopted TWFE DiD regression and given the known pitfalls associated with TWFE, we find it important to re-examine this topic to increase research rigor as well as to provide more accurate insights for managers and policymakers in the airline industry.

Third, among the few airline research that has investigated merger effect on OTP, either a binary measure was used, i.e., a dummy variable 1 was assigned to all post-merger periods and 0 otherwise (Steven et al. 2016); or a simplified dummy coding was used, such as two years post-merger was coded as one dummy and 3-5 years post-merger was coded as the second dummy (Prince and Simon 2017). Binary and dummy measures only capture the overall average effect of mergers for the entire post-merger periods and this average effect is accordingly called "static effect" in the literature (Sun and Abraham 2021). However, policymakers may be more interested to know how the effect of mergers evolves over time (Sun and Abraham 2021, Callaway and Sant'Anna 2021). The effect of mergers over time is referred to as "dynamic effect" or "event study" in OM literature. No prior airline merger research has investigated how merger effect on OTP unravels over time. Therefore, there is also another need to re-investigate this research topic.

3. DATA

3.1 Data Handling

We collect relevant data from the Department of Transportation (DOT). To avoid the impact of 9/11, DOT reporting format change in 2003, and the global pandemic starting from 2020, we construct a panel data from 2004Q1 to 2019Q4, consisting of 64 quarters in total. On-time performance data is available at flight-day-route-carrier level. During the 64 quarters, we observe 34,514,934 rows of data. Following extant airline literature (Prince and Simon 2017), we aggregate our analysis at carrier-route-quarter level. After the official merger completion date, DOT still allows a grace period for the acquirer and the target to report to DOT as two individual carriers till eventually reporting as one. At the route level, therefore,

necessary data cleaning is required. A route is defined as a one-way origin-destination pair in our study following the practice in related literature. Following Prince and Simon (2017), we handle three different route level scenarios for the two merged carriers. First, on those routes where both the acquirer and the target operate under their individual carrier brand both before and after merger, we keep those routes as they are as long as they still report to DOT as two individual carriers. This helps to estimate the merger impact for both the acquirer and the target. Second, on those routes where both the acquirer and the target operate under their individual brand before the merger but operate under the acquirer's brand post-merger, we assign these routes, both before and after merger, to the acquirer by taking a weighted average of the pre-merger data following Prince and Simon (2017). Third, on those routes where only the target operates before the merger but changed to operate under the acquirer's brand post-merger, we re-assign these routes, both before and after merger, to the target till the quarter when these the target eventually reported to DOT under the acquirer's name. After the three-step route reclassifying process, we have 27 carriers in our data, 467,481 carrier-route-quarter observations, and 10,007 unique origin-destination pairs. Table 2 summarizes the carriers in our data.

Table 2 Airlines in the Dataset

No.	Airline	First quarter in the sample	Last quarter in the sample	Total quarters in the sample
1	AIRTRAN	2004 Q1	2011 Q1	33
2	ALASKA	2004 Q1	2019 Q4	64
3	ALEEGIANT	2018 Q1	2019 Q4	8
4	AMERICA WEST	2004 Q1	2005 Q4	8
5	AMERICAN	2004 Q1	2019 Q4	64
6	ATA	2004 Q1	2006 Q4	12
7	ATLANTIC SOUTHEAST	2004 Q1	2011 Q4	32
8	COMAIR	2004 Q1	2010 Q4	28
9	CONTINENTAL	2004 Q1	2011 Q4	32
10	DELTA	2004 Q1	2019 Q4	64
11	ENDEAVOR	2010 Q4	2019 Q4	13
12	ENVOY	2004 Q1	2019 Q4	56
13	EXPRESSJET	2004 Q1	2019 Q4	62
14	FRONTIER	2005 Q2	2019 Q4	59
15	HAWAIIAN	2004 Q1	2019 Q4	64
16	INDEPENDENCE	2004 Q1	2005 Q4	8
17	JETBLUE	2004 Q1	2019 Q4	64
18	MESA	2006 Q1	2019 Q4	40
19	NORTHWEST	2004 Q1	2009 Q4	24
20	PSA	2018 Q1	2019 Q4	8
21	REPLUBLIC	2018 Q1	2019 Q4	8
22	SKYWEST	2004 Q1	2019 Q4	64
23	SOUTHWEST	2004 Q1	2019 Q4	64
24	SPIRIT	2015 Q1	2019 Q4	20
25	UNITED	2004 Q1	2019 Q4	64
26	US AIRWAYS	2004 Q1	2015 Q2	41
27	VIRGIN AMERICA	2012 Q1	2017 Q4	24

3.2 Measures

On-Time Performance

As discussed in the previous session, we elect to use carrier delay to measure the impact of airline mergers on OTP as only this measure is within carrier's control and is probably the best measure to reflect any impact that truly resulted from a merger between two carriers. We use carrier-induced delay in this article to refer to carrier delay to more specifically define the nature of delay.

Merger Event

To determine merger dates for the merged carriers, we review various news bulletins (Appendix 2), such as airline news releases and CNN, and identify seven recent mergers which occurred from 2008 to 2016 as shown in Table 3. After the official merger completion date, DOT allows a grace period during which the acquirer and the target can still report as two individual carriers. Table 3 shows that the acquirer and the target still reported to DOT as two individual carriers up to 10 quarters during the grace period, which enables us to estimate the impact of mergers on both the acquirer and the target.

Table 3 List of Merged Carriers

Carrier	Merger Announcement Date	Merger Completion Date	Merger Completion Quarter	Quarters Before Mergers – Acquirer	Quarters Before Mergers – Target	Quarters After Mergers – Acquirer	Quarters After Mergers – Target
Alaska/Virgin America	April 4, 2016	December 14, 2016	2016Q4	51	19	12	7
American/US Airways	February 14, 2013	December 9, 2013	2013Q4	39	39	24	0
Delta/Northwest	April 14, 2008	October 29, 2008	2008Q4	17	19	44	4
ExpressJet/Atlantic Southeast	August 4, 2010	December 31, 2011	2011Q4	31	31	30	3
Frontier/Midwest	April 13, 2010	October 1, 2010	2010Q4	27	27	36	-
Southwest/Air Tran	September 27, 2010	May 2, 2011	2011Q2	29	29	33	10
United/Continental	May 3, 2010	October 1, 2010	2010Q4	27	27	36	4

Note: First carrier is the acquirer.

Control Variables

To address potential endogeneity bias, an area of challenge in airline research (Scotti and Dresner 2015), we control for factors that might influence the relationships between merger events and carriers' OTP following current practices in the airline literature (Prince and Simon 2017, Alan and Lapré 2018). We briefly discuss the reasons to include the following control variables.

First, several operational factors may impact carriers' OTP. Load factor refers to how full an aircraft is. An overloaded aircraft might necessitate longer boarding time which will consequently impact OTP. Similarly, fleet utilization, i.e., a highly-utilized and a lowly-utilized aircraft may expect different OTP due to very different routes and time schedules. Fleet heterogeneity and network sparsity, two frequently used measures in airline research, may also impact individual carrier's OTP. Second, human factors may impact OTP. The number of enplaned passengers directly impacts OTP. Greater numbers of passengers impose greater operational challenges, potentially leading to worsened OTP. In addition, the airline industry is a labor-intensive industry where airlines' ground operational efficiency, a key contributor to OTP, heavily depends on the number of employees, whereas ground staff account for 85% of an airline's employees

(DOT 2023). Third, market share also has a direct influence on OTP. Greater market share indicates more complicated networks to manage, potentially hindering carriers from achieving better OTP. Fourth, ground operations and congestion, such as Taxi-in time, Taxi-out time, and total number of flights at origin and destination airport, will also impact OTP. We, therefore, include all afore-mentioned variables in our analysis. Table 4 defines each variable and its data source. Following extant econometrics practice (Wooldridge 2010) and current airline merger literature (Stevens et al. 2016, Prince and Simon 2017), we also control for carrier fixed effects, time fixed effects, and route fixed effects.

Table 4 Variables Used in Analysis

Variable	Definition	Data Source
Carrier Delays	The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, etc.)	DOT On-time Reporting On-time Performance
Load Factor	Revenue passenger miles divided by available seat miles	DOT Schedule T1
Fleet Utilization	Block Aircraft Hours divided by Aircraft Days	DOT Schedule P52
Fleet Heterogeneity	Blau index of different aircraft types within an airline's fleet in each quarter	DOT Schedule T100
Network Sparsity	Sum of squared proportions of flights originating from each airport in an airline's network in each quarter	DOT Schedule T100
Enplaned Passengers	Number of enplaned passengers	DOT Schedule T1 in Form 41
Full-time Employees	Number of Full-Time Equivalent Employees	DOT Schedule P1(a) in Form 41
Market Share	The ratio of a carrier's quarterly revenue passenger miles to the sum of revenue passenger miles of the total 7 carriers in that quarter	DOT Schedule T1 in Form 41
Taxi-In	Taxi-in at destination airport	DOT On-time Reporting On-time Performance
Taxi-Out	Taxi-out at origin airport	DOT On-time Reporting On-time Performance
Total Number of Flights	Total number of flights at origin and destination airports	DOT On-time Reporting On-time Performance

4. METHODOLOGY – STAGGERED DIFFRENCE-IN-DIFFERENCE DESIGN

From Table 2 and Table 3, we see that our data includes seven acquirers and seven target that have been engaged in mergers and 13 carriers that have not experienced any mergers. This provides us a natural experiment for us to draw causal inference of merger effects. In econometrics terminology, the merger event is known as the "treatment"; carriers that have experienced mergers in our data are referred to as "treated" units or eventually-treated units while carriers that have never experienced mergers in our data are called "control" units or never-treated units; the impact of merger on the carriers who experienced merger events is accordingly called "treatment effect". Table 3 also shows that the seven merger events occurred at different times – commonly referred to as "staggered design", meaning there is variation in treatment timing (i.e., different merger dates) across multiple time periods (i.e., 64 quarters). Note that the current study is a single airline industry study (Tsikriktsis 2007, Alan and Lapré 2018) where 27 carriers are the total available population in this industry.

Also known as quasi-experiment, natural experiment research design is widely applied in the OM field by using two way fixed effects (TWFE) Difference-in-difference (DiD) regression to draw causal inference for treatment effects (Dong et al. 2019, Li and Wu 2020, Cui et al. 2022). However, with the

recent advancement in econometrics, TWFE DiD regression has been proven to produce biased estimates in a staggered design (Callaway and Sant'Anna, 2021; Baker et al., 2022). Choosing the correct method and estimator is paramount to provide pertinent implications for airline industry decision makers. Therefore, we explain the rationale of our choice of method in this section. As extant airline merger literature (Appendix 1) examines both an average effect and a long term effect. We also investigate these two effects.

4.1 Static Effect – the Average 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). All extant airline merger research estimated the static effect of mergers on OTP using TWFE DiD regressions (Steven et al. 2016, Prince and Simon 2017, Rupp and Tan 2019). A standard static specification, in its simplest form, is illustrated in Equation 1.

$$Y_{it} = \alpha_i + \lambda_t + \delta^{DD} D_{it} + \varepsilon_{it} \qquad \text{Equation 1}$$

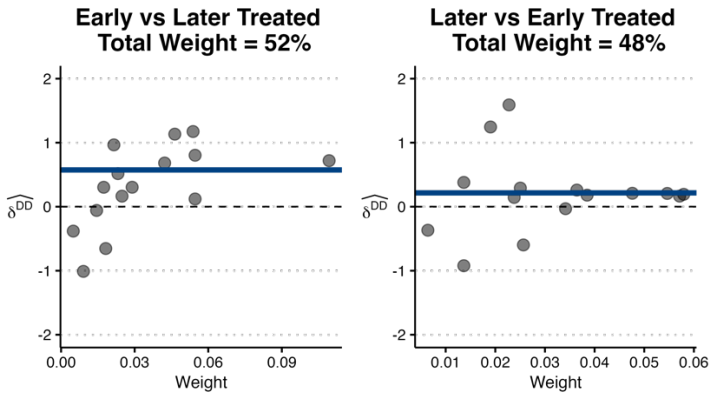
α_i is the unit or group fixed effect while λ_t is the time fixed effect. ε_{it} are error terms. D_{it} is a binary variable taking the value of 1 if unit i is treated in time period t and taking the value of 0 otherwise. The parameter of interest in the static specification is δ^{DD}, which is typically interpreted as the overall treatment effect for all units that have participated in the treatment up to a certain period T.

When computing static effect from staggered DiD design, ordinary least squares (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.

We use our data as an example to decompose TWFE weights following Goodman-Bacon (2021) to further illustrate the potential problems in the static TWFE specification. As Bacon decomposition requires a balanced panel data, we construct a balanced panel using available data for carriers that have been involved in mergers only, i.e., we did not include non-treated groups for this illustration as treated VS

non-treated groups are clean controls. We use carrier-induced delays to decompose TWFE weights into all possible 2×2 DiD estimates from pooled OLS regression in Figure 1: Early treated (as treatment) compared to later treated (as controls) as one group; and later treated (as treatment) compared to early treated (as controls) as another group. Each gray dot represents a comparison between treatment-timing cohorts. For example, ExpressJet experiencing a merger event in 2011Q4 compared to Frontier experiencing a merger event in 2010Q4. The weighted average of all these gray dots in each group is represented by the bold blue line. We see that the two weighted averages are both slightly above 0. The total weights applied by TWFE to each group is labeled as "Total Weights" in the chart title. The weights applied by TWFE to early VS later group is 52% and the weights applied by TWFE to later VS early group is 48%. Then, the ATT, the overall average treatment effect across groups, is the weighted sum of each weighted average of all comparisons. The problematic 2×2 comparisons, in this illustrative data, are those later treated VS early treated comparisons which receive a weight of 48% in the TWFE estimate. The potential contamination and biased estimate for the static single average treatment effect would arise from these 48% of comparisons.

Figure 1 Goodman-Bacon (2021) Decomposition – Carrier-Induced Delays



Note: Figure 1 reports the decomposition the weights of TWFE static regression. For the seven mergers, TWFE computes all possible 2×2 DiD estimates from pooled OLS regression in two categories: Early treated (as treatment) compared to later treated (as controls) and later treated (as treatment) compared to early treated (as controls). Each grey dot represents a comparison between treatment-timing cohorts, such as Delta in 2008Q4 VS Frontier 2010Q4. 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.

4.2 Dynamic Effect – The Long-Term Effect

Long-term effect seeks to explore how treatment effect evolves over time. This is typically known as the dynamic effect. Often referred to as event study methodology, dynamic effect allows treatment effects to vary over time non-parametrically and is a widespread practice to test the dynamic treatment effect of policy intervention (Sun and Abraham 2021, Baker et al. 2022). Researchers typically are interested in estimating the coefficients of relative time indicators after the treatment. 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 2:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{l=-K}^{-1} \mu_l \boldsymbol{D_{it}^l} + \sum_{l=0}^{L} \mu_l \boldsymbol{D_{it}^l} + \beta X_{it} + \varepsilon_{it} \quad \text{Equation 2}$$

α_i and λt are vectors of carrier 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. D_{it}^l is a set of relative time indicators. l = (–K, ... , L) represents the length of time periods relative to the time period when the merger occurred, such as (–2, –1, 0, 1, 2, 3) where 0 is when the merger started, –2 is two time periods before the merger, and 2 is two time periods after the merger. 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 such as $t = 0$ (Sun and Abraham 2021). In the current study, time period of $t = 0$ was accordingly omitted to avoid multicollinearity. $\sum_{l=-K}^{-1} \mu_l D_{it}^l$, therefore, captures the time periods up to the first quarter before merger. $\sum_{l=1}^{L} \mu_l D_{it}^l$ represents the time periods after merger ($l = 1$). The main parameters of interest in Equation 2 are the μ_ls which captures the differences in the outcome Y_{it} between treated and untreated carriers l time periods apart from the time when the merger started.

The recent advancement in econometrics also reveals major pitfalls of TWFE regression of dynamic treatment effect when there is variation in treatment timing and treatment heterogeneity (Callaway and Sant'Anna 2021, Baker et al. 2022). A decomposition of the regression coefficient $\sum_{l=0}^{L} \mu_l D_{it}^l$ on the relative time indicators reveal 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 (Sun and Abraham 2021). In addition, the weights associated with these coefficients are non-linear functions of the distribution of the groups, and these weights are prone to the issue of negative weights (Sun and Abraham 2021). These two factors together contaminate the estimation of $\sum_{l=0}^{L} \mu_l D_{it}^l$.

4.3 Choice of Method

Given the known pitfalls of TWFE estimator, different alternatives have been proposed. Although the econometrics literature has settled on the drawbacks of TWFE DiD in estimating treatment effects in a staggered design, which alternative estimator is the best still remains debated. The advancement of econometrics may continue furthering this topic. However, at the time of writing, following Baker et al. (2022), we recommend three alternatives that can be used to avoid the potential bias associated with TWFE in a staggered design: Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Stacked Regression (Deshpande and Li 2019, Cengiz et al. 2019). The commonality of these three alternatives is that they all modify those units that can serve as effective controls (i.e., avoid using already-treated units as effective controls). The difference between the three alternatives is that neither Callaway and Sant'Anna (2021) nor Sun and Abraham (2021) requires data restructuring while stacked regression requires researchers to restructure the data. In addition, it is relatively easy to incorporate covariates when using Callaway and Sant'Anna's (2021) estimator and stacked regression. In this study, we elect to use stacked regression to

estimate the effects of mergers as this alternative may be the most familiar to empirical researchers. Next, we explain how stacked regression works when used to estimate Equation 2.

Note that stacked regression per se is not an estimator. Stacked regression also estimates the model using TWFE but modifies the data such that only clean control groups are used in the regression to avoid the contamination issues. The only difference between applying TWFE directly on the original data and applying TWFE on a stacked data is that a stacked data defines all variables in a clean event-specific dataset, i.e., already-treated units are not used as effective controls. Following airline merger literature (Prince and Simon 2017), we estimate merger effect up to 20 quarters (5 years) after merger and 4 quarters before merger. Table 5 is an example of the stacked data used in our analysis. We first create a unique data identifier (column *dt* in Table 5) for each carrier. Data identifier is different from carrier identifier (column *Carriercode* in Table 5). United Airlines was assigned a data identifier 26 as is shown in the column *dt* in Table 5. United merged with Continental on October 1, 2010 (2010Q4) – the 28th occasion in our data. Because we elect to estimate merger effect 4 quarters prior and 20 quarters post-merger, the *Occasion* column starts from 24 (4 quarters prior to the merger occasion 28) and ends at 46 (20 quarters after the merger occasion 28); while the relative occasion column (*rel_occasion*) starts from – 4 to 20. Relative time indicators (columns rel_–4 .... rel_20) are the *l* time periods in Equation 2. The time period when merger occurred (rel_0) was omitted in estimation to avoid multicollinearity.

Table 5 Illustrative Stacked Data using US Airways

Carriercode	Occasion	rel_occasion	dt	rel_-4	rel__-3	... rel_19	rel_20
21	24	-4	26	**1**	0	0	0
21	25	-3	26	0	**1**	0	0
21	26	-2	26	0	0	0	0
21	27	-1	26	0	0	0	0
21	**28**	**0**	**26**	**0**	**0**	**0**	**0**
21	29	1	26	0	0	0	0
21	30	2	26	0	0	0	0
21	31	3	26	0	0	0	0
21	...	...	...	...	...	...	...
21	44	18	26	0	0	0	0
21	45	19	26	0	0	**1**	0
21	46	20	26	0	0	0	**1**

To use stacked regression, we need to create a clean stacked data. First, we need to identify all possible clean 2×2 comparable carriers for United Airlines. Among the remaining carriers that are eligible to be compared with United Airlines as a clean control, Alaska and Frontier were identified as clean effective controls for United based on the criteria of 1) the timing of the merger event; and 2) the pre-4 and post-20 quarters time frame. Then, all relevant variables are constructed and added as additional columns in Table 5 for the three carriers, i.e., each carrier will have an exact copy of Table 5. Next, the three data tables of United, Alaska, and Frontier are stacked together. As we have seven mergers, the same process repeats for the remaining six acquirers. The final data to be used for regression is a stacked data consisting of all possible clean 2×2 comparisons for all the seven acquirers. Hence, the name stacked regression.

When TWFE estimates the DiD on each clean 2×2 dataset, the unit and time fixed effects (α_i and λ_t) are saturated with dataset indicators. Then TWFE applies variance weighting to combine the treatment effects across groups and time periods (Deshpande and Li 2019, Cengiz et al. 2019), thus, avoiding the contamination issues. Step-by-step detail to build stacked data is provided in the supplementary R code. Interested readers can refer to the R code for more information on how the staked data was constructed.

5. ANALYSIS

All analysis in the current study was conducted in R 4.3.1. How each model was constructed and tested was also provided in the supplementary R code. From Table 3, we see that target airlines continue to report to DOT as an individual carrier up to 10 quarters during the post-merger grace period, which enables us to estimate the merger impact for both the acquirer and the target. Following our airline merger literature review (Appendix 1), we assess both static effect and dynamic effect (event study) of merger impact.

5.1 Impact of Mergers on OTP for Acquirers

5.1.1 Static Effect – Acquirer

We first estimate static effect for acquirers using our restructured stacked data. All analysis in the current study was estimated using data up to 4 quarters prior merger and 20 quarters post-merger following Prince and Simon (2017). We create a dummy variable "Treated" and assign all 20 post-merger quarters as 1 and all 4 pre-merger quarters as 0. Therefore, the variable "Treated" captures the static effect. Table 6 presents the results. In our stacked regression, all statistical inferences use clustered standard errors at the carrier level to account for autocorrelation in the data (Wooldridge 2010). From Table 6, we see that using 20 post-merger quarters as the time period to measure merger impact, the static effect (Treated, $\delta^{DD} = 0.03$, $t = 0.50$) is not statistically significant for acquirers. Therefore, we can conclude that when measured by 20 post-merger quarters and measured by average effect during these 20 quarters, there is no statistically significant impact on OTP for acquirers following a merger.

Table 6 Static Effect of Merger on Acquirers – Estimated using Stacked Regression

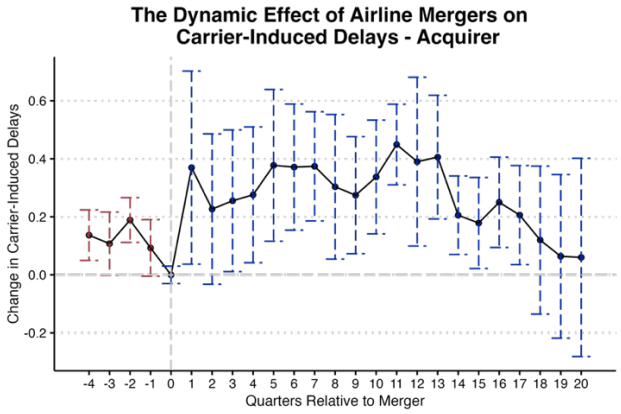
Variables	***δ^{DD}***	***Standard Error***	***t-Value***
Dependent variable is carrier-controllable delays			
Treated	0.03	0.07	0.50
Load factor	3.05*	1.27	2.38
Fleet Utilization	0.15	0.08	1.72
Fleet Heterogeneity	0.04	0.11	0.37
Network Sparsity	-0.50***	0.09	-5.32
# of Enplaned Passengers	0.72°	0.37	1.90
# of Employees	0.61*	0.25	2.45
Market Share	-1.17°	0.59	-1.96
Taxi In (Min)	0.08*	0.03	2.72
Taxi Out (Min)	0.59***	0.07	7.71
Total Flights	0.26*	0.09	2.71
Fixed Effects Included: Carrier, Occasion, Route, Year, Quarter			
$R^2 = 0.617$ Observations = 333,830			

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

5.1.2 Dynamic Effect – Acquirer

Researchers and policymakers may be more interested to know how the merger effect unravels over time. Therefore, we continue to estimate the dynamic effect, or event study, of merger impact on OTP. We estimate Equation 2 using the stacked data created from Table 5. Figure 2 and Table 7 report the results. In Figure 2 (and the rest of the Figures in this paper), red dots and red dashed lines represent point estimates and 95% confidence bands for *pre-treatment* periods. Blue dots and blue dashed lines represent point estimates and 95% confidence bands for *post-treatment* periods. Standard errors are clustered at the carrier level.

Figure 2 Long Term Effect of Mergers on Acquirers – Estimated using Stacked Regression



Note: Figure 2 reports results of dynamic treatment effect for 20 quarters post-merger using stacked regression. Red dots and red dashed lines represent point estimates and 95% confidence intervals for pre-treatment periods. Blue dots and blue dashed lines represent point estimates and 95% confidence intervals for the treatment effect of mergers on OTP following mergers, allowing for clustering at the carrier level.

From Table 7, we see that except for Occasion 2, Occasion 18 to 20, all the coefficients of the post-merger occasions are statistically significant. Therefore, despite an insignificant static effect, when estimated using an event study methodology, the story is completely different. After eliminating other delays that cannot be attributed to carriers, such as security delay, weather delay, and National Aviation System delay, our results are distinctively different from extant airline merger research. When OTP was narrowed down to carrier-induced delays only, we see that in the four quarters following merger, OTP worsened (i.e., carrier-indued delay increased), in contrast to Rupp and Tan (2019) who found that OTP (including non-carrier-induced delays) improved immediately in the four quarters following mergers. Prince and Simon (2017) used travel time (including non-carrier-induced delays) to measure OTP and found that in 1-2 years post-merger, travel time was not impacted; in 3-5 years post-merger, travel time even shortened. In sharp contrast to Prince and Simon (2017), our results indicate that OTP keeps worsening (carrier-induced delays keeps increasing) from year one up to the first quarter in year five post-merger. Merger impact on carrier-induced delays become non-significant starting from the second quarter of year five post-merger. Although our results verified Steven et al. (2016)'s findings that on-time arrivals

worsened in a three-year post-merger window, we extend Steven et al. (2016)'s findings by offering new insights in that the worsening effect lasts beyond three years and lingers into the first quarter of year five post-merger.

Table 7 Long Term Effect of Mergers on Acquirers – Estimated using Stacked Regression

Dependent variable is carrier-controllable delays			
l Wave relative to mergers	$\widehat{\mu_l}$	Standard Error	t-Value
-4	0.13*	0.05	2.45
-3	0.10	0.06	1.68
-2	0.18**	0.04	3.77
-1	0.09	0.05	1.56
1	0.37*	0.17	2.17
2	0.23	0.13	1.74
3	0.25°	0.13	1.97
4	0.27*	0.12	2.22
5	0.37*	0.13	2.74
6	0.37**	0.11	3.26
7	0.37**	0.10	3.70
8	0.30*	0.12	2.37
9	0.27*	0.10	2.54
10	0.34**	0.09	3.43
11	0.45***	0.07	6.06
12	0.39*	0.14	2.67
13	0.40**	0.11	3.68
14	0.20**	0.06	3.04
15	0.18*	0.08	2.23
16	0.25*	0.08	2.91
17	0.20*	0.09	2.19
18	0.11	0.13	0.89
19	0.06	0.15	0.41
20	0.06	0.18	0.36
Load factor	3.59**	0.93	3.82
Fleet Utilization	0.45°	0.23	1.93
Fleet Heterogeneity	-0.08	0.06	-1.27
Network Sparsity	-0.60**	0.15	-3.94
# of Enplaned Passengers	1.45*	0.53	2.70
# of Employees	0.62**	0.19	3.18
Market Share	-2.04**	0.62	-3.26
Taxi In (Min)	0.05	0.04	1.36
Taxi Out (Min)	0.64***	0.07	8.28
Total Flights	0.22°	0.10	2.11
Fixed Effects Included: Carrier, Occasion, Route, Year, Quarter			
R^2 = 0.695 Observation = 155,976			

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

5.2 Impact of Mergers on OTP for Target

As DOT allows a grace period for both the acquirer and the target to still report as two individual carriers after the official merger completion date, we are able to utilize this grace period to model merger impact on the target as well, which was not examined in airline merger literature before. From Table 3, we see that target airlines continue to report to DOT as an independent carrier up to 10 quarters post-merger, ranging from 0 to 10 quarters. 0 indicates that the target airline immediately reported to DOT under the acquirer's brand (the second and third scenario in Section 3.1). We follow the same process of constructing stacked

data for target airlines as described in Section 4.3. Note that the maximum post-merger quarters reported by target airlines under their own brand are 10 quarters. As a result, we estimate a maximum 10 post-quarter impact for the target airlines.

5.2.1 Static Effect of Mergers on Target Airlines

Table 8 reports the static effect of 10 post-merger quarters for the target airlines. The coefficient (δ^{DD} = 0.19, t = 1.29) is statistically insignificant. Therefore, we can draw the conclusion that when measured by 10 post-merger quarters and measured by the average effect during these 10 quarters, there is no statistically significant impact on OTP for target airlines following a merger.

Table 8 Static Effect of Merger on Target – Estimated using Stacked Regression

Variables	δ^{DD}	***Standard Error***	***t-Value***
Dependent variable is carrier-controllable delays			
Treated	0.19	0.15	1.29
Load factor	3.26°	1.76	1.85
Fleet Utilization	-0.08	0.30	-0.27
Fleet Heterogeneity	0.02	0.04	0.71
Network Sparsity	0.14	0.29	0.49
# of Enplaned Passengers	0.93	0.79	1.16
# of Employees	0.27	0.44	0.63
Market Share	-1.48	0.97	-1.51
Taxi In (Min)	0.12*	0.04	2.56
Taxi Out (Min)	0.57***	0.13	4.19
Total Flights	0.20	0.13	1.52
Fixed Effects Included: Carrier, Occasion, Route, Year, Quarter			
R^2 = 0.668 Observations = 65,872			

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

5.2.2 Dynamic Effect of Mergers on Target Airlines

For an event study analysis of the merger impact on target airlines, we estimate up to 10 quarters post-merger. Figure 3 and Table 9 report the results. Apart from a marginally significant post-merger occasion 1 and a significant occasion 9, all other coefficients are statistically insignificant. Therefore, we can conclude that for target airlines, there is no strong statistical evidence to support that OTP of target airlines has been impacted following a merger.

Figure 3 Long Term Effect of Mergers on Target – Estimated using Stacked Regression

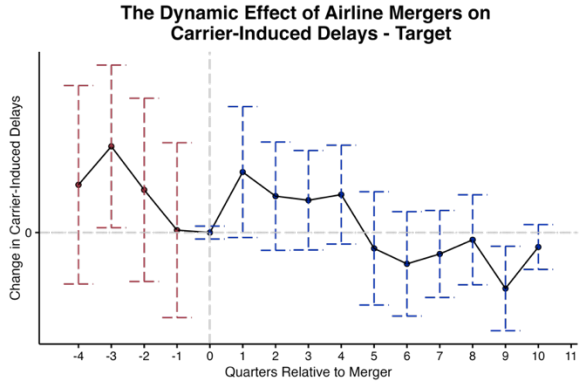


Table 9 Long Term Effect of Mergers on Acquirers – Estimated using Stacked Regression

Dependent variable is carrier-controllable delays			
l Wave relative to mergers	$\hat{\mu}_l$	Standard Error	t-Value
-4	0.22	0.22	0.96
-3	0.40°	0.19	2.10
-2	0.19	0.21	0.93
-1	0.01	0.20	0.05
1	0.28°	0.15	1.80
2	0.16	0.13	1.28
3	0.15	0.12	1.23
4	0.17	0.12	1.45
5	-0.07	0.13	-0.53
6	-0.14	0.12	-1.14
7	-0.09	0.10	-0.91
8	-0.03	0.11	-0.30
9	-0.25*	0.10	-2.46
10	-0.06	0.06	-1.07
Load factor	2.84°	1.57	1.80
Fleet Utilization	-0.08	0.21	-0.41
Fleet Heterogeneity	-0.001	0.03	-0.04
Network Sparsity	-0.02	0.30	-0.08
# of Enplaned Passengers	1.42*	0.57	2.44
# of Employees	0.03	0.48	0.07
Market Share	-1.75*	0.71	-2.46
Taxi In (Min)	0.12°	0.06	1.96
Taxi Out (Min)	0.64***	0.13	4.97
Total Flights	0.13	0.15	0.91
Fixed Effects Included: Carrier, Occasion, Route, Year, Quarter			
R^2 = 0.682 Observation = 65,872			

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

6. PARALLEL TREND ASSUMPTION AND ROBUSTNESS TEST

6.1 The Parallel Trend Assumption

The validity of all DiD design relies on the parallel trend assumption (i.e., the control group and the treatment group should trend at similar paces on the outcome variable before the treatment). There are two common practices to test the parallel trend assumption in the OM field. The first practice is to compare the trendline of the average outcome on the control group and the treatment group before the treatment to visually detect if the two trendlines are trending parallelly. However, visually parallel trendlines do not equal to statistical significance. The second practice is to test the coefficients on the leads (i.e., relative time indicators leading to the treatment) using event study. If no statistical significance is found on the leads, researchers normally conclude that there is no statistical difference between the trend of the treatment group and the trend of the control group before the treatment. Hence, the parallel trend assumption holds. However, TWFE has also been proved to produce biased estimates on relative time indicators of both leads and lags (Sun and Abraham 2021) in a staggered design, leading to either Type I or Type II errors (Baker et al. 2022) – meaning that the widespread practice of using TWFE event study to test for the parallel trend assumption is also problematic when used on the original untransformed data.

de Chaisemartin and D'Haultfoeuille (2020) accordingly proposed an alternative placebo estimator (DID_M^{pl}) to test for pretrends which "essentially compares the outcome's evolution from $t-2$ to $t-1$, in groups that switch and do not switch treatment between $t-1$ and t" (p. 2989). In our case, DID_M^{pl} computes the change in carrier-induced delay between carriers experiencing mergers and those not experiencing mergers one time period before the merger event. We also extend DID_M^{pl} and compute $DID_M^{pl,2}$, $DID_M^{pl,3}$, and $DID_M^{pl,4}$, which compare the change in carrier-induced delays two, three, and four periods before the merger event. We report the results in Figure 4 and Table 10. We see from Table 10 that the estimates of all the four placebo estimates ($t=-4,-3,-2,-1$) are not significantly different from 0, indicating that the parallel trend assumption was not violated: carriers who engaged in mergers do not experience different trends before the merger event compared with the carriers who did not experience mergers. Therefore, we can draw the conclusion that the parallel trend assumption holds in our data.

Figure 4 Parallel Trend Assumption Test

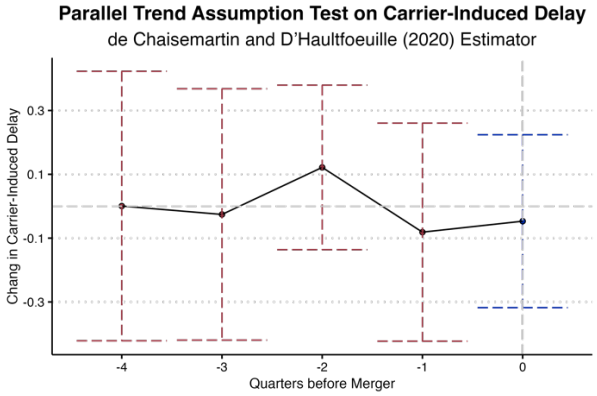


Table 10 Parallel Trend Assumption Test

	Estimate	Standard Error
DID_M^{pl}	-0.081	0.174
$DID_M^{pl,2}$	0.121	0.131
$DID_M^{pl,3}$	-0.025	0.200
$DID_M^{pl,4}$	0.001	0.215
(Covariates included in model but omitted in reporting)		

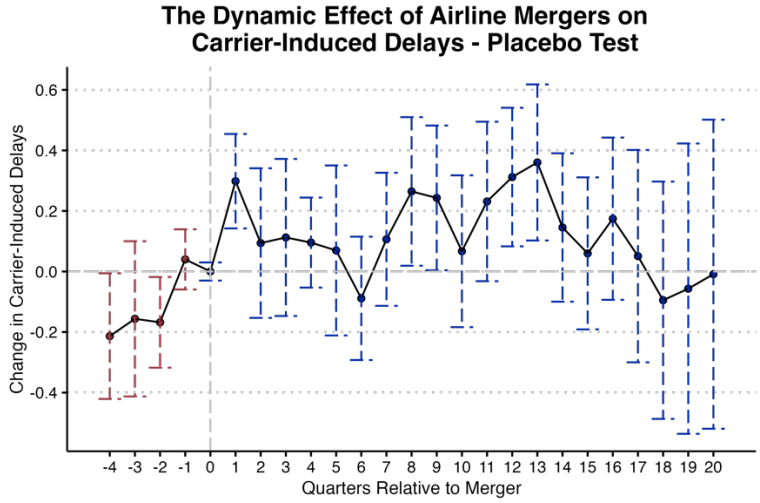
Note: This table reports the estimates for the four placebos.
*P value: *** 0.001, ** 0.01, * 0.05, °0.1*

6.2 Placebo Test Using Merger Announcement Date

From Table 3, we see that the time gap between the merger announcement date and the merger completion date spans over 1-3 quarters. Carriers may start integration after the merger announcement date, potentially triggering a treatment effect before the official merger completion date that was used in our analysis. Therefore, we use the merger announcement date as a placebo "treatment" date to run a placebo test. We run placebo test for the dynamic effect for both the acquirer and the target and our results indicate that there

is no strong statistical evidence supporting the impact of merger announcement on OTP. We show the graph of the placebo test for acquirers in Figure 5. The associated table is omitted to save space.

Figure 5 Using Merger Date as Placebo to Test the Long-term Effect for Acquirers



7. POST-HOC SIMULATION – WHY TRUSTING STACEKD REGRESSION?

Extant airline merger literature presents mixed and conflicting findings regarding the merger impact on OTP, we attribute these mixed findings to two potential causes. The first potential cause is that when calculating various measures of OTP, current airline merger research included non-carrier-induced delays – a staggering 71.9% of total delays in our data, which is highly likely to produce biased results. The second potential cause is that extant airline merger research all adopted TWFE DiD to estimate treatment effect on the original untransformed data, which may also add another layer of bias to the findings given the pitfalls of TWFE estimator when applied on the original data.

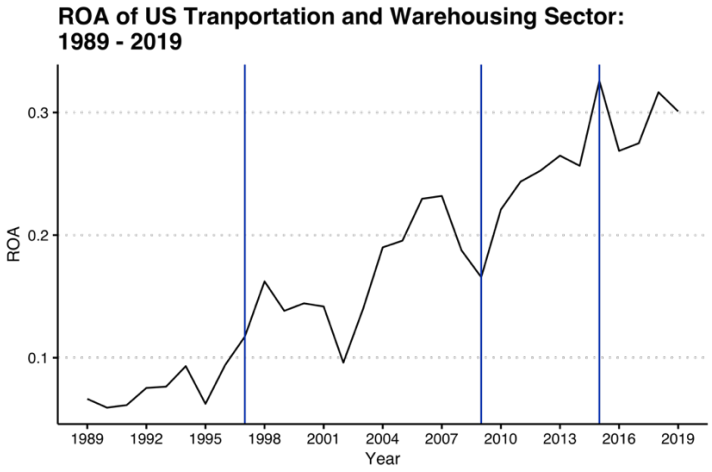
In our current study, we only focus on measuring OTP using the 28.1% carrier-induced delays in our data. Therefore, our findings should more accurately reflect the true effect on OTP caused by the merger event itself. To make sure that our methodological approach does not add another layer of "noise" to extant mixed findings, we perform a post-hoc simulation to demonstrate the difference between using TWFE on the raw data and using TWFE on the restructured stacked data.

Compustat is a widely used data source to study operations management topics in recent years (Dong et al. 2020). Therefore, we use Compustat data to simulate a staggered difference-in-difference 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. All data retrieving and simulation process are attached in the supplementary R code, too.

We then create three artificial exogeneous shocks as is shown in Figure 6. The solid black line in Figure 6 is the actual ROA trend while the three solid blue vertical lines represent three exogeneous shocks. The three exogenous shocks create 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 6. Following recent econometrics literature (Baker et al. 2022), our 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, our simulation has no pretrends (flat) for the treated groups. The choice of the annual increase in ROA in terms of number of σ_{ROA} is random. We also randomly tested different combinations of random σ_{ROA} and our simulation results do not change.

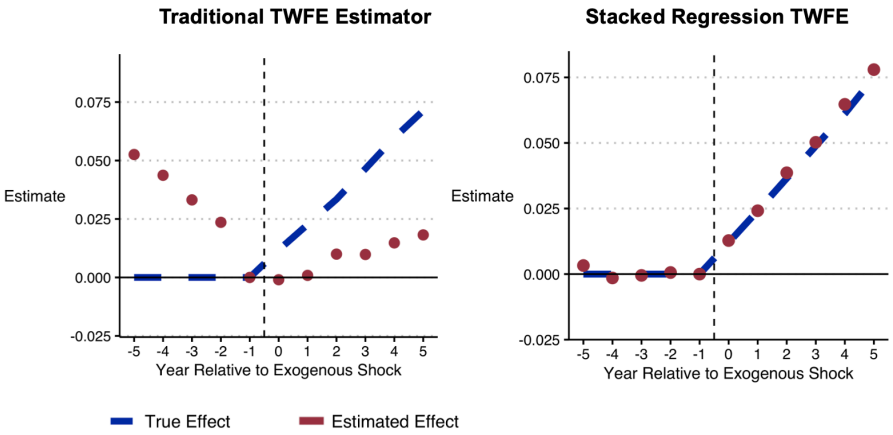
Figure 6 ROA of US Transportation and Warehousing Sector: 1989-2019



We then use TWFE event study (Equation 2) to estimate the coefficients of relative year time indicators (5 years pre and 5 years post the treatment) on the original data. We plot the result on the left in Figure 7. Blue line represents the actual simulated trend (true effect) and the red line represents TWFE estimates. We then create a stacked data and use TWFE to estimate the leads and lags again on the restructured stacked data and plot the graph on the right in Figure 7. From the two side-by-side graphs, we see that for the coefficients on the leads, TWFE, using the raw data, estimated a downward going pre-trend when there is actually no pretrend; while a stacked regression, using restructured stacked data, produces almost perfect estimates. For the coefficients on the lags, we also see that the estimates of TWFE on the raw data deviates from the true effect while a stacked regression produces an almost perfect fit. Therefore, we can conclude that our analysis, using restructured stacked data of clean controls to assess merger impact

on OTP, does not just add another layer of "noise" to the literature. But rather, our study provides more accurate estimates on the treatment effects.

Figure 7 Simulation to Test the Parallel Trend Assumption



8. CONTRIBUTION AND CONCLUSION

Our study makes several distinctive contributions to the OM literature. First, our study reminds airline researchers about the potential bias to include non-carrier-induced delays to assess policy impact on OTP. This can be a serious issue in drawing causal inference as non-carrier-induced delays account for 71.9% of all delays in our data (72.3% of all delays as reported by Nicolae et al. 2017). When evaluating policy impact on OTP, we call for airline researchers to use carrier-induced delays to measure OTP to increase research rigor as well as to provide more accurate managerial insights for airline decision makers. At a minimum, on-time arrivals, when used to assess any policy impact, should have nothing to do with extreme weather delays which account for almost half of all delays.

Second, our research extends the understanding of merger impact on OTP in the airline industry. Different from related research, our study presents new findings when narrowing down OTP to carrier-induced delays only. Different from Rupp and Tan (2019) who found OTP improved in the immediate four quarters following a merger, our results indicate that OTP worsened (carrier-induced delays increased) in the first four quarters following a merger. Unlike Prince and Simon (2017) who found no impact of merger on travel time in the first two years and travel time even improved from year three to year five following a merger, our analysis shows that OTP keeps worsening throughout year one to the first quarter of year five. Supplementary to Steven et al. (2016) whose findings show that OTP has deteriorated in a three-year time window post-merger, our results indicate that the worsening effect lasts longer – up to the first quarter in year five post-merger. All these nuanced findings not only contribute to knowledge accumulation in airline merger literature but also provide more accurate managerial insights – partially due to the exclusion of the 71.9% non-carrier-induced delays and also partially due to our different methodological approach, which leads to our next contribution.

Our simulation analysis in Section 7 illustrates the serious bias associated with TWFE when applied on the original data in a staggered DiD design. But when TWFE was applied on the restructured stacked data, it produces almost perfect estimates for both the leads and the lags. To this end, our research contributes to advancing research validity of DiD research in OM field, specifically for staggered DiD design where TWFE produces biased estimates for both average treatment effect and dynamic effects using the original untransformed data (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021). In a staggered DiD design, we call for researchers to adopt the latest estimators (Callaway and Sant'Anna 2021 estimator; Sun and Abraham 2021 estimator) and alternatives (stacked regression using restructured stacked data) to estimate treatment effects to increase research rigor. In addition, using TWFE event study to test the parallel trend assumption has a long track record in OM research, which is also problematic as in a staggered design, using an event study to test the parallel trend assumption can result in both Type I and Type II errors (Baker et al. 2022), as is partially shown in Figure 7. Therefore, we also call for researchers to use alternative estimators to test the parallel trend assumption as this assumption is paramount to the validity of all DiD designs in the OM field.

Next, our analysis of static and dynamic effect reminds airline researchers the importance of assessing treatment effects from different perspectives. Take static effect and dynamic effect for the acquirers in Table 5 and Table 6 for example: if researchers only examined static effect using 20 post-merger quarters, the conclusion would be that mergers do not have a statistically significant impact on OTP ($\delta^{DD} = 0.03$, $t = 0.50$). However, an event study analysis using stacked data reveals that mergers do have a negative impact on OTP through year one to the first quarter of year five. The two different analyses reveal two different stories, reinforcing the importance of assessing treatment effect from different perspectives.

Lastly, our study also provides more accurate insights to airline executives and policymakers to design appropriate operations strategies. Despite the struggles and diminished prospects in the airline industry during Covid-19, "interest of mergers shows no sign of flagging" (Louge 2021). In the wake of a potential new wave of carrier consolidation post Covid-19 (Primack 2022), our new findings regarding the merger impact on carrier-induced delays can provide more pertinent guidance for airline policymakers to design proper post-merger strategies.

In sum, by investigating the impact of mergers on OTP using seven recent U.S. airline mergers, we provide new insights by excluding non-carrier-induced delays and by using the latest advancement in econometrics. We call for researchers to consider using the latest estimators to increase research rigor to draw treatment effect in a staggered DiD design, or at least use these new estimators as a robustness test if researchers continue to report traditional TWFE DiD results produced from the untransformed data. In addition, the track-record practice of using event study to test the parallel trend assumption in a staggered

DiD design is also problematic (Baker et al. 2022). We, therefore, also call for researchers to adopt alternatives to test this assumption to increase research validity.

REFERENCES

Alan Y, Lapré MA (2018) Investigating operational predictors of future financial distress in the US airline industry. Production Oper. Management. 27(4):734-755.

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

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

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.

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

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

Department of Justice (2019) Statutory provisions and guidelines of the antitrust division. Accessed August 25, 2022, https://www.justice.gov/atr/file/761131/download.

Department of Transportation (2023) Data directory: list of databases. https://www.transtats.bts.gov/DataIndex.asp. Accessed on December 30, 2023.

Deshpande M, Li Y (2019) Who is screened out? Application costs and the targeting of disability programs. Am. Econ. J. Econ. Policy. 11(4):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. Manuf. Serv. Oper. Manag. 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.

Dresner M, Xu K (1995) Customer service, customer satisfaction, and corporate performance. J. Bus. Logist. 16(1):23-40.

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

Hüschelrath K, Müller K (2014) Airline networks, mergers, and consumer welfare. J. Transp. Econ. Policy. 48(3):385–407.

Josephs L (2018) Years after airline mergers, flight attendants are finally flying together. Retrieved from https://www.cnbc.com/2018/10/01/years–after–airline–mergers–flight–attendants–start–to–fly–together.html. Accessed on December 30, 2023.

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

Lichtenber FR, Kim M (1989) The effects of mergers on prices, costs, and capacity utilization in the US air transportation industry, 1970–84 (No. w3197). National Bureau of Economic Research.

Louge F (2021) How the pandemic is changing the outlook for airline mergers. https://www.frontier-economics.com/uk/en/news-and-articles/articles/article-i8918-how-the-pandemic-is-changing-the-outlook-for-airline-mergers/. Accessed on July 30, 2023.

McCartney S (2010) An airline report card: Fewer delays, hassles last year, but bumpy times may be ahead. Wall Street Journal. January 7, D1–D3.

Mouawad J (2012) For United, Big Problems at Biggest Airline. https://www.nytimes.com/2012/11/29/business/united-is-struggling-two-years-after-its-merger-with-continental.html. Accessed on December 30, 2023.

Nicolae M, Arıkan M, Deshpande V, Ferguson M (2017) Do bags fly free? An empirical analysis of the operational implications of airline baggage fees. Management Sci. 63(10):3187-3206.

Primack D (2022) Spirit Airlines takeover fight heads to a vote. https://www.axios.com/2022/06/28/spirit-airlines-takeover-fight-heads-to-a-vote. Accesses on December 30, 2023.

Prince JT, Simon DH (2017) The impact of mergers on quality provision: Evidence from the airline industry. J. Ind. Econ. 65(2):336–362.

Rupp NG, Tan KM (2019) Mergers and product quality: A silver lining from de–hubbing in the U.S airline industry. Contemp. Econo. Policy. 37(4):652-672.

Scotti D, Dresner M (2015) The impact of baggage fees on passenger demand on US air routes. Transp. Policy. 43: 4-10.

Singal V (1996)a. Airline mergers and multimarket contact. Manage Decis. Econ. 17(6):559–574.

Singal V (1996)b. Airline mergers and competition: an integration of stock and product price effects. J. Bus. 69(2):233–268.

Steven AB, Yazdi AA, Dresner M (2016) Mergers and service quality in the airline industry: A silver lining for air travelers? Transport. Res. Part E Logistics Transport. Rev. 89:1–13.

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

Tsikriktsis N (2007) The effect of operational performance and focus on profitability: A longitudinal study of the US airline industry. Manuf. Serv. Oper. Manag. 9(4):506-517.

Vaze V, Luo T, Harder R (2017) Impacts of airline mergers on passenger welfare. Transport. Res. Part E Logistics Transport. Rev. 101:130–154.

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