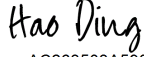


## Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

DocuSigned by:  
Signature:   
AC263503A590411...

Hao Ding \_\_\_\_\_  
Name

4/5/2024 | 10:50 AM EDT  
Date


**Title** Essays on Social, Healthcare, and Digital Operations

**Author** Hao Ding

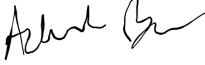
**Degree** Doctor of Philosophy

**Program** Business


**Approved by the Committee**

DocuSigned by:  
  
BC1DDA0D03AC493...

Ruomeng Cui  
*Advisor*

DocuSigned by:  
  
97FAC4E5A57648E...

Achal Bassamboo  
*Committee Member*

DocuSigned by:  
  
2E1D2C8717A74B7...

Donald Lee  
*Committee Member*

DocuSigned by:  
  
E9E629FBADBB45A...

Daniel McCarthy  
*Committee Member*

DocuSigned by:  
  
9CE5C09A38E44AD...

Feng Zhu  
*Committee Member*

*Committee Member*

**Accepted by the Laney Graduate School:**

---

Kimberly Jacob Arriola, Ph.D, MPH  
Dean, James T. Laney Graduate School

---

Date

Essays on Social, Healthcare, and Digital Operations

By

Hao Ding  
B.S., Indiana University, 2017

Advisor: Ruomeng Cui, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the James T.  
Laney School of Graduate Studies of Emory University  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in  
Business  
2024

Abstract  
Essays on Social, Healthcare, and Digital Operations  
By Hao Ding

This dissertation includes three essays, each focusing on an important element of my research. The first essay explores a long-standing societal issue, gender inequality. In particular, the essay addresses how working from home during the COVID pandemic lockdown impacted the productivity of male and female academic researchers differently. The second essay focuses on healthcare operations. In particular, we develop a new framework to study state- and time-dependent service processes in healthcare. The third essay is on platform and digital operations. We study how business buyers in the Business-to-Business (B2B) leverage online reviews strategically on B2B trading platforms.

Essays on Social, Healthcare, and Digital Operations

By

Hao Ding  
B.S., Indiana University, 2017

Advisor: Ruomeng Cui, Ph.D.

A dissertation submitted to the Faculty of the James T.  
Laney School of Graduate Studies of Emory University  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy  
in  
Business  
2024

## Acknowledgement

I would like to thank many people through my Ph.D. journey and beyond, but particularly to the following list of people:

My Mom, Guohua Xu

My Aunt, Yihua Xu

My Advisor and Mentor, Ruomeng Cui

My Co-author, Donald Lee

My great friend and co-author, Sokol Tushe

My faithful friends, F.D.C. Xiaomai

F.D.C. Beibei, and F.D.C. Midnight

## LIST OF TABLES AND FIGURES for Essay I

### List of Figures

1	Time Trends of US Preprints from December 2019 to May 2020 . . . . .	E1 – 10
2	Impact of Lockdown on Gender Inequality across Countries . . . . .	E1 – 16
A.1	Time Trends of US Preprints from December 2018 to May 2019 . . . . .	E1 – 26

### List of Tables

1	Summary Statistics . . . . .	E1 – 12
2	Impact of Lockdown on Gender Inequality with the Discipline Fixed Effect . . . . .	E1 – 13
3	Impact of Lockdown on Gender Inequality in Aggregation . . . . .	E1 – 14
4	Impact of Lockdown on Gender Inequality by Academic Ranks . . . . .	E1 – 14
5	Impact of Lockdown on Gender Inequality by University Ranking . . . . .	E1 – 15
6	Impact of Lockdown on Abstract Views . . . . .	E1 – 18
7	Impact of Lockdown on Downloads . . . . .	E1 – 18
8	Impact of Lockdown on Gender Inequality among All-male and All-female Preprints	E1 – 19
9	Impact of Lockdown on Gender Inequality Using the Balanced Panel . . . . .	E1 – 20
A.1	Robustness to Different University Rankings . . . . .	E1 – 27
A.2	Parallel Trends Test . . . . .	E1 – 28
A.3	Impact of Lockdown on Gender Inequality by Academic Ranks in Aggregation . . .	E1 – 28
A.4	Impact of Lockdown on Gender Inequality among All Male or All Female Preprints in Aggregation . . . . .	E1 – 29
A.5	Summary Statistics for December 2018 - May 2019 . . . . .	E1 – 29
A.6	Falsification Test . . . . .	E1 – 30
A.7	Summary Statistics for Downloads and Abstract Views . . . . .	E1 – 30
A.8	Impact of Lockdown on Abstract Views . . . . .	E1 – 30
A.9	Impact of Lockdown on Downloads . . . . .	E1 – 30

## LIST OF TABLES AND FIGURES for Essay II

### List of Figures

1	A stylized service process. . . . .	E2 – 12
2	An Illustration of Workload Measure . . . . .	E2 – 17
3	Curve of simulated average service time per patient . . . . .	E2 – 26
4	Time-varying revenue trends . . . . .	E2 – 26
5	Static revenue trends . . . . .	E2 – 26

### List of Tables

1	Workload Literature . . . . .	E2 – 5
2	Static workload effects on patient LOS (w/o quadratic workload terms) . . . . .	E2 – 19
3	Static workload effects on patient LOS (w/ quadratic workload terms) . . . . .	E2 – 19
4	Log-normal service rate estimation of (8) . . . . .	E2 – 21
5	Costs savings comparison across different specifications . . . . .	E2 – 27
6	Costs savings comparison across different specifications . . . . .	E2 – 27



## LIST OF TABLES AND FIGURES for Essay III

### List of Figures

1	Search Result of a Product . . . . .	E3 – 10
2	A Seller’s Page on Alibaba.com . . . . .	E3 – 11
3	Transaction History of a Seller . . . . .	E3 – 12
4	Review History of a Seller . . . . .	E3 – 13
5	Illustration of the DID Identification . . . . .	E3 – 19

### List of Tables

1	Summary Statistics . . . . .	E3 – 17
2	Impact of Online Reviews on B2B Sales . . . . .	E3 – 21
3	Impact of Sellers’ Recent Transaction Volumes on Buyers’ Review Probability . . . . .	E3 – 21
4	Impact of Sellers’ Recent Transaction Volumes on Buyers’ Review Ratings . . . . .	E3 – 22
5	Negative Reviews and Repeat Purchase . . . . .	E3 – 23
6	Falsification Test . . . . .	E3 – 24
7	Temporal Trend Not Affecting the Results . . . . .	E3 – 25
8	Impact of Sellers’ Recent Transaction Value on Buyers’ Review Probability . . . . .	E3 – 26

# Gender Inequality in Research Productivity during the COVID-19 Pandemic

Ruomeng Cui

Goizueta Business School, Emory University, ruomeng.cui@emory.edu

Hao Ding

Goizueta Business School, Emory University, hao.ding@emory.edu

Feng Zhu

Harvard Business School, Harvard University, fzhu@hbs.edu

**Problem definition:** We study the disproportionate impact of the lockdown as a result of the COVID-19 outbreak on female and male academics' research productivity in social science. **Academic/practical Relevance:** The lockdown has caused substantial disruptions to academic activities, requiring people to work from home. How this disruption affects productivity and the related gender equity is an important operations and societal question. **Methodology:** We collect data from the largest open-access preprint repository for social science on 41,858 research preprints in 18 disciplines produced by 76,832 authors across 25 countries over a span of two years. We use a difference-in-differences approach leveraging the exogenous pandemic shock. **Results:** Our results indicate that, in the 10 weeks after the lockdown in the United States, although total research productivity increased by 35 percent, female academics' productivity dropped by 13.2 percent relative to that of male academics. We also show that this intensified productivity gap is more pronounced for assistant professors and for academics in top-ranked universities and is found in six other countries. **Managerial implications:** Our work points out the fairness issue in productivity caused by the lockdown, a finding that universities will find helpful when evaluating faculty productivity. It also helps organizations realize the potential unintended consequences that can arise from telecommuting.

*Key words:* Gender inequality, research productivity, telecommuting, COVID-19

## 1. Introduction

The Coronavirus 2019 (COVID-19) pandemic has significantly changed the way people live and work. Many people have been forced to work from home through telecommuting, potentially affecting their productivity. We study how this pandemic shock affected academics' research productivity using data from the world's largest open-access repository for social science—the Social Science Research Network (SSRN).<sup>1</sup> We provide evidence that, as a result of the lockdown in the United States, female researchers' productivity dropped significantly relative to that of male researchers.

<sup>1</sup> Source: [https://en.wikipedia.org/wiki/Social\\_Science\\_Research\\_Network](https://en.wikipedia.org/wiki/Social_Science_Research_Network), accessed June 2020.

In response to the pandemic, the US and many other countries required their citizens to stay at home. As a result of the lockdown, many people have had to perform work duties at home along with household duties. Most countries have also closed schools and daycare centers, massively increasing childcare needs. With the childcare provided by grandparents and friends limited due to the social distancing protocol, parents must care for their children themselves. In addition, restaurants have been either closed or do not allow dine-ins, which has increased the need for food preparation at home. Given that women, on average, carry out disproportionately more childcare, domestic labor, and household responsibilities (Bianchi et al. 2012), they are likely to be more affected than men by the lockdown.

The lockdown has also disrupted how academics carry out their activities. Many countries have closed their universities, so faculty have to teach and conduct research from home. A researcher's productivity is jointly determined by his or her available time for research and research efficiency (KC 2019). First, given the unequal distribution of domestic duties, the pandemic is more likely to burden female researchers with more home-related tasks, leaving them less time to dedicate to research. Second, scientific research generally requires a quiet and interruption-free environment. As a result of the pandemic, female researchers are more likely to multitask between research and home-related tasks and thus to have lower efficiency in conducting research. Together, these factors suggest that female academics' productivity is likely to be disproportionately affected compared with that of male academics.

Anecdotal evidence provides mixed support for this prediction (Dolan and Lawless 2020). A recent survey involving 4,500 principal investigators reported significant and heterogeneous declines in the time they spend on research (Myers et al. 2020). Several journal editors have noticed that, while there was a 20–30-percent increase in submissions after the pandemic started, most was from male academics (Beck 2020). Amano-Patiño et al. (2020) find a particularly large number of senior male economists, rather than mid-career economists, exploring research questions arising from the COVID-19 shock. Others have seen no change or have been receiving comparatively more submissions from women since the lockdown (Kitchener 2020). Overall, there is a dearth of systematic evidence on whether and to what extent the shock has affected gender inequality in academia. We provide such systematic evidence, showing an unequal impact on productivity for female and male researchers.

It is an important operations and societal question to understand the change in productivity and the related gender equity caused by the reorganization of work and care at home. In this paper, we use a large dataset on female and male academics' production of new research papers to systematically study whether COVID-19 and the subsequent lockdown have had a disproportionate

---

effect on female academics' productivity. We also identify the academic ranks, universities, and countries in which this inequality is intensified.

We collect data on all research papers uploaded to SSRN in 18 disciplines from December 2018 to May 2019 and from December 2019 to May 2020. We extract information on paper titles and the authors' names, affiliations, and addresses, which we use to identify the authors' countries and academic ranks, and the ranking of their institutions. We also use their names and faculty webpages to identify their gender. In particular, we use a large database to predict authors' gender. For author names with a prediction confidence level lower than 80 percent, we use Amazon Mechanical Turk to identify gender manually. The final dataset includes 41,858 papers written by 76,832 authors from 25 countries. Our main analysis focuses on US academics; we then perform the same analysis for other countries.

We use a difference-in-differences (DID) approach to estimate the effect. We compute the number of papers produced by female and male academics in each week, then compare the variations in the productivity gap between genders before and after the start of the lockdown and show that the gap increased after the start of the lockdown. We also validate that female and male authors' preprint volumes followed the same parallel time trend before the lockdown and that there was no significant change in the research productivity gap in 2019 during the same time of year. Taken together, these results suggest that the intensified disparity has primarily been driven by the pandemic shock.

We find that during the 10 weeks since the lockdown began in the US, female academics' research productivity dropped by 13.2 percent compared to that of male academics. The effect persists when we vary the time window since the pandemic outbreak. Our findings show that when female and male academics face a reorganization of care and work time, women become significantly less productive by producing fewer papers. We also find that the quality of their uploaded papers, measured by the download and view rates, does not change. Finally, we find that the effect is more pronounced among assistant professors and among researchers in top-ranked research universities and that it exists in six other countries.

While gender inequality has been long documented for academics in terms of tenure evaluation (Antecol et al. 2018), coauthoring choices (Sarsons 2017), and number of citations (Ghiasi et al. 2015), the COVID-19 pandemic brings this issue to the forefront. Our study is among the first to rigorously quantify such inequality in research productivity as a result of the pandemic and our results highlight that this disruption has exacerbated gender inequality in the academic world. Because all academics participate together in open competition for promotions and positions, these short-term changes in productivity will affect their long-term career outcomes (Minello 2020). Thus, institutions should take this inequality into consideration when evaluating faculty members.

Our paper contributes to the literature on productivity, a central topic in operations management. Previous studies have examined key determinants of workers' productivity, such as workers' capacity (Tan and Netessine 2014, KC 2019), multitasking (KC 2014, Bray et al. 2016, KC 2020), peer effects (Song et al. 2018, Tan and Netessine 2019), and task sequences (Ibanez et al. 2018). We contribute to the literature by showing that the disruption due to the pandemic has significantly enlarged the productivity gap between female and male researchers, highlighting fairness as an important factor in performance evaluation.

Our work also sheds light on the fairness issues that could arise from telecommuting, an operations choice faced by companies. Since working from home can provide a flexible work schedule for employees and reduce office-related costs for companies, an increasing number of companies are choosing this operating model. Between 2005 and 2015, the number of US employees who chose to telecommute increased by 115% (Abrams 2019). By 2019, about 16% of the total workforce in the US was working remotely full time or part time (U.S. Bureau of Labor Statistics 2019). During the pandemic, telecommuting is a constraint rather than a choice; many companies were forced to allow telecommuting. But going forward, an increasing number of companies may choose to offer this operating model to provide flexible work schedule to employees and reduce office-related costs. For example, Twitter and Facebook have already announced that their employees could work from home permanently (McLean 2020) and JP Morgan planned to expand its telecommuting program (Kelly 2020). Despite the growing popularity of telecommuting, scholars and practitioners still lack a comprehensive understanding of the managerial and societal impact of this operational choice (Nicklin et al. 2016). We contribute to the literature by pointing out the productivity inequality caused by the lockdown and telecommuting, which might lead to a longer-term unemployment risk for women, an unintended consequence that companies and society should take into account when making their operational choices or designing policies for performance evaluation. Our findings help institutions and firms understand the potential implications in designing and implementing telecommuting.

## **2. Literature Review**

Our work is closely related to three streams of literature: productivity, social operations, and telecommuting.

### **2.1. Productivity**

According to the productivity literature (KC and Terwiesch 2009, Tan and Netessine 2014), working in different environments causes significant changes in operational factors that drive worker's productivity. In our context, due to the pandemic, researchers have to change to working from

---

home, potentially affecting several drivers of productivity identified by research, such as multitasking, workload, task sequence, and peer effects.

Multitasking is particularly relevant to our research context. When working from home during the pandemic, researchers may need to allocate their limited cognitive capacity across home-related and work-related tasks, thus dealing with more distractions arising from multitasking. Prior studies have shown mixed effects of multitasking on workers' productivity, such as an increased service speed with a lower service quality for restaurant waiters (Tan and Netessine 2014) and a slower processing for bank associates (Staats and Gino 2012). The productivity losses could be greater for jobs requiring greater cognitive capacity. For example, in the judiciary system, reducing multitasking has been shown to help judges focus on the most recent cases, reduce the switching costs between cases, and increase the case completion rate (Bray et al. 2016). In healthcare, excessive multitasking and frequent interruptions in the work flow have been shown to undermine the productivity of discharging (KC 2014), processing (Berry Jaeker and Tucker 2017), and medication delivery (Batt and Terwiesch 2017).

Workload, task sequences, and peer effects are examples of other operations drivers in researchers' productivity during the pandemic. Workers have been shown to adjust their productivity based on their workload, slowing down when facing more workload and speeding up when facing less because they internalize the congestion cost (KC and Terwiesch 2009, Tan and Netessine 2014). However, the extra workload could make workers fatigued, which could reduce productivity (Salvendy 2012, KC 2019). Technology—such as tabletop technology in restaurants—has been shown to improve workers' productivity by reducing their non-value-added tasks, enabling them to focus on more important tasks (Tan and Netessine 2020). The literature has also identified workers' choice of task sequence as a productivity factor (Staats and Gino 2012, Ibanez et al. 2018). Workers tend to deviate to suboptimal task sequences when facing a higher workload or when fatigued (KC et al. 2020). Another driver of workers' productivity is peer effects—workers adapt their own productivity to that of their peers' (Schultz et al. 1998). For example, having a particularly capable worker on a shift could motivate slower workers to speed up but could also discourage good performers as it becomes more difficult for them to outperform their peers and reach their goals (Tan and Netessine 2019). Displaying peers' productivity publicly has been shown to improve productivity (Song et al. 2018).

## **2.2. Social Operations**

This paper sheds light on a key social issue—fairness and equity—in research productivity, adding to the growing literature on the social impact of operational choices. Several recent influential papers by Tang and Zhou (2012), Lee and Tang (2018), and Dai et al. (2020) encourage OM

---

researchers to work on socially responsible topics that are important to corporations and to society at large. Papers have examined the use of review information to reduce racial discrimination arising in the sharing economy (Cui et al. 2020, Mejia and Parker 2020) and the gender inequality driven by specific compensation schemes (Pierce et al. 2020). The literature on gender bias has shown evidence that female researchers and students tend to be discredited when they are evaluated alongside equally competent male candidates (Moss-Racusin et al. 2012, Sarsons 2017), that women are more likely to be assigned more service-oriented and less desirable tasks (Guarino and Borden 2017) with fewer promotion opportunities (Babcock et al. 2017), and that women are often responsible for more housework and childcare (Schiebinger and Gilmartin 2010, Misra et al. 2012).

In our context, when working from home, the unequal distribution of housework means that women are more likely to deal with non-work-related tasks during the lockdown and lose productivity. A recent survey involving 4,500 principal investigators shows that female scientists self-reported a sharper reduction in research time during the lockdown, primarily due to childcare needs (Myers et al. 2020). We contribute to the literature by providing evidence that the lockdown affects productivity and exacerbates gender inequity in the workplace, potentially leading to a long-term career risk for women, an unintended consequence that organizations should consider when designing their operations models and performance evaluation policies.

### **2.3. Telecommuting**

Our work is also related to the emerging literature on organizations' telecommuting choices. Transitioning from traditional in-office work to telecommuting might affect workers' behavior and productivity through team-work and peer effects. For example, Dutcher and Saral (2012) observe that workers do not indulge in free-riding behavior when a team is made up of in-office workers and telecommuters, and Bloom et al. (2015) demonstrates that telecommuting can improve productivity when it is carried out in a quiet environment. Our work adds to this stream of literature by demonstrating an unexpected social issue of fairness arising from this operating model.

## **3. Theory Development**

A researcher's productivity can be measured as the product of the amount of time he or she can dedicate to research, *Time Available for Research*, and how efficiently he or she conducts research, *Research Efficiency*. This definition is consistent with the key insights from the literature that productivity is determined by two elements: (1) the capacity constraint due to physical or cognitive limitations and (2) efficiency variations with changes in operations factors in the working environment (KC 2019).

For many researchers, the outbreak of COVID-19 has affected their productivity both in terms of time available for research and research efficiency. In response to the pandemic, most countries have

---

closed schools and daycare centers and required that their citizens to be quarantined at home. As a result, researchers from more than 1,100 colleges and universities had to carry out both work and household duties at home (National Conference of State Legislatures 2020). We next illustrate how the disruptions change researchers' time available for research and research efficiency separately. We then argue how these changes might be unequal among female and male researchers.

The pandemic changes researchers' working environment, resulting in a need to reallocate their time across research, work-related tasks (such as commuting, social interaction, and service to the academic community), and home-related tasks (such as childcare and house chores). Certain tasks have an in-person nature, such as commuting to the office, serving administrative duties, and interacting socially with colleagues. These tasks are significantly reduced during the lockdown and the time savings could be substantial; for example, workers in the US on average spend an hour commuting each work day (United States Census Bureau 2018). During the pandemic, researchers could use this additional capacity for research. At the same time, they might have to allocate more time to home-related tasks. Even without the pandemic, working from home constantly exposes researchers to the home environment and home duties. During the lockdown, to make matters worse, many services and amenities such as schools, day-care centers, hospitals, and restaurants have either been closed or operating at a lower capacity. Researchers might need to allocate extra time to domestic duties. For example, childcare could be particularly time-demanding since parents have diminished access to their regular childcare support network, such as professional caregivers, relatives, and friends. Consequently, researchers who are responsible for more house works and childcare are more likely to allocate additional time to domestic duties. In conclusion, a researcher could have either more or less time available due to the pandemic, which depends on the new allocation of home responsibilities.

Besides the time available for research, the pandemic affects drivers of research efficiency, such as multitasking, task sequence, fatigue, and peer effects. Conducting scientific work often requires hours of interruption-free environment. When working from home, although researchers are not distracted by activities like commuting, administrative duties, and social interactions with colleagues, they are likely to be distracted by childcare and housework, resulting in excessive multitasking. Multitasking means that workers have to allocate their limited cognitive capacity across multiple tasks. The setup cost associated with switching between tasks and the difficulty of focusing on and organizing relevant information hinder efficiency (KC 2014). Multitasking has also been shown to induce stress and frustration (Mark et al. 2008), make people more easily distracted (Levitin 2014), and exhaust their cognitive capacity (KC 2014). Home duties and urgency often require researchers' immediate attention, forcing them to deviate from their optimal task sequences, which would in turn reduce their research efficiency. Researchers may also encounter a heavier workload



---

from increased housework and thus experience fatigue, which could also reduce efficiency. Last but not least, working from home makes it difficult to observe one’s colleagues’ productivity and to discuss research topics with their peers, both of which reduce the positive influence of peers in motivating researchers, inspiring research ideas, and improving efficiency (Song et al. 2018, Tan and Netessine 2019). In conclusion, a researcher’s research efficiency could go up and down due to the pandemic, depending on the level of impact they get from multitasking, fatigue, altered task sequence, and altered peer effects.

We next demonstrate how the changes in research time and efficiency can be different for women and men. Women are on average, disproportionately burdened with childcare and household responsibilities (Bianchi et al. 2012). In the US, they are shown to spend almost twice as much time as men on housework and childcare in the US (Bianchi et al. 2012). Moreover, there are 8.5 million more single mothers than single fathers (Alon et al. 2020). Even in the gender-egalitarian countries of northern Europe, women are responsible for almost two-thirds of the unpaid work (European Commission 2016). Among heterosexual couples with female breadwinners, women still do most of the care work (Chesley and Flood 2017). The same pattern exists in academia (Schiebinger and Gilmartin 2010, Andersen et al. 2020). Female professors are shown to spend more time doing housework and carework than male professors across various ranks; for example, 34.1 hours per week versus 27.6 hours for lecturers, 29.6 hours per week versus 25.1 hours for assistant professors, and 37.7 hours per week versus 24.5 hours for associate professors (Misra et al. 2012).

The lockdown has caused a surge in domestic duties. The unequal distribution of domestic duties means that the pandemic might further enlarge the gap between women and men’s domestic workload and thus might affect female and male researchers’ productivity unequally. First, in terms of time available for research, female researchers are likely to reallocate more time to domestic duties due to the pandemic than male researchers do, leaving them with less capacity for research. Second, in terms of efficiency, female researchers are more likely to be disrupted by multitasking between research and home-related tasks and consequently deviate from their optimal task sequences. Taken these factors together, female researchers tend to suffer more from a reduced amount of time available for research and diminished efficiency compared to male researchers, which suggests a disadvantage in women’s productivity during the pandemic. We therefore hypothesize that, during the pandemic, female researchers are more likely to be disproportionately affected in their productivity compared with male researchers.

#### **4. Data and Summary Statistics**

We collect data from the Social Science Research Network (SSRN), a repository of preprints with the objective of rapidly disseminating scholarly research in social science. We gather data on *all*

social science preprints submitted from December 2018 to May 2019 and from December 2019 to May 2020. We extract information on paper titles and the authors' names, affiliations, and addresses. We use the authors' addresses to identify their countries. The COVID-19 outbreak began at different times in different countries, so we collect each country's lockdown start date from news sources and a United Nations report.<sup>2</sup> We drop authors without addresses or with addresses in more than one country because we cannot determine when these authors were affected by the lockdown. We also drop countries with fewer than 800 submissions during our study period. The final data set consists of a total of 41,858 papers in 18 disciplines produced by 76,832 authors from 25 countries.

To identify the authors' genders, we first use a database called *Genderize*,<sup>3</sup> which predicts gender based on first name and provides a confidence level. About 78 percent of the authors' genders were identified with confidence levels over 80 percent. For the remaining authors, we use Amazon Mechanical Turk to manually search for their professional webpages based on names and affiliations and then infer their genders from their profile photos. Our dataset contains 21,733 female academics and 55,099 male academics.

We aggregate the number of new preprints at the weekly level. We then count the number of papers uploaded by each author in each week. To measure the *effective* productivity for preprints with multiple authors, when a preprint has  $n$  authors,<sup>4</sup> each author gets a publication count of  $1/n$ .<sup>5</sup> Finally, we aggregate the effective number of papers to the gender level: in each week, we count the total number of papers produced by male and female authors separately in each social science discipline.

Figure 1 plots the time trend of preprints in aggregation from December 3, 2019 to May 19, 2020 in the US. The vertical line represents the week of March 11, 2020, on which nationwide lockdown measures began in the US.<sup>6</sup> We can observe that male academics, on average, have submitted more preprints than female academics, and that female and male academics' research productivity evolved in parallel before the lockdown. After the lockdown started, however, male academics significantly boosted their productivity, whereas female academics' productivity did not change much, indicating an increased productivity gap.

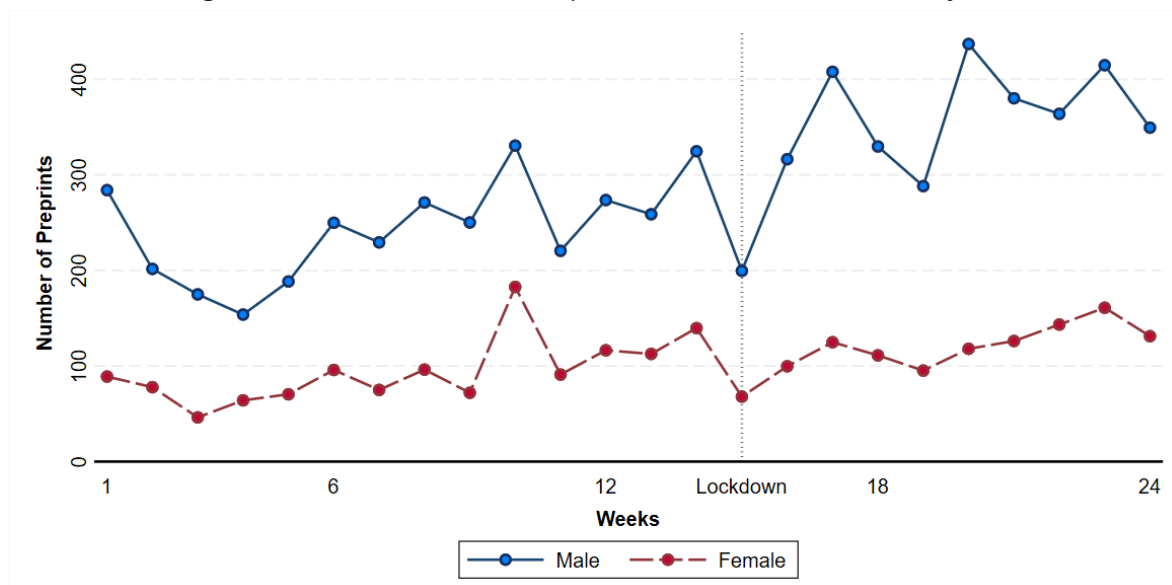
<sup>2</sup> Source: <https://en.unesco.org/covid19/educationresponse>, accessed June 2020.

<sup>3</sup> Source: <https://genderize.io/>, accessed June 2020.

<sup>4</sup> Note that in many social science disciplines, author names are listed in alphabetical order.

<sup>5</sup> Note that the validity of this measure relies on the assumption that female and male researchers' relative contributions to the paper do not change significantly after the lockdown. If female researchers have decreased their contributions to the teamwork since the lockdown, this measure would underestimate gender inequality during the pandemic. In Section 7.4, we study *all-male* and *all-female* preprints as an alternative measure, which minimizes potential work shifting across genders.

<sup>6</sup> Most universities were closed in the week of March 11, 2020. Source: <https://gist.github.com/jessejanderson/09155afe313914498a32baa477584fae?from=singlemessage&isappinstalled=0>, accessed June 2020.

**Figure 1** Time Trends of US Preprints from December 2019 to May 2020

This graph plots the time trend of the number of preprints for female and male academics. The vertical line represents the start of the lockdown due to COVID-19 in the US.

To ensure that our results are not driven by seasonality, we plot the time trend of preprints during the same time window in 2019 in Appendix Figure A.1. We observe a similar pattern before the week of March 11, 2019, but there is no significant change in the productivity gap after that week.

We use authors' professional information to identify academic ranks (e.g., PhD student or assistant, associate or full professor). Specifically, we ask workers on Amazon Mechanical Turk to find each researcher's curriculum vitae or professional webpage. To study how the pandemic affects researchers with different academic ranks, we categorize researchers into *students* (which includes PhD and postdoctoral students), *assistant professors*, *associate professors*, and *full professors*, each accounting for 20.3 percent, 16.1 percent, 19.7 percent, and 44.0 percent of the population, respectively. Next, we use authors' affiliations to classify the ranking of their universities. To ascertain whether the productivity gap is intensified or weakened across top-ranked and lower-ranked research universities, we collect social science research rankings from three sources: QS University Ranking,<sup>7</sup> Times Higher Education,<sup>8</sup> and Academic Ranking of World University.<sup>9</sup> We then use these data to rank US universities in order to study the heterogeneous effects across university rankings.

<sup>7</sup> Source: <https://www.topuniversities.com/university-rankings/university-subject-rankings/2020/social-sciences-management>, accessed June 2020.

<sup>8</sup> Source: [https://www.timeshighereducation.com/world-university-rankings/2020/subject-ranking/social-sciences#!/page/0/length/25/sort\\_by/rank/sort\\_order/asc/cols/stats](https://www.timeshighereducation.com/world-university-rankings/2020/subject-ranking/social-sciences#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats), accessed June 2020.

<sup>9</sup> Source: <http://www.shanghairanking.com/FieldSOC2016.html>, accessed June 2020.

Table 1 reports the summary statistics for the weekly number of preprints by gender and discipline from December 3, 2019 to May 19, 2020, spanning 24 weeks, as well as split-sample statistics prior to or after the lockdown. This sample includes 9,943 preprints produced by 15,494 authors in the US and 21,065 preprints produced by 37,997 authors across all countries. The average number of submissions per week is 444.6 in the US and 877.7 across all 25 countries. Notably, while research productivity in the US increased by 35 percent after the lockdown, male authors seem to be the main contributors to this increase.

About 78 percent of the preprints fall under multiple disciplines.<sup>10</sup> Note that when computing the total preprints, we count the paper only once when aggregating across disciplines to avoid multiple counting. When computing the number of preprints in each discipline, we separately count all of the papers classified under each one. We observe substantial variations across disciplines. Among the 18 disciplines, Political Science, Economics, and Law received the most submissions, whereas Geography, Criminal Justice and Education the fewest. While there was a large increase in productivity in several disciplines, such as Economics, Political Science, Finance, Health Economics, and Sustainability, after the COVID-19 outbreak, other disciplines showed no obvious increase. A few disciplines, such as Anthropology, Cognitive, and Information Systems, even experienced a decline.

## 5. Identification

Our identification exploits the lockdown in response to the COVID-19 outbreak as an exogenous shock that has caused substantial disruption to academic activities, requiring academics to teach, conduct research, and carry out household duties at home. The validity of our approach depends on the assumption that the shock is exogenous with respect to the researchers' anticipated responses. If a particular gender group of researchers anticipated and strategically prepared for the shock by accelerating the completion of their research papers, this could confound the treatment effect. In reality, this possibility is unlikely because of the rapid development of the situation. COVID-19 was regarded as a low risk and not a threat to the US in late January (Moreno 2020), and no significant actions had been taken other than travel warnings issued until late February (Franck 2020). It quickly turned into a global pandemic after the declaration of the World Health Organization on March 11, 2020, followed by the nationwide shelter-in-place orders within a week.<sup>11</sup>

We adopt the difference-in-differences (DID) methodology, a common approach used to evaluate people's or organizations' responses to natural shocks (e.g., Seamans and Zhu 2013, Calvo et al.

<sup>10</sup> Authors self-classify their own preprints into disciplines when they upload their papers. SSRN reviews and approves these classifications.

<sup>11</sup> Source: <https://www.cdc.gov/nchs/data/icd/Announcement-New-ICD-code-for-coronavirus-3-18-2020.pdf>, accessed June 2020.

**Table 1** Summary Statistics

Level	Weekly No. of Preprints	All observations					Before lockdown		After lockdown	
		Mean	Std. dev	Max	Min	Total	Mean	Std. dev	Mean	Std. dev
All Disciplines (US only)	All	444.6	109.4	617	224	9,934	378.8	88.0	511.4	86.0
	Female authors	111.3	30.8	186	47	2,493	103.4	36.2	119.3	21.4
	Male authors	333.3	85.3	480	161	7,441	275.4	55.4	392.1	68.6
By Discipline (US only)	Accounting	19.5	7.2	40	9	468	17.9	6.3	21.8	8.2
	Anthropology	85.0	21.5	141	63	2,040	93.9	24.0	72.5	6.9
	Cognitive	11.3	9.2	31	1	271	14.1	11.1	7.4	3.2
	Corporate	14.1	6.5	27	3	339	12.2	6.5	16.8	5.8
	Criminal	15.4	6.7	27	4	370	12.8	6.7	19.1	4.9
	Economics	133.2	54.2	237	37	3,197	106.6	39.1	170.5	51.6
	Education	17.9	7.0	36	7	429	16.9	7.4	19.2	6.7
	Entrepreneurship	9.9	5.3	22	2	238	10.2	4.9	9.5	5.9
	Finance	91.7	34.5	139	25	2,201	78.5	35.5	110.2	24.0
	Geography	8.2	3.3	17	3	196	7.5	2.7	9.1	4.0
	Health Economics	8.4	10.1	47	0	202	3.0	2.1	16.0	12.1
	Information Systems	15.6	7.3	39	7	374	17.4	8.6	13.1	4.2
	Law	98.5	24.3	142	44	2,365	94.1	26.7	104.7	20.1
	Management	33.4	11.4	56	12	802	33.4	13.3	33.4	8.6
	Organization	20.5	11.5	44	3	491	16.9	10.2	25.5	11.7
	Political Science	167.9	50.5	255	85	4,030	142.1	39.0	204.1	42.8
	Sustainability	22.8	11.9	66	8	546	18.1	5.9	29.3	15.1
Women/Gender	18.0	4.7	28	10	431	17.2	4.4	19.0	5.2	
All countries	All	877.7	199.3	1,175	487	21,065	779.1	177.5	1015.8	140.4
	Female authors	246.5	53.9	347	165	5,916	231.0	57.0	268.2	42.9
	Male authors	631.2	152.0	866	322	15,149	548.1	124.4	747.6	104.3

The table summarizes the weekly number of papers from December 2019 to May 2020. The sample includes 15,494 authors from the United States and 37,997 authors across all countries. There are 9,934 preprints produced by US authors, 2,493 of which are produced by 3,877 female researchers and 7,441 by 11,617 male researchers. We gather the country-specific lockdown time to split our sample to before and after the lockdown for each country.

0). We perform the DID analysis using outcome variables on two levels: the number of preprints in each discipline and the number of preprints aggregated across all disciplines.<sup>12</sup>

We compare the productivity gap between female and male researchers before and after the pandemic outbreak using the following model specification with discipline-level panel data:

$$\log(\text{Preprints}_{igt}) = c + \text{Female}_g + \beta \text{Female}_g \times \text{Lockdown}_t + \gamma_t + \delta_i + \epsilon_{igt}, \quad (1)$$

where  $i$  denotes discipline,  $g$  denotes the gender group,  $t$  denotes the week,  $\log(\text{Preprints}_{igt})$  represents the logged number of preprints uploaded for discipline  $i$  for gender  $g$  during week  $t$ ,  $\gamma_t$  is the time fixed effect,  $\delta_i$  is the discipline fixed effect that captures the time-invariant characteristics of discipline  $i$ , and  $\epsilon_{igt}$  is the error term. The time fixed effect  $\gamma_t$  includes weekly time dummies that control for time trends. The dummy variable  $\text{Female}_g$  equals 1 if gender  $g$  is female, and 0 otherwise. The dummy variable  $\text{Lockdown}_t$  equals 1 if week  $t$  occurs after the lockdown measure was adopted (that is, the week of March 11, 2020), and 0 otherwise. Its main effect is absorbed by the time fixed effects. The coefficient  $\beta$  estimates the effect of the lockdown on female academics' research productivity relative to male academics' productivity.

<sup>12</sup> We also perform a DID analysis with the country-level panel data. For this, we assign a lockdown dummy to each country and combine data across countries to form a country-level panel. This analysis yields consistent results.

We also use aggregate-level data to estimate the effect with the following DID specification:

$$\log(\text{Preprints}_{gt}) = c + \text{Female}_g + \beta \text{Female}_g \times \text{Lockdown}_t + \gamma_t + \epsilon_{gt}, \quad (2)$$

where  $g$  denotes the gender group,  $t$  denotes the week,  $\log(\text{Preprints}_{gt})$  represents the logged number of preprints uploaded for gender  $g$  during week  $t$ , and  $\epsilon_t$  is the error term. As before, we include the time fixed effect  $\gamma_t$ .

## 6. Empirical Results

In this section, we report the effect of the COVID-19 outbreak on research productivity. We first show the average effect of the pandemic on gender inequality. We then show the heterogeneous effects across academic disciplines, faculty ranks, university rankings, and countries.

### 6.1. Main Results

Table 2 reports the estimates with the discipline fixed effect using Equation (1). Table 3 reports the estimated effect of the pandemic shock on research productivity at the aggregate level using Equation (2). In each analysis, we use 14 weeks before the lockdown as the pre-treatment period and 6 to 10 weeks after the lockdown as the post-treatment periods. The analyses yield consistent results. First, in line with our summary statistics, the results show that fewer preprints are produced by female academics than by male academics in general. Second, since the lockdown began, there has been a significant reduction in female academics' productivity relative to that of their male colleagues, indicating an exacerbated productivity gap in gender. The coefficient of the interaction term in Column (5) suggests a 13.2-percent reduction in females' productivity over the 10-week period after the lockdown relative to the males'.<sup>13</sup>

**Table 2 Impact of Lockdown on Gender Inequality with the Discipline Fixed Effect**

Variables	Dependent variable: No. of Preprints (in logarithm) by discipline				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	−0.791*** (0.042)	−0.791*** (0.042)	−0.791*** (0.042)	−0.791*** (0.042)	−0.791*** (0.042)
<i>Female × Lockdown</i>	−0.140* (0.076)	−0.148** (0.072)	−0.162** (0.068)	−0.157** (0.065)	−0.142** (0.063)
Discipline fixed effects	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	720	756	792	828	864
$R^2$	0.837	0.836	0.839	0.841	0.841

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (1) with the discipline fixed effect. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

<sup>13</sup> Because the outcome variable is logged, the percentage change in the outcome variable is computed as  $e^{\text{coefficient}} - 1$ .

**Table 3 Impact of Lockdown on Gender Inequality in Aggregation**

Variables	Dependent variable: No. of Preprints (in logarithm) in aggregation				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	-1.013*** (0.054)	-1.013*** (0.054)	-1.013*** (0.053)	-1.013*** (0.053)	-1.013*** (0.053)
<i>Female</i> × <i>Lockdown</i>	-0.197** (0.068)	-0.199*** (0.064)	-0.173** (0.067)	-0.159** (0.066)	-0.150** (0.064)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	40	42	44	46	48
$R^2$	0.981	0.982	0.982	0.982	0.983

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (2) at the aggregation level. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 6.2. Heterogeneous Effects

In this section, we study the heterogeneous effects across faculty ranks, university rankings, disciplines, and countries.

**6.2.1. Effects across academic rank.** We study how the pandemic affects researchers with different academic ranks (such as PhD student, or assistant, associate or full professor). Because assistant professors often face more pressure than senior professors to publish papers in order to get tenure, they are more motivated to devote as much time as possible to research. They are also at a stage at which many have young children. As a result, the pandemic’s effect on the productivity gap is likely to be more pronounced for this group. We repeat the DID analysis for each academic rank and report results based on Equations (1) and (2) in Table 4 and Table A.3, respectively. The two tables show consistent results. Table 4 shows that female assistant professors experienced the most significant drop in research productivity (compared to male junior faculty) since the lockdown.

**Table 4 Impact of Lockdown on Gender Inequality by Academic Ranks**

Academic rank	Dependent variable: No. of Preprints (in logarithm) by Academic Ranks				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
Student	-0.002	-0.016	-0.014	-0.038	-0.071
Assistant professor	-0.529***	-0.419***	-0.448***	-0.419***	-0.441***
Associate professor	0.038	-0.026	0.004	-0.044	-0.034
Full professor	-0.181	-0.163	-0.063	-0.044	-0.064
Observations	720	756	792	828	864

This table reports the estimated coefficients based on Equation (1) with the discipline fixed effect for academics within each rank. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Time and discipline fixed effects at the weekly level are included in each regression. Standard errors and estimates of other variables are omitted for brevity. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**6.2.2. Effects across university ranking.** Table 5 replicates the DID analysis using Equation (1) for a subset of academics based on the rankings of their universities.<sup>14</sup> Due to our focus on

<sup>14</sup> For authors affiliated with more than one academic institutions, we use the highest-ranked institution.

social science, we use the 2020 QS World University Ranking for social sciences and management for the main analysis. We separately analyze academics in universities ranked in the top 10, 20, ..., and 100. The results show that the COVID-19 effect is more pronounced in top-tier universities and that this effect in general decreases and becomes less significant as we include more lower-ranked universities. We find similar results when using the two other rankings, as shown in Appendix Table A.1.

**Table 5 Impact of Lockdown on Gender Inequality by University Ranking**

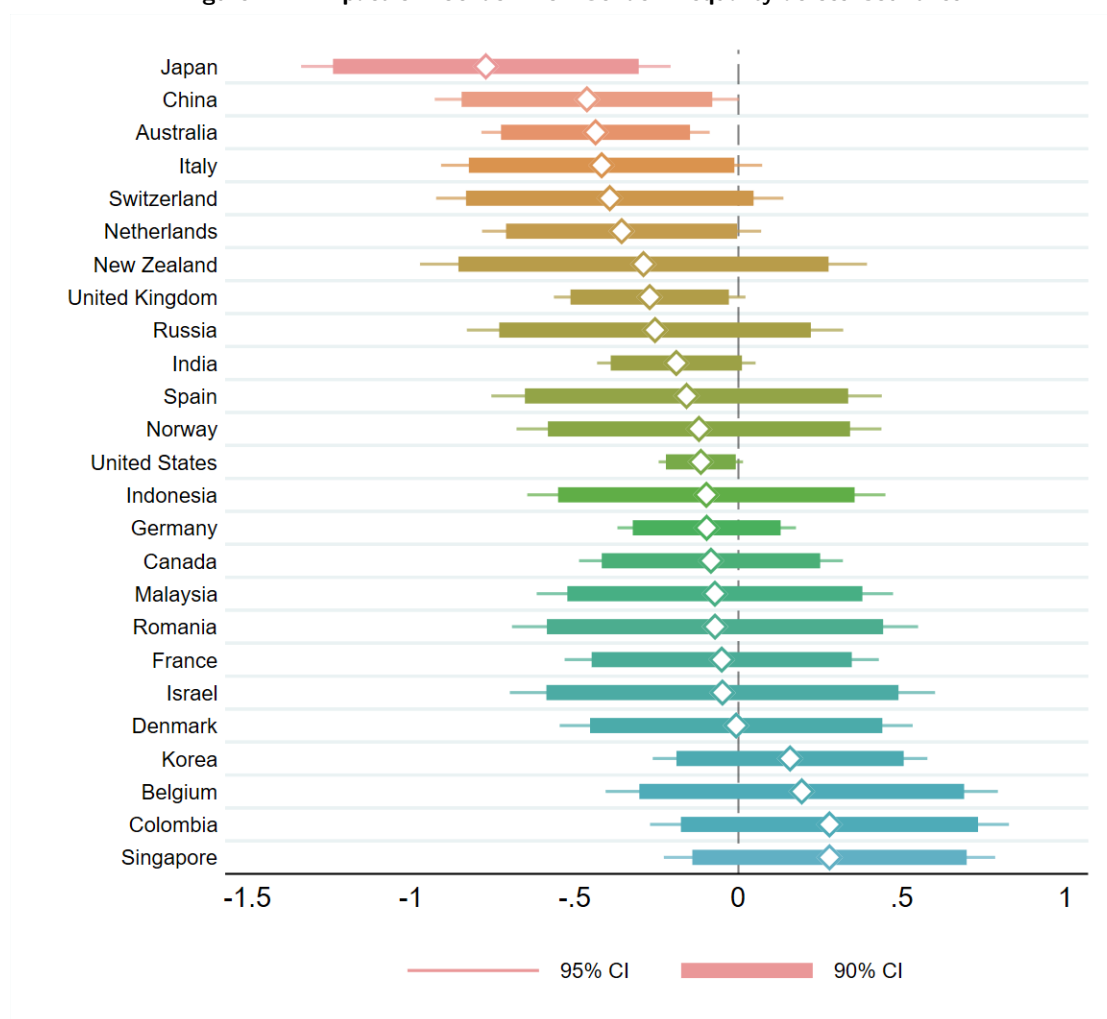
Universities by QS Ranking	Dependent variable: No. of Preprints (in logarithm) by discipline				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
Top 10	-0.169**	-0.199***	-0.158**	-0.153**	-0.165**
Top 20	-0.181**	-0.215***	-0.183**	-0.179***	-0.183***
Top 30	-0.189**	-0.210**	-0.167**	-0.168**	-0.170**
Top 40	-0.218***	-0.238***	-0.200***	-0.191***	-0.194***
Top 50	-0.197**	-0.214***	-0.180***	-0.179***	-0.182***
Top 60	-0.138*	-0.163*	-0.145*	-0.143**	-0.155**
Top 70	-0.142*	-0.155*	-0.132*	-0.122*	-0.127*
Top 80	-0.139*	-0.149**	-0.130*	-0.123*	-0.126*
Top 90	-0.118	-0.124*	-0.101	-0.097	-0.097
Top 100	-0.100	-0.102	-0.083	-0.082	-0.090
Observations	720	756	792	828	864

This table reports the estimated coefficients based on Equation (2) at the aggregate level for universities within each rank group. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Time fixed effects at the weekly level are included in each regression. Standard errors and estimates of other variables are omitted for brevity. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

As junior faculty, researchers in higher-ranked universities often face more pressure to publish papers in order to get tenure and therefore devote more time to research. The constraint caused by the pandemic therefore has a bigger impact on researchers from top-tier universities, resulting in a greater gender inequality among them.

**6.2.3. Effects across countries.** Finally, we examine how the estimated gender inequality varies across countries by replicating the analysis for academics in each country. Figure 2 illustrates the impact on the productivity gap graphically by plotting the estimates of the interacted term with 90-percent and 95-percent confidence intervals; a negative value represents a drop in female academics' research productivity relative to that of male academics. We can observe that most countries—21 out of 25 countries—have experienced a decline in female researchers' productivity. In addition to the US, six countries have shown statistically significant declines: Japan, China, Australia, Italy, the Netherlands, and the United Kingdom. Note that because SSRN is primarily used by US researchers, its preprints for other countries might be limited, which might weaken our ability to detect changes.



**Figure 2** Impact of Lockdown on Gender Inequality across Countries

This graph plots the estimates of the interacted term with 90-percent and 95-percent confidence intervals in each country. The negative values represent female academics' research productivity drop relative to that of male academics.

## 7. Robustness Checks

In this section, we report the results of several robustness tests. Specifically, we check the parallel trends assumption and conduct falsification tests to ensure that our estimated effects are not idiosyncratic. We also test gender inequality in research quality, examine the co-authorship issue, and rule out the data censoring concern.

### 7.1. Parallel trends

The key identification assumption for the DID estimation is the parallel trends assumption; namely, that before the COVID-19 shock, female and male researchers' productivity would follow the same time trend. In Appendix Figure A.1, which presents the time trends of preprints in 2019, visual inspection shows the two genders' parallel evolution before the shock. We then test this assumption by performing an analysis similar to that of Seamans and Zhu (2013) and Calvo et al. (0), in

which we expand Equation (1) to estimate the treatment effect week by week before the shock. Specifically, we replace  $Lockdown_t$  in Equation (2) with the dummy variable  $Time_\tau^t$ , where  $\tau \in \{-14, -13, \dots, -2, -1, 0\}$  and  $Time_\tau^t = 1$  if  $\tau = t$  and 0 otherwise, indicating the relative  $\tau$ th week of the outbreak,

$$\log(Preprint_{igt}) = c + Female_g + \sum_{\tau=-14}^{-1} Time_\tau^t + \sum_{\tau=-14}^{-1} \beta_\tau Female_g \times Time_\tau^t + \delta_i + \epsilon_{igt}. \quad (3)$$

The benchmark group is the week of the pandemic outbreak. The coefficients  $\beta_{-14}$  to  $\beta_{-1}$  identify any week-by-week pre-treatment difference between the female and male researchers, which we expect to be insignificant. We then repeat the same analysis with our aggregate-level data.

Appendix Table A.2 presents the estimation results. They show no pre-treatment differences in the research productivity trends of female and male academics, which supports the parallel trends assumption.

## 7.2. Falsification test

To show that our estimate effects are not an artifact of seasonality, we test whether such a decline in female productivity also occurred in 2019. Appendix Table A.5 reports the summary statistics in 2019. We repeat the same analysis specified in Equation (2) using data in 2019 for the same time window used in 2020. If our results simply capture seasonality, we would be able to find significant effects in 2019. Appendix Table A.6 reports the falsification test results. The placebo-treated average treatment effects are insignificant, implying that women’s productivity did not decline significantly in the previous year.

## 7.3. Research Quality

So far, we study the research quantity— using *Number of Preprints* as the dependent variable. One might question whether the difference in productivity is because, since the lockdown began, male researchers increased the volume of their production at the expense of quality. If this is true, the relative quality of female researchers’ preprints should have increased since the lockdown. We test this possibility using data on how many times the abstract has been viewed and the preprint has been downloaded for each preprint, the two primary quality indicators used by SSRN to rank preprints. Appendix Table A.7 reports the summary statistics of these two variables. We compare the average number of abstract views per preprint and the average number of downloads per preprint for preprints from male and female researchers prior to and after the pandemic outbreak using the same specification as in Equation (1) with the discipline fixed effect:

$$\log(Abstract\ Views_{gt}\ or\ Downloads_{gt}) = c + Female_g + \beta Female_g \times Lockdown_t + \gamma_t + \delta_i + \epsilon_{igt}. \quad (4)$$

**Table 6 Impact of Lockdown on Abstract Views**

Variables	Dependent variable: No. of Abstract Views (in logarithm) by discipline				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	-0.083* (0.047)	-0.083* (0.047)	-0.083* (0.048)	-0.083* (0.048)	-0.083* (0.049)
<i>Female × Lockdown</i>	0.103 (0.085)	0.112 (0.079)	0.090 (0.074)	0.087 (0.070)	0.050 (0.068)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Discipline fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	720	756	792	828	864
$R^2$	0.828	0.829	0.830	0.831	0.834

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (4) using the discipline fixed effect, with  $\log(\text{abstract views})$  as the dependent variable. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table 7 Impact of Lockdown on Downloads**

Variables	Dependent variable: No. of Downloads (in logarithm) by discipline				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	-0.049 (0.042)	-0.049 (0.043)	-0.049 (0.043)	-0.049 (0.043)	-0.044 (0.067)
<i>Female × Lockdown</i>	-0.077 (0.085)	-0.101 (0.077)	-0.111 (0.072)	-0.128* (0.068)	-0.131* (0.065)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	720	756	792	828	864
$R^2$	0.853	0.855	0.855	0.858	0.861

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (4) using the discipline level fixed effect, with  $\log(\text{downloads})$  as the dependent variable. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

Tables 6 and 7 report the the effect of the lockdown on the number of abstract views and the number of downloads, respectively. In general, the average treatment effects are insignificant, suggesting that after the lockdown, female and male researchers’ research quality did not change significantly, and that our findings are unlikely to be driven by shifts in research quality. In addition, column (4) and (5) of Table 7 suggest that female researchers’ research quality in terms of the number of downloads per preprint decreased compared to male researchers’ after the lockdown. We repeat our analysis using Equation 2 and find consistent results. The estimation results are reported in Appendix Tables A.8 and A.9.

#### 7.4. Coauthorship

We next study how coauthorship across genders affects our results. Recall that we measure the *effective* productivity for preprints with  $n$  authors by allocating  $1/n$  preprint to each coauthor. That is, our measure implicitly assumes equal productivity across female and male authors. To alleviate the concern that this assumption may not hold, we conduct a sub-sample analysis focusing on preprints that have either *all-male* or *all-female* authors, excluding preprints that have both

male and female authors. We repeat our DID analysis for this sub-sample using Equations (1) and (2) to compare the productivity gap between all-male and all-female preprints.

Table 8 reports the estimation results with the discipline fixed effect using Equation (1). The results show that the number of all-female preprints has significantly dropped since the lockdown, compared to all-male preprints. Note that the coefficients of  $Female \times Lockdown$  in Table 8 are more significant and larger than those in Table 2, suggesting that gender inequality is more pronounced when a research team has only female authors. Intuitively, an all-female research team's capacity is more severely constrained, resulting in a lower productivity. Table A.4 reports the results at the aggregate-level and the results are consistent.

**Table 8 Impact of Lockdown on Gender Inequality among All-male and All-female Preprints**

Variables	Dependent variable: No. of Preprints (in logarithm) by discipline				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	-1.002*** (0.047)	-1.002*** (0.047)	-1.002*** (0.047)	-1.002*** (0.047)	-1.002*** (0.047)
<i>Female</i> $\times$ <i>Lockdown</i>	-0.233*** (0.088)	-0.263*** (0.084)	-0.267*** (0.079)	-0.255*** (0.076)	-0.232*** (0.074)
Discipline fixed effects	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	720	756	792	828	864
$R^2$	0.810	0.810	0.812	0.813	0.814

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (1) with the discipline level fixed effect. We restrict our sample to preprints that have either all-male authors, or all-female authors. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 7.5. Data Censoring

One might be concerned that because research takes time, many researchers only post one preprint during our sample period. As a result, our data sample before and after the lockdown may contain a different set of authors.

To address this concern, we perform two analyses. First, we collect additional data on US authors, extending the time window of our main analysis to 40 weeks. By adding 16 weeks to the post-treatment period, we capture more researchers in our sample, especially those who were not able to post preprints within 10 weeks of the lockdown. We repeat the DID analysis and find consistent results.

Second, we construct a balanced panel by including only authors who posted preprints before the lockdown to compare the productivity of the same group of authors before and after the lockdown. This approach ensures an apple-to-apple comparison and helps us rule out potential biases introduced by different author samples in the pre-treatment and post-treatment periods. Table 9 reports the estimated results using the balanced panel. The findings are consistent with

our main results: within the same group of authors, female researchers’ productivity dropped significantly after the lockdown compared to male researchers’ productivity.

**Table 9 Impact of Lockdown on Gender Inequality Using the Balanced Panel**

Variables	Dependent variable: No. of Preprints (in logarithm) in aggregation				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	−0.920*** (0.065)	−0.920*** (0.065)	−0.920*** (0.065)	−0.920*** (0.065)	−0.920*** (0.065)
<i>Female × Lockdown</i>	−0.518*** (0.138)	−0.638*** (0.170)	−0.589*** (0.159)	−0.556*** (0.148)	−0.641*** (0.159)
Discipline fixed effects	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	40	42	44	46	48
$R^2$	0.991	0.988	0.988	0.988	0.985

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (2) at the aggregate level. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in Columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 8. Conclusions

Our paper adds to the long-standing literature on gender equality, an important topic in social science. For example, the literature has shown evidence of fairness in parental leaves (Lundquist et al. 2012) and inequality in tenure evaluation (Sarsons 2017, Antecol et al. 2018), recognition (Ghiasi et al. 2015), and compensation (Pierce et al. 2020). Researchers have also investigated business innovations to help empower women (Plambeck and Ramdas 2020). The COVID-19 crisis brings to the forefront a long existing issue: the inequities faced by women who often do more of the childcare and housework. We contribute to the literature by providing direct tests of the impact of the pandemic shock on gender inequality in academia.

We show that in the US, since the lockdown began, women have produced 13.2–13.9-percent fewer social science research papers than men. We also find that the effect is especially significant for junior faculty and for researchers at top-ranked universities. Finally, we find that the increase in productivity inequality is significant in seven countries. The results are robust when we repeat our analysis over papers with same-gender authors, a balanced sample with the same group of authors before and after the lockdown, and an extended data sample.

Our findings indicate that, if the lockdown is kept in place for too long, female academics in junior positions and at top-ranked universities are likely to be significantly disadvantaged—a fairness issue that may expose women to a higher unemployment or career risk in the future. We hope our findings increase awareness of this issue. Actions could be taken to balance domestic responsibilities among spouses. Recently, many universities have taken actions such as granting tenure clock extensions to both female and male faculty. Recall that our paper finds an overall 35-percent increase in productivity and a 13-percent increase in gender gap among social science

researchers. Therefore, our findings do not provoke a concern for overall productivity but rather for gender inequality. As a result, universities could consider providing additional support, such as childcare support, to female researchers whose productivity has been disproportionately affected. Universities and letter writers should keep this inequality in mind when evaluating professors for promotion. We also hope our work will inspire researchers to explore other forms of inequality arising from the COVID-19 pandemic.

Our findings also suggest that telecommuting may have unintended consequences for gender inequality. As the COVID-19 outbreak accelerates the trend toward telecommuting, institutions and firms should take gender equality into consideration when designing and implementing telecommuting policies. We hope that our work could serve as a stepping stone to stimulate more research on the synergy between operations and social issues.

Our study has a few limitations. First, since it focuses on social science disciplines, and thus the findings may not be generalizable to other disciplines. Second, we have limited information about the researchers in our dataset. Future research could collect additional data—such as parental status, whether they are allocating more or less time to research than they did before the pandemic, whether they multitask at home, and who performs household duties—to pinpoint the exact mechanism underlying the observed empirical patterns.

**Acknowledgements:** The authors thank Christopher S. Tang, the anonymous associate editor, and anonymous referees for their constructive and helpful feedback. The authors appreciate the valuable feedback provided by colleagues at the Goizueta Business School.

---

## References

- Abrams, Zara. 2019. The future of remote work <https://bit.ly/2W9LPvI>.
- Alon, Titan M, Matthias Doepke, Jane Olmstead-Rumsey, Michele Tertilt. 2020. The impact of COVID-19 on gender equality. Tech. rep., National Bureau of Economic Research.
- Amano-Patiño, Noriko, Elisa Faraglia, Zeina Hasna. 2020. Who is doing new research in the time of COVID-19? Not the female economists <https://voxeu.org/article/who-doing-new-research-time-covid-19-not-female-economists>.
- Andersen, Jens Peter, Mathias Wullum Nielsen, Nicole L Simone, Resa E Lewiss, Reshma Jagsi. 2020. Meta-research: Is COVID-19 amplifying the authorship gender gap in the medical literature? *arXiv preprint arXiv:2005.06303*.
- Antecol, Heather, Kelly Bedard, Jenna Stearns. 2018. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review* **108**(9) 2420–2441.
- Babcock, Linda, Maria P Recalde, Lise Vesterlund, Laurie Weingart. 2017. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review* **107**(3) 714–47.
- Batt, Robert J, Christian Terwiesch. 2017. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.
- Beck, Dani. 2020. The COVID-19 pandemic and the research lab <https://www.neuro-central.com/the-covid-19-pandemic-and-the-research-lab/>.
- Berry Jaeker, Jillian A, Anita L Tucker. 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* **63**(4) 1042–1062.
- Bianchi, Suzanne M, Liana C Sayer, Melissa A Milkie, John P Robinson. 2012. Housework: Who did, does or will do it, and how much does it matter? *Social Forces* **91**(1) 55–63.
- Bloom, Nicholas, James Liang, John Roberts, Zhichun Jenny Ying. 2015. Does working from home work? evidence from a chinese experiment. *The Quarterly Journal of Economics* **130**(1) 165–218.
- Bray, Robert L, Decio Coviello, Andrea Ichino, Nicola Persico. 2016. Multitasking, multiarmed bandits, and the italian judiciary. *Manufacturing & Service Operations Management* **18**(4) 545–558.
- Calvo, Eduard, Ruomeng Cui, Laura Wagner. 0. Disclosing product availability in online retail. *Manufacturing & Service Operations Management* **0**(0) null. doi: 10.1287/msom.2020.0882. URL <https://doi.org/10.1287/msom.2020.0882>.
- Chesley, Noelle, Sarah Flood. 2017. Signs of change? at-home and breadwinner parents’ housework and child-care time. *Journal of Marriage and Family* **79**(2) 511–534.
- Cui, Ruomeng, Jun Li, Dennis J Zhang. 2020. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science* **66**(3) 1071–1094.

- 
- Dai, Yue, Tianjun Feng, Christopher S Tang, Xiaole Wu, Fuqiang Zhang. 2020. Twenty years in the making: The evolution of the journal of manufacturing & service operations management. *Manufacturing & Service Operations Management* **22**(1) 1–10.
- Dolan, Kathleen, Jennifer Lawless. 2020. It takes a submission: Gendered patterns in the pages of ajps <https://ajps.org/2020/04/20/it-takes-a-submission-gendered-patterns-in-the-pages-of-ajps/#comments>.
- Dutcher, E Glenn, Krista Jabs Saral. 2012. Does team telecommuting affect productivity? an experiment. Tech. rep., Working Papers in Economics and Statistics.
- European Commission. 2016. Women and unpaid work: recognise, reduce, redistribute! <https://ec.europa.eu/social/main.jsp?catId=89&furtherNews=yes&newsId=2492&langId=en>.
- Franck, Thomas. 2020. Us expands iran travel restrictions over coronavirus, raises advisory for south korea and italy <https://www.cnbc.com/2020/02/29/us-expands-iran-travel-restrictions-over-coronavirus-raises-advisory-for-regions-in-south-korea-and-italy.html>.
- Ghiasi, Gita, Vincent Larivière, Cassidy R Sugimoto. 2015. On the compliance of women engineers with a gendered scientific system. *PloS one* **10**(12) e0145931.
- Guarino, Cassandra M, Victor MH Borden. 2017. Faculty service loads and gender: Are women taking care of the academic family? *Research in Higher Education* **58**(6) 672–694.
- Ibanez, Maria R, Jonathan R Clark, Robert S Huckman, Bradley R Staats. 2018. Discretionary task ordering: Queue management in radiological services. *Management Science* **64**(9) 4389–4407.
- KC, Diwas Singh. 2014. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.
- KC, Diwas Singh. 2019. Worker productivity in operations management. *Available at SSRN 3466947* .
- KC, Diwas Singh. 2020. Heuristic thinking in patient care. *Management Science* **66**(6) 2545–2563.
- KC, Diwas Singh, Bradley R Staats, Maryam Kouchaki, Francesca Gino. 2020. Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science*.
- KC, Diwas Singh, Christian Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.
- Kelly, Jack. 2020. The massive work-from-home covid-19 test was a great success and will be the new norm <https://www.forbes.com/sites/jackkelly/2020/05/11/the-massive-work-from-home-covid-19-test-was-a-great-success-and-will-be-the-new-norm/#7240f4094e74>.
- Kitchener, Caroline. 2020. Women academics seem to be submitting fewer papers during coronavirus. ‘never seen anything like it,’ says one editor. <https://www.thelily.com/women-academics-seem-to-be-submitting-fewer-papers-during-coronavirus-never-seen-anything-like-it-says-one-editor>.

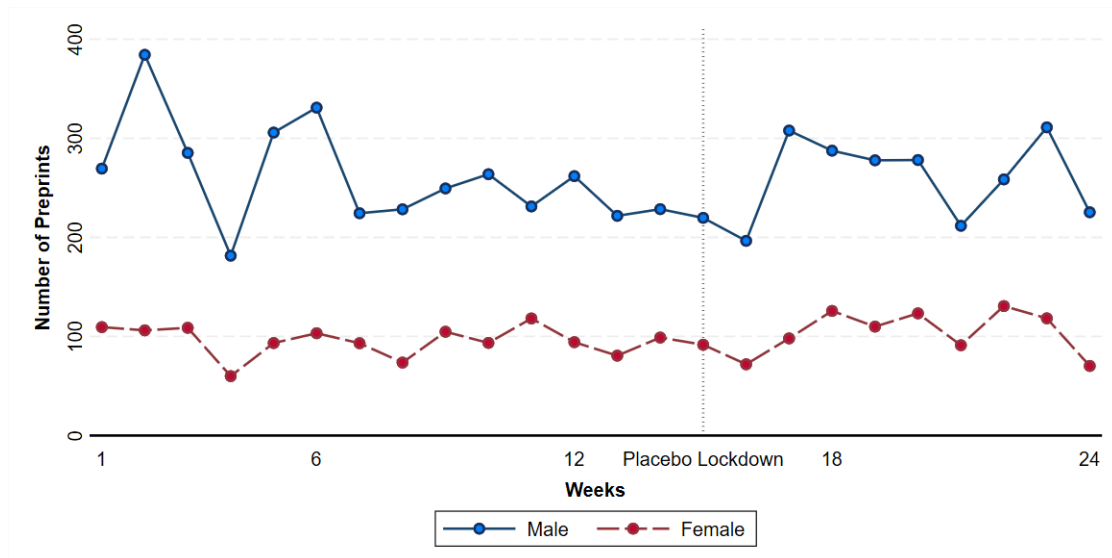


- 
- Lee, Hau L, Christopher S Tang. 2018. Socially and environmentally responsible value chain innovations: New operations management research opportunities. *Management Science* **64**(3) 983–996.
- Levitin, Daniel J. 2014. *The Organized Mind: Thinking Straight in the Age of Information Overload*. Penguin.
- Lundquist, Jennifer H, Joya Misra, KerryAnn O’Meara. 2012. Parental leave usage by fathers and mothers at an American university. *Fathering* **10**(3) 337–363.
- Mark, Gloria, Daniela Gudith, Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110.
- McLean, Rob. 2020. These companies plan to make working from home the new normal. as in forever <https://www.cnn.com/2020/05/22/tech/work-from-home-companies/index.html>.
- Mejia, Jorge, Chris Parker. 2020. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* .
- Minello, Alessandra. 2020. The pandemic and the female academic <https://www.nature.com/articles/d41586-020-01135-9>.
- Misra, Joya, Jennifer Hickes Lundquist, Abby Templer. 2012. Gender, work time, and care responsibilities among faculty. *Sociological Forum*, vol. 27. Wiley Online Library, 300–323.
- Moreno, Edward J. 2020. Government health agency official: Coronavirus ‘isn’t something the american public need to worry about’ <https://thehill.com/homenews/sunday-talk-shows/479939-government-health-agency-official-corona-virus-isnt-something-the>.
- Moss-Racusin, Corinne A, John F Dovidio, Victoria L Brescoll, Mark J Graham, Jo Handelsman. 2012. Science faculty’s subtle gender biases favor male students. *Proceedings of the national academy of sciences* **109**(41) 16474–16479.
- Myers, Kyle R, Wei Yang Tham, Yian Yin, Nina Cohodes, Jerry G Thursby, Marie C Thursby, Peter Schiffer, Joseph T Walsh, Karim R Lakhani, Dashun Wang. 2020. Unequal effects of the covid-19 pandemic on scientists. *Nature Human Behaviour* 1–4.
- National Conference of State Legislatures. 2020. Higher education responses to coronavirus (covid-19) <https://www.ncsl.org/research/education/higher-education-responses-to-coronavirus-covid-19.aspx>.
- Nicklin, Jessica M, Christopher P Cerasoli, Katie L Dydyn. 2016. Telecommuting: What? why? when? and how? *The impact of ICT on work*. Springer, 41–70.
- Pierce, Lamar, Laura W Wang, Dennis J Zhang. 2020. Peer bargaining and productivity in teams: Gender and the inequitable division of pay. *Manufacturing & Service Operations Management*.
- Plambeck, Erica, Kamalini Ramdas. 2020. Alleviating poverty by empowering women through business model innovation: M&som insights and opportunities. *Manufacturing & Service Operations Management*.
- Salvendy, Gavriel. 2012. *Handbook of human factors and ergonomics*. John Wiley & Sons.

- 
- Sarsons, Heather. 2017. Recognition for group work: Gender differences in academia. *American Economic Review* **107**(5) 141–45.
- Schiebinger, Londa, Shannon K Gilmartin. 2010. Housework is an academic issue. *Academe* **96**(1) 39–44.
- Schultz, Kenneth L, David C Juran, John W Boudreau, John O McClain, L Joseph Thomas. 1998. Modeling and worker motivation in jit production systems. *Management Science* **44**(12-part-1) 1595–1607.
- Seamans, Robert, Feng Zhu. 2013. Responses to entry in multi-sided markets: The impact of craigslist on local newspapers. *Management Science* **60**(2) 476–493.
- Song, Hummy, Anita L Tucker, Karen L Murrell, David R Vinson. 2018. Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science* **64**(6) 2628–2649.
- Staats, Bradley R, Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* **58**(6) 1141–1159.
- Tan, Tom, Serguei Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.
- Tan, Tom, Serguei Netessine. 2019. When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance. Forthcoming at Management Science.
- Tan, Tom, Serguei Netessine. 2020. At your service on the table: Impact of tabletop technology on restaurant performance. *Management Science* **66**(10) 4496–4515.
- Tang, Christopher S, Sean Zhou. 2012. Research advances in environmentally and socially sustainable operations. *European Journal of Operational Research* **223**(3) 585–594.
- United States Census Bureau. 2018. 2014-2018 American community survey 5-year estimates <https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2018/5-year.html>.
- U.S. Bureau of Labor Statistics. 2019. Employed persons working at home, workplace, and time spent working at each location by full- and part-time status and sex, jobholding status, and educational attainment, 2019 annual averages <https://www.bls.gov/news.release/atus.t06.htm>.

## Appendix

Figure A.1 Time Trends of US Preprints from December 2018 to May 2019



This graph plots the time trend of the number of preprints for female academics and male academics. The vertical line represents the placebo lockdown week (the week of March 11) in 2019.

**Table A.1 Robustness to Different University Rankings**

Dependent variable: No. of Preprints (in logarithm) by discipline					
Universities by Times ranking	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
Top 10	-0.209***	-0.230***	-0.198***	-0.185***	-0.181***
Top 20	-0.177**	-0.222***	-0.205***	-0.204***	-0.214***
Top 30	-0.227***	-0.253***	-0.228***	-0.228***	-0.228***
Top 40	-0.157**	-0.211***	-0.196***	-0.196***	-0.202***
Top 50	-0.114	-0.147**	-0.130*	-0.138**	-0.146**
Top 60	-0.126*	-0.143*	-0.131*	-0.137**	-0.147**
Top 70	-0.142*	-0.157**	-0.141**	-0.143**	-0.143**
Top 80	-0.139*	-0.154**	-0.140**	-0.131*	-0.130**
Top 90	-0.134*	-0.146**	-0.137**	-0.133*	-0.135**
Top 100	-0.124	-0.129*	-0.125*	-0.118*	-0.118*
Observations	720	756	792	828	864

Dependent variable: No. of preprints (in logarithm) by discipline					
Universities by ARWU ranking	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
Top 10	-0.232***	-0.255***	-0.233***	-0.214***	-0.222***
Top 20	-0.259**	-0.297***	-0.271***	-0.260***	-0.256***
Top 30	-0.261***	-0.305***	-0.268***	-0.264***	-0.259***
Top 40	-0.136*	-0.188**	-0.171**	-0.176***	-0.171***
Top 50	-0.104	-0.156**	-0.132*	-0.133**	-0.139**
Top 60	-0.171**	-0.154***	-0.154***	-0.143***	-0.114*
Top 70	-0.080	-0.125*	-0.109	-0.113*	-0.120*
Top 80	-0.123	-0.128*	-0.117*	-0.118*	-0.120*
Top 90	-0.099	-0.105	-0.095	-0.093	-0.096
Top 100	-0.090	-0.094	-0.086	-0.084	-0.089
Observations	720	756	792	828	864

This table reports the estimated coefficients in Equation (2) across universities with different rankings. The coefficients for 6, 7, 8, 9 and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Time fixed effects at the weekly level are included in all regressions. Note that we omit reporting standard errors and estimates of other variables for brevity. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.2 Parallel Trends Test**

Variables	No. of Preprints (in logarithm) in aggregation		No. of Preprints (in logarithm) by discipline	
	(1)	(2)	(1)	(2)
<i>Female</i> × <i>Time</i> <sub>-14</sub>	-0.231 (0.430)	-0.189 (0.352)		
<i>Female</i> × <i>Time</i> <sub>-13</sub>	-0.013 (0.430)	0.157 (0.335)		
<i>Female</i> × <i>Time</i> <sub>-12</sub>	-0.377 (0.430)	-0.202 (0.309)		
<i>Female</i> × <i>Time</i> <sub>-11</sub>	0.060 (0.430)	0.219 (0.302)		
<i>Female</i> × <i>Time</i> <sub>-10</sub>	-0.030 (0.430)	-0.054 (0.210)		
<i>Female</i> × <i>Time</i> <sub>-9</sub>	-0.028 (0.430)	-0.213 (0.243)		
<i>Female</i> × <i>Time</i> <sub>-8</sub>	-0.144 (0.430)	-0.146 (0.258)		
<i>Female</i> × <i>Time</i> <sub>-7</sub>	-0.101 (0.430)	-0.031 (0.234)		
<i>Female</i> × <i>Time</i> <sub>-6</sub>	-0.363 (0.430)	-0.413** (0.250)		
<i>Female</i> × <i>Time</i> <sub>-5</sub>	0.355 (0.430)	0.314* (0.214)		
<i>Female</i> × <i>Time</i> <sub>-4</sub>	0.130 (0.430)	0.063 (0.224)		
<i>Female</i> × <i>Time</i> <sub>-3</sub>	0.098 (0.430)	-0.051 (0.218)		
<i>Female</i> × <i>Time</i> <sub>-2</sub>	0.069 (0.430)	0.056 (0.239)		
<i>Female</i> × <i>Time</i> <sub>-1</sub>	0.092 (0.430)	0.190 (0.219)		
Observations	24	540		
<i>R</i> <sup>2</sup>	0.894	0.808		

This table reports the estimated coefficients of the parallel trends test using Equation (3). The results at the aggregate-level and discipline-level are presented in Columns (1) and (2), respectively. Note that we omit reporting estimates of other variables for brevity. Time fixed effects at the weekly level are included in all regressions. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.3 Impact of Lockdown on Gender Inequality by Academic Ranks in Aggregation**

Researchers by Academic Ranks	Dependent variable: No. of Preprints (in logarithm) in aggregation				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
Student	-0.030	-0.042	-0.039	-0.061	-0.091
Assistant Prof.	-0.485***	-0.384**	-0.408**	-0.383**	-0.405**
Associate Prof.	0.046	-0.013	0.016	-0.029	-0.019
Full Prof.	-0.175	-0.154	-0.055	-0.035	-0.052
Observations	40	42	44	46	48

This table reports the estimated coefficients based on Equation (2) for researchers within each rank group. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Time fixed effects at the weekly level are included in each regression. Standard errors and estimates of other variables are omitted for brevity. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.4 Impact of Lockdown on Gender Inequality among All Male or All Female Preprints in Aggregation**

Variables	Dependent variable: No. of Preprints (in logarithm) in aggregation				
	6 weeks (1)	7 weeks (2)	8 weeks (3)	9 weeks (4)	10 weeks (5)
<i>Female</i>	-1.253*** (0.066)	-1.253*** (0.066)	-1.253*** (0.066)	-1.253*** (0.066)	-1.253*** (0.065)
<i>Female</i> × <i>Lockdown</i>	-0.285** (0.099)	-0.297*** (0.092)	-0.254** (0.096)	-0.233** (0.093)	-0.220** (0.089)
Discipline Fixed Effects	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	40	42	44	46	48
$R^2$	0.978	0.979	0.978	0.978	0.979

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (2) at the aggregate level. We restrict our sample to those preprints that have either all-male authors, or all-female authors. The coefficients for 6, 7, 8, 9, and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.5 Summary Statistics for December 2018 - May 2019**

Level	Weekly No. of Preprints	All observations					Before March 2019		After March 2019	
		Mean	Std. dev	Max	Min	Total	Mean	Std. dev	Mean	Std. dev
All	All	401.0	69.6	535	267	9,333	406.4	75.8	393.3	58.9
Disciplines (US only)	Female authors	103.0	17.2	131	62	2,413	102.1	15.1	104.4	19.7
	Male authors	298.0	57.9	424	205	6,920	304.3	65.7	288.9	42.7
By Discipline (US only)	Accounting	21.0	6.3	34	10	505	21.9	6.6	19.9	6.2
	Anthropology	76.3	19.9	115	41	1,832	69.4	20.9	86.1	14.0
	Cognitive	17.0	7.7	38	7	407	20.5	7.9	12.0	3.7
	Corporate	17.5	5.9	30	8	420	17.2	5.6	17.9	6.4
	Criminal	16.3	5.6	32	6	390	14.9	6.4	18.2	3.8
	Economics	212.0	50.9	348	133	5,089	225.7	55.7	192.9	37.9
	Education	15.3	5.2	29	6	366	15.3	5.2	15.2	5.6
	Entrepreneurship	16.1	5.6	28	8	387	18.7	5.3	12.5	3.6
	Finance	89.7	21.3	148	66	2,153	95.0	25.2	82.3	11.8
	Geography	13.6	6.3	29	5	327	11.9	4.9	16.0	7.5
	Health Economics	4.3	4.2	22	0	104	3.3	1.7	5.8	6.1
	Information Systems	20.2	5.8	36	10	485	22.0	6.4	17.7	3.9
	Law	143.1	32.6	211	76	3,434	135.4	36.3	153.8	24.4
	Management	32.4	11.8	57	8	778	34.7	11.1	29.2	12.5
	Organization	24.8	7.8	43	15	594	27.2	8.4	21.3	5.7
Political Science	166.3	28.3	225	124	3,991	172.5	30.9	157.6	22.8	
Sustainability	38.8	23.9	105	14	930	34.1	16.7	45.2	31.3	
Women/Gender	19.4	8.4	40	4	466	20.9	9.9	17.4	5.8	

The table summarizes the weekly number of papers from December 2018 to May 2019. In total, there are 9,333 preprints produced by 14,767 US authors, 2,413 of which are produced by 3,876 female researchers and 6,920 are produced by 10,891 male researchers. We gather the country-specific lockdown time to split our sample to before and after the lockdown for each country.

**Table A.6 Falsification Test**

	Dependent variable: No. of Preprints (in logarithm) in aggregation				
	6 weeks	7 weeks	8 weeks	9 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
<i>Female</i> × <i>Lockdown</i>	0.042	0.061	0.088	0.080	0.057
Observations	40	42	44	46	48
$R^2$	0.980	0.980	0.979	0.980	0.980

	Dependent variable: No. of Preprints (in logarithm) by discipline				
	6 weeks	7 weeks	8 weeks	9 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
<i>Female</i> × <i>Lockdown</i>	0.092	0.094	0.103*	0.085	0.070
Observations	720	756	792	828	864
$R^2$	0.877	0.877	0.871	0.873	0.873

This table reports the estimated coefficients of the interacted term, *Female* × *Lockdown*, in Equation (2). The coefficients for 6, 7, 8, 9 and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Note that we omit reporting estimates of other variables for brevity. Time fixed effects at the weekly level are included in all regressions. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.7 Summary Statistics for Downloads and Abstract Views**

Level	Groups	All observations				Before Lockdown		After Lockdown	
		Mean	Std. dev	Min	Max	Mean	Std. dev	Mean	Std. dev
No. of downloads per preprint	All	40.9	18.6	13.7	84.6	54.2	14.6	26.9	10.2
	Female authors	39.2	21.1	10.2	85.6	53.0	18.1	24.6	12.5
	Male authors	41.7	18.6	14.9	84.2	54.8	14.6	27.8	10.7
No. of abstract views per preprint	All	144.5	47.6	57.67	226.1	184.0	19.4	102.7	29.2
	Female authors	139.1	49.0	44.7	243.1	176.4	26.6	99.6	34.1
	Male authors	146.8	48.2	62.1	232.3	187.1	19.8	104.1	28.6

The table summarizes the weekly average number of downloads and abstract views per preprint from December 2019 to May 2020. The sample includes 9,934 preprints from authors in the United States.

**Table A.8 Impact of Lockdown on Abstract Views**

Variables	Dependent variable: No. of Abstract Views (in logarithm) in aggregation				
	6 weeks	7 weeks	8 weeks	9 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
<i>Female</i>	-0.054 (0.048)	-0.054 (0.048)	-0.054 (0.048)	-0.054 (0.048)	-0.054 (0.048)
<i>Female</i> × <i>Lockdown</i>	0.086 (0.074)	0.088 (0.068)	0.074 (0.065)	0.067 (0.062)	0.044 (0.058)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	40	42	44	46	48
$R^2$	0.894	0.913	0.935	0.948	0.955

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (2), with *abstract views* as the dependent variable. The coefficients for 6, 7, 8, 9 and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

**Table A.9 Impact of Lockdown on Downloads**

Variables	Dependent variable: No. of Downloads (in logarithm) in aggregation				
	6 weeks	7 weeks	8 weeks	9 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
<i>Female</i>	-0.044 (0.067)	-0.044 (0.067)	-0.044 (0.067)	-0.044 (0.067)	-0.044 (0.067)
<i>Female</i> × <i>Lockdown</i>	-0.027 (0.175)	-0.057 (0.157)	-0.068 (0.141)	-0.085 (0.130)	-0.087 (0.120)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	40	42	44	46	48
$R^2$	0.836	0.866	0.891	0.910	0.927

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (2), with *downloads* as the dependent variable. The coefficients for 6, 7, 8, 9 and 10 weeks since the lockdown are presented in columns (1)–(5), respectively. Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

# On state- and time-dependent service processes in healthcare

Hao Ding<sup>1</sup>

Ruomeng Cui<sup>1</sup>

Donald K.K. Lee<sup>1,2</sup>

<sup>1</sup>Goizueta Business School, Emory University

<sup>2</sup>Department of Biostatistics & Bioinformatics, Emory University

**Abstract** The empirical operations literature has shown that the service process of customers is state-dependent, with a primary focus on the relationship between patient length-of-stay (LOS) and workload in healthcare settings. However, the service process is also time-dependent because physician workload is not a static state, but varies continuously over the course of a patient's stay. By recognizing that the service process is both state- and time-dependent, we argue that the instantaneous service rate is the more natural quantity of interest in place of LOS. Conducting analyses based on the instantaneous service rate rather than LOS allows us to potentially resolve a longstanding question regarding the relationship between service speed and workload. Specifically, the empirical literature has identified contradictory findings among analyses based on LOS. We show, through the setting of an academic emergency department, that a consistent and intuitive relationship emerges when the analysis is based on the instantaneous service rate. We also use a naturalistic simulation to demonstrate meaningful performance gains from employing a service rate model to set staffing, relative to staffing according to results from analyses based on LOS.



---

## 1 Introduction

Understanding the drivers of efficiency in healthcare service processes has long been of interest to the Operations Management (OM) community, especially the determinants of efficiency in the Emergency Departments (ED). A key finding in the literature is that services rates are state-dependent. In the context of healthcare, this means the productivity of individual physicians, and therefore the service rates of individual patients, depend on the states of the physicians and the ED. For example, workload has been shown to have a significant impact on ED service speeds, although there is less agreement on how. Indeed, a wide spectrum of relationships have been reported that range from slowdown as workload increases (Lucas et al., 2009; Armony et al., 2015; Batt et al., 2017), to workload having no effect (Lucas et al., 2009; McCarthy et al., 2009), to speed up (Kc and Terwiesch, 2009; Anderson et al., 2011), to slow down and then speed up (Batt and Terwiesch, 2014; Kuntz et al., 2015), and to still more complicated patterns (Berry Jaeker and Tucker, 2017). These findings stand in stark contrast to theories from the neuroscience and psychology literature, which unequivocally suggest that multitasking leads to a decrease in performance. Human brains are not designed to handle multiple tasks simultaneously (Worringer et al., 2019). Multitasking requires frequent switching between tasks, which reduces efficiency and performance (Pashler, 1994; Sigman and Dehaene, 2006). It also increases the production of stress hormones such as cortisol and adrenaline, overstimulating the brain and leading to mental fatigue, reduced focus, and impaired cognitive abilities (Dux et al., 2006).

The complicated relationship between workload and service speed is curious for two reasons. First, why do we observe vastly different empirical workload effects within the OM literature? Second, theories in other fields such as psychology and neuroscience predict that a higher workload should have a simple yet intuitive relationship with service speed—a higher workload should slow workers down. Why do the findings in the OM literature differ from this prediction?

To explain the disparate findings in the literature, we hypothesize that a potential source comes from the resolution of patient flow data. Previous works typically use a single snapshot of the ward census (at patient arrival or discharge) or its average level as the measure of workload for the patient’s entire stay. Preliminary evidence has since emerged to suggest that service speeds may vary dynamically with the state of the system represented by variables such as workload (Armony et al., 2015). Thus, even on the same dataset, the observed relationship between speed and workload

---

at the time of arrival, may be quite different from the one between speed and workload at the time of discharge. Second, the literature often focuses on the relationship between service speed and one particular dimension of workload, with most work looking at the aggregate ward census (system workload). However, other dimensions may also matter, perhaps even more so: the service speed for a particular patient may also depend on the loads of their assigned health providers. Given that provider workloads are correlated with the system workload, current estimates of the latter’s impact may be confounded if the former are excluded from the analysis, with the direction of confounding varying from one study to the next according to the case mix of provider loads.

This work attempts to improve our empirical understanding of the workload effect on speed by analyzing patient flow through the ED at a higher resolution. In particular, we take advantage of a novel data set from an academic ED in the United States that tracks workload in real time from April 2017 to March 2019. Workload data sampled at this frequency have not been exploited before in the literature, and it allows us to empirically model the service rate  $\mu(t, X_t)$  for a patient as a function of their time-varying variables  $X_t$  (e.g., workload). To take advantage of real-time workload data, we use time-dependent survival analysis to estimate patient service rates  $\mu(t, X_t)$  as a function of the patient’s time-in-service  $t$  and the time-dependent state variable  $X_t$ . Specifically, we first take advantage of a recent survival machine learning approach (Lee et al., 2021) to perform an exploratory analysis on the first year of data non-parametrically. Then we confirm the findings using parametric survival analysis, which is more established in the literature.

We build up the analysis from static workload to time-varying workload to demonstrate the importance of the time-varying nature. In particular, we first follow the literature and use static workload measures in the analysis, in which we show results similar to those shown in the literature. When using static workload, the service speed appears to have a complicated relationship with the workload/multitasking level. Next, we analyze the same data, but by replacing static workload measures with the actual time-varying workload measures. The results reveal a simple and intuitive relationship that agrees with the neuroscience and psychology literature—patient service speed is monotonically negatively associated with workload level. In addition to the theories specifically pertaining to workload and multitasking, we also develop a simple model to demonstrate that the average workload level could lead to misleading results. We show that with one patient, under prevalent conditions, using average workload to estimate the probability of a patient being discharged could lead to misleading results.

---

Finally, we perform naturalistic simulations to demonstrate and prescriptive value of the state- and time-dependent framework. In particular, we use real-world conditions to show how average workload can lead to suboptimal staffing levels in the ED and to prescribe a usable tool for hospital managers. To achieve this, we simulate three months of ED operations to two service rate functions: one derived from the state-dependent model and the other from the state- and time-dependent model, respectively. For each model, we assess the average patient service time at various physician staffing levels. This allows us to determine the optimal staffing levels based on the economic trade-off between increased ED revenue through faster patient service time and the cost of physician staffing. Beyond identifying optimal staffing levels from simulations, we also extract the optimal physician level from the OLS model. We then compare the optimal levels from each of the three models—the state- and time-dependent, the state-dependent, and the OLS models—using out-of-sample data. Our findings reveal that state- and time-dependent model outperforms the others, leading to a financial advantage of \$2.7MM and \$3.5MM.

In conclusion, this study highlights that time-dependency, akin to state-dependency, is a first order concern. We substantiate this claim theoretically, empirically, and prescriptively. This paper focuses on workload, a widely examined “state” in existing literature. Through this lens, we accomplish two main objectives: first, we illustrate the necessity of incorporating time-dependency to model service rates accurately. Second, we provide a plausible explanation that could potentially reconcile the mixed findings prevalent in workload-related studies. Furthermore, while our focus is on workload and in a healthcare context, the framework we propose has broader implications. It can be generalized to other service processes reliant on human interaction, such as those in call centers. As high-resolution data becomes more readily available to practitioners and researchers, the applicability and utility of our framework are likely to continue to expand.

## 2 Literature Review

In this section, we review the workload and multitasking literature. We first examine the OM literature, focusing on the mixed findings within the literature and the workload measures employed. Next, we discuss studies on workload and multitasking from neuroscience and psychology literature. Researchers in these disciplines have explored how multitasking impacts human brain and cognitive processing. The consensus across these studies is that multitasking tends to hinder rather than enhance our efficiency, resulting in slower performance. This reveals a gap in the literature between

empirical findings in operations and theoretical predictions from neuroscience and psychology.

Note that we discuss workload and multitasking literature together as they share commonalities both in practice and in the literature. In practice, as workload in hospitals increases, physicians find themselves caring for more patients, leading to higher degrees of multitasking (Berry Jaeker and Tucker, 2017). In the literature, these two concepts often employ analogous measurements. For example, at the physician level, both workload and multitasking levels are frequently gauged by the count of patients assigned to a given physician. Furthermore, both streams of literature focus on state-dependency in service rate, highlighting the variability of productivity, its drivers, and their underlying mechanisms.

## 2.1 Workload and Multitasking in OM

This section presents the literature on the impact of workload and multitasking on service rates, a key topic in OM research. Table 1 provides a summary of selected studies investigating the relationship between workload and patient service rate. Although we strive for a comprehensive review, the extensive nature of workload literature means that our coverage cannot be complete.

**Table 1: Workload Literature**

Paper	Workload level	Workload Measure	Results
<b>Workload Literature</b>			
Kc and Terwiesch (2009)	bed occupancy	snapshot	speed-up then slow-down
Kuntz et al. (2011)	bed occupancy	snapshot	slow-down
Anderson et al. (2011)	bed occupancy	daily capacity	speed-up
Kc and Terwiesch (2012)	bed occupancy	snapshot	speed-up
Tan and Netessine (2014)	server workload	grouped by check	slow-down then speed-up
Kuntz et al. (2015)	system workload	daily capacity	slow-down then speed-up
Armony et al. (2015)	system workload	daily capacity	slow-down
Batt and Terwiesch (2017)	system workload	multiple snapshots	inverted U-shape
Berry Jaeker and Tucker (2017)	system workload	multiple snapshots	N-shape
Batt et al. (2019)	physician workload	hourly workload	
Xu et al. (2021)	banker workload	multiple snapshots	

This table summarizes a selection of paper on the effect of workload. Note to be consistent with the focus of this paper, which is service speed, the results are translated into the effect of workload on service speed.

The initial assumption in the healthcare queuing literature was that service time—the time required to care for a patient—is independent of the system’s current state, including workload (Kc and

---

Terwiesch, 2009). Subsequent experimental evidence began to reveal that the processing time of human workers might actually depend on the system's state, including factors like inventory levels and the pace of their coworkers (Schultz et al., 1998). Powell and Schultz (2004) further develop models to represent this state-dependent behavior in production workers. Their models demonstrate a significant shift in effects when considering state-dependency: longer production lines, previously thought to slow workers down under the assumption of state-independent behavior, are shown to actually accelerate workers' processing rates when their state-dependent behavior is accounted for.

Empirical evidence soon highlighted workload as a crucial state of the system. The first wave of studies presented contradictory relationships, indicating that higher workload leads to either a speed-up or slow-down effect on service rate. One of the first empirical evidence comes from Kc and Terwiesch (2009). Focusing on patient transport services and cardiothoracic surgery, the authors demonstrate that increased workload could enhance processing speed at the individual level, but this speed-up effect was found to be unsustainable. Anderson et al. (2011) examine this phenomenon at a system level (i.e., bed utilization), and observe that higher workloads often resulted in faster patient discharges, effectively reducing the length-of-stay (LOS). This pattern is also observed in an Intensive Care Unit (ICU) setting (Kc and Terwiesch, 2012). But their paper noted a downside: higher workload led to more patients returning to the ED, ultimately reducing the overall service rate. In contrast, Kuntz et al. (2011) report that higher workload actually slows down patient discharge, attributing this slowdown to mental fatigue and queuing for resources.

Recent research then reveals more intricate relationship between workload and service rate. Namely, a U-shape relationship. In other words, as workload increases, patient LOS first increases due to the congestion effects, then decreases as the workload reaches a threshold, after which the physician may elect to discharge patients faster to free up the space (Kc, 2014). In a restaurant setting, similar findings have shown that as servers' workload increases, customers' meal duration first increases and then decreases (Tan and Netessine, 2014).

Given these diverse findings, OM scholars have been actively working to reconcile the literature. The focus has been on explaining different findings by understanding the mechanism driving workload effect. For instance, Batt and Terwiesch (2017) propose that the observed instability of the aggregate effect at the system level might be due to multiple server-level mechanisms working simultaneously. They argue that one should view complex systems like ED as multi-stage processes, instead of single stage ones. Their research supports the notion of a U-shaped impact of workload

---

on patient service rate. Another study by Berry Jaeker and Tucker (2017) investigates how the effect of workload changes as workload increases and approaches the system capacity. Their findings reveal an inverted N-shape relationship, with two distinct tipping points. Initially, service rates decrease as the workload rises, which authors attribute to congestion effects. Then, as workload increases further, service rates increase temporarily as healthcare workers strive to discharge patients more quickly, potentially by working harder or cutting corners. However, as workload approaches maximum capacity, service rates decline again, indicating the system’s inability to compensate for extremely high workloads.

This paper aims to bridge the gaps in the existing literature by developing a state- and time-dependent framework to study the effect of workload in a time-dependent manner. In contrast to the focus of Batt and Terwiesch (2017) and Berry Jaeker and Tucker (2017), our study concentrates on the temporal aspect of workload. We acquire high-resolution data to construct time-varying workload measures and analyze the workload effect in a time-dependent survival analysis. Our results highlight that time-dependency is a first order concern in accurately assessing the effects of workload, offering a new dimension to understanding the complex dynamics of such effect.

Since this paper primarily focuses on estimating service processes in a time-dependent manner, our discussions of the literature have focused on workload measure and its impact on service speed. It is worth noting that these studies also explore outcomes beyond service rate and investigate drivers other than workload. For instance, the past research has examined patient outcomes in hospitals (Kc, 2014; Kuntz et al., 2015; Berry Jaeker and Tucker, 2017), efforts to up-sell in restaurants (Tan and Netessine, 2014), and operational error rates in banks (Xu et al., 2021), among other factors. In terms of the independent variable, Batt et al. (2019) focus on how physician shifts impact efficiency and quality. Thus, while the primary focus of this paper is on estimating service processes in a time-dependent manner, the broader literature encompasses a range of outcomes and drivers beyond just workload.

## **2.2 Workload and Multitasking in Psychology and Neuroscience**

Multitasking has been extensively studied in the fields of psychology and neuroscience as well. Researchers have investigated its effects on cognitive processes, attention, and performance. In this literature, researchers use experiments to study the effect of workload and multitasking on speed and accuracy in completing simple tasks. Most studies have shown that high workload and multitasking

---

can impair performance and productivity. To explain the findings, neuro-imaging techniques such as functional magnetic resonance imaging (fMRI) has been used to study how human brains respond to working on multiple tasks. Neuroimaging have revealed that when individuals multitask, there is increased activation in brain regions associated with attention, such as the prefrontal cortex (PFC). However, this does not necessarily translate into improved performance since increasing demands on attention and cognitive resources can lead to overall decreased performance.

Two main mechanisms have been established to explain the slowdown from multitasking. First, the bottleneck in human information processing. As humans, we are not build to multitask. In fact, when processing information, we can only process information sequentially and not in parallel (Worringer et al., 2019). Numerous behavior experiments have shown that when an individual is presented with two tasks in parallel, response to the second task becomes significantly delayed (Pashler, 1994; Sigman and Dehaene, 2006). With modern techonology, studies have also been able to explain the root of this bottleneck, with theories favoring the frontal areas (IFJ) (Dux et al., 2006), parietal areas (Sigman and Dehaene, 2008), and finite neural resources as a limitation factor (Stelzel et al., 2006).

Another key mechanism behind the slowdown from multitasking comes from task-switching, when we alternate between two or more tasks repeatedly or in successions. Compared to performing a single task, having to switch between two tasks significantly reduces the participants' performance with respect to both speed and accuracy. This slowdown is true even when participants have to switch between two simple cognitive tasks in fixed sequence (Rogers and Monsell, 1995). Performance further decreases if participants have to switch between more than two tasks and when facing tasks in random sequences (Monsell, 2003; Kiesel et al., 2010). In addition, while the effect is stronger when intervals between tasks are shorter, switch costs remain significant even with longer tasks or longer intervals between tasks (Rogers and Monsell, 1995). Similar to the bottleneck effect, researchers have also managed to explain why task-switching slows us down. In essence, switching between two tasks requires additional cognitive resources to allow both task sets to be maintained and worked on without interference between each other. Or, as Kc (2014) argue in their revolutionary paper on the effect of multitasking on patient service rate—multitasking can result in poorer throughput due to mental strain and increased setup times when switching between patients.

In summary, the psychology and neuroscience literature has more consistent findings when it comes to workload and multitasking. Compared to the operations literature, this stream of literature

predicts that high workload and resulting multitasking will slow down workers. Their theory predictions and findings based on lab environments and simple tasks are currently not in line with empirical findings in real-world settings in the OM literature.

### 3 Theory

In this section, we discuss the necessity of using time-varying workload measures estimate instantaneous service rate from a theory perspective. First, we explore how using static workload measures to estimate LOS may yield misleading results in general. Then, we present a stylized model that illustrates the potential inaccuracies arising from the application of various static workload measures, even in a simple setup.

#### 3.1 General Theory

**Ground truth.** Let  $T$  be the service completion time for a patient. The instantaneous service speed experienced by the patient at time  $t$ , whose current state is represented by a vector of covariates  $X_t$ , is

$$\mu(t, X_t). \quad (1)$$

Mathematically, the service rate is equivalent to the hazard rate in survival analysis. In our context, this is the instantaneous probability of completing service in the time interval  $[t, t + dt)$  conditional on still being in service at  $t$ :

$$\mu(t, X_t)dt \approx P(T \in [t, t + dt) | T \geq t, X_t). \quad (2)$$

Combining this fundamental quantity with knowledge of the evolution dynamics of  $X_t$ , we can completely specify the data generating process for the patient’s service time. To avoid onerous technical details, we keep things simple by treating the covariate trajectory  $X_{(\cdot)}$  as fixed, and let it represent workload alone. Then from the start of treatment at  $t = 0$ , the service duration is determined by successive coin tosses at infinitesimal increments  $t = 0, dt, 2dt, \dots$ , and service completion occurs when the first heads is tossed. It follows from (2) that the probability of tossing



a heads at time  $t$  is  $\mu(t, X_t)dt$ , so the likelihood of observing the service complete at time  $T$  is

$$\begin{aligned} & \{1 - \mu(0, X_0)dt\} \times \{1 - \mu(dt, X_{dt})dt\} \times \cdots \times \mu(T, X_T) \\ & \xrightarrow{dt \downarrow 0} e^{-\int_0^T \mu(t, X_t)dt} \mu(T, X_T), \end{aligned}$$

where the limit is a product integral.

Given this setup, the impact of workload on service speed at a given point in time  $\tau$  is naturally the derivative

$$\mu^{(0,1)}(\tau, X_\tau) := \left. \frac{d}{d\alpha} \mu(\tau, \alpha) \right|_{\alpha=X_\tau}. \quad (3)$$

Note that this ground truth for the workload effect (3) does not depend on the workload level at any other point in time. As we shall see, this is not necessarily the case when it comes to the measure of workload effect commonly used in the literature.

**Existing literature.** Until very recently, real-time workload data was not readily available to researchers, so existing studies focus on time-static proxies of service speed such as LOS. Hence existing efforts to quantify the workload effect is via its influence on the expected LOS

$$\mathbb{E}[T|X_{(\cdot)}] = \int_0^\infty \mathbb{P}(T > t|X_{(\cdot)})dt = \int_0^\infty e^{-\int_0^t \mu(u, X_u)du} dt.$$

In other words, the interest is in the change in expected LOS after perturbing the workload trajectory by  $dX_{(\cdot)}$ , i.e.  $X_{(\cdot)} \mapsto X_{(\cdot)} + \alpha dX_{(\cdot)}$  for  $\alpha \ll 1$ . This is essentially the Gateaux derivative

$$\begin{aligned} \partial_X LOS(dX) &:= \left. \frac{d}{d\alpha} \mathbb{E}[T|X_{(\cdot)} + \alpha dX_{(\cdot)}] \right|_{\alpha=0} \\ &= - \int_0^\infty \left\{ e^{-\int_0^t \mu(u, X_u)du} \int_0^t dX_s \cdot \mu^{(0,1)}(s, X_s) ds \right\} dt. \end{aligned} \quad (4)$$

If it is somehow possible to only increase workload at time  $\tau$  without changing the workload at any other point in time, then  $dX_t$  can be represented by a Dirac delta distribution  $\delta_\tau(t)$  centred at  $\tau$ ,

i.e.  $\int \delta_\tau(t)f(t)dt = f(\tau)$  for a well-behaved function  $f$ . In which case

$$\begin{aligned}\partial_X LOS(\delta_\tau) &= - \int_0^\infty \left\{ e^{-\int_0^t \mu(u, X_u) du} \int_0^t \delta_\tau(s) \mu^{(0,1)}(s, X_s) ds \right\} dt \\ &= - \int_0^{\tau^-} \left\{ e^{-\int_0^t \mu(u, X_u) du} \cdot 0 \right\} dt - \int_{\tau^+}^\infty \left\{ e^{-\int_0^t \mu(u, X_u) du} \cdot \mu^{(0,1)}(\tau, X_\tau) \right\} dt, \\ &= -\mu^{(0,1)}(\tau, X_\tau) \underbrace{\int_\tau^\infty e^{-\int_0^t \mu(u, X_u) du} dt}_{>0},\end{aligned}$$

so the direction is consistent with the natural measure of workload effect (3): If  $\mu^{(0,1)}(\tau, X_\tau) < 0$ , this implies that increasing workload will slow down service, hence expected LOS should increase, i.e.  $\partial_X LOS(\delta_\tau) > 0$ . Conversely if  $\mu^{(0,1)}(\tau, X_\tau) > 0$  we would expect  $\partial_X LOS(\delta_\tau) < 0$ .

However, all existing studies are based on observational data, so it is not possible to localize the change in workload to a single neighbourhood of time. Imagine a simple example whereby an increase in workload at  $\tau$  is correlated with a decrease in workload at  $\tau' \neq \tau$  in the observed data. Then the variation used to identify the workload effect at  $\tau$  might resemble  $dX_t = \delta_\tau(t) - \delta_{\tau'}(t)$ , so by linearity

$$\begin{aligned}\partial_X LOS(dX) &= -\mu^{(0,1)}(\tau, X_\tau) \int_\tau^\infty e^{-\int_0^t \mu(u, X_u) du} dt \\ &\quad + \mu^{(0,1)}(\tau', X_{\tau'}) \int_{\tau'}^\infty e^{-\int_0^t \mu(u, X_u) du} dt.\end{aligned}$$

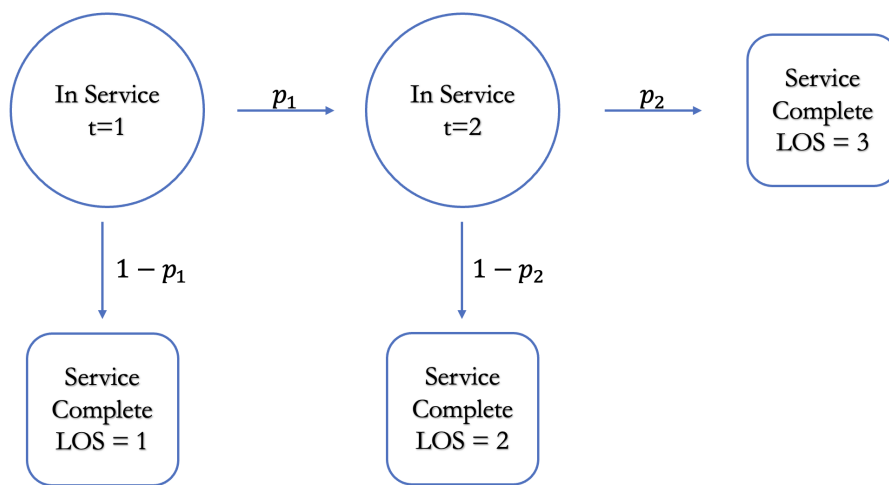
If an increase in workload at a given point in time always slows down service ( $\mu^{(0,1)} < 0$ ), but  $\mu^{(0,1)}(\tau', X_{\tau'})$  is sufficiently more negative than  $\mu^{(0,1)}(\tau, X_\tau)$  is, then the change in the expected LOS will be negative, i.e. service speedup. Similarly, if an increase in workload actually speeds up service ( $\mu^{(0,1)} > 0$ ), it is possible for the change in the expected LOS to be positive, i.e. service slowdown.

This simple example illustrates that the inability to time-localize the workload variation can cause the LOS-based workload effect (4) to diverge from the ground truth (3). On the other hand, as alluded to earlier, the ground truth measure  $\mu^{(0,1)}(\tau, X_\tau)$  is unaffected by whether the variation in workload at  $\tau$  is correlated with the variation in workload at another point in time. By acknowledging that the service process is time-dependent and thus focusing on the instantaneous service rate, we can avoid issues associated with LOS-based analyses.

### 3.2 Stylized Model

We next present a stylized discrete-time model to demonstrate that using the starting workload, average workload, or ending workload can lead to misleading service speed relationships.

In this model, a patient commences the service at  $t = 0$ . One time unit later at  $t = 1$ , the patient either completes the service with probability  $1 - p_1$ , or continues treatment into the next period with probability  $p_1$ . At  $t = 2$ , the patient either completes the service with probability  $1 - p_2$ , or continues treatment into the next period with probability  $p_2$ . The patient will always complete the service and leave the hospital by  $t = 3$ . Figure 1 depicts the dynamics of this service process.



**Figure 1: A stylized service process.**

Note that at  $t \in \{1, 2\}$ , the discrete-time service rate  $\mu_t$  is  $1 - p_t$ . Hence if  $p_t$  represents the workload in period  $t$ , then as the workload increases, the patient will have a lower service rate and is less likely to complete the service in that period. From the perspective of patient length-of-stay  $LOS$ , this quantity follows the distribution

$$LOS = \begin{cases} 1 & \text{w.p. } 1 - p_1 \\ 2 & \text{w.p. } p_1(1 - p_2) \\ 3 & \text{w.p. } p_1p_2 \end{cases}$$

Conditional on the workloads  $(p_1, p_2)$  for all periods, the patient's expected *LOS* is

$$\begin{aligned}\mathbb{E}(LOS|p_1, p_2) &= 1 - p_1 + 2p_1(1 - p_2) + 3p_1p_2 \\ &= 1 + p_1 + p_1p_2,\end{aligned}\tag{5}$$

Hence the expected *LOS* is increasing in the workloads  $p_1$  and  $p_2$ , i.e. higher workloads slow down service.

Now suppose that  $(p_1, p_2)$  is drawn from a simple distribution parameterized by two parameters  $0 \leq v, w \leq 1/2$ , and which admits just two possible outcomes:

$$(p_1, p_2) = \begin{cases} (1/2 - v, 1/2 + w) & \text{w.p. } 1/2 \\ (1/2 + v, 1/2 - w) & \text{w.p. } 1/2 \end{cases}\tag{6}$$

The expected *LOS* conditional on only the starting workload  $p_1$  is

$$\mathbb{E}(LOS|p_1) = \begin{cases} 7/4 - vw + (w - 3v)/2 & \text{if } p_1 = 1/2 - v \\ 7/4 - vw - (w - 3v)/2 & \text{if } p_1 = 1/2 + v \end{cases}$$

When  $w - 3v > 0$ , the expected *LOS* is decreasing in (starting) workload, which misleadingly conveys the opposite of the true relationship (5). The conditional expected *LOS* does not depend on  $p_1$  at all when  $w - 3v = 0$ , which is also misleading. Note that  $\mathbb{E}(LOS|p_1)$  is the true conditional mean and not an estimated one.

The same issue occurs when the discrete-time service rate  $\mu_t$  at  $t = 2$  is conditioned on only the starting workload  $p_1$ :

$$\mu_2(p_1) \equiv \mathbb{P}(LOS = 2 | LOS \geq 2, p_1) = \begin{cases} 1/2 - w & \text{if } p_1 = 1/2 - v \\ 1/2 + w & \text{if } p_1 = 1/2 + v \end{cases}$$

which suggests that the service speed at  $t = 2$  increases with the starting workload.

Using the average workload  $\bar{p}$  can also lead to misleading relationships. Bearing in mind that  $\bar{p} = p_1$

if service is completed at  $t = 1$ , and  $\bar{p} = (p_1 + p_2)/2$  otherwise,

$$\mathbb{E}(LOS|\bar{p}) = \begin{cases} 1 & \text{if } \bar{p} = \bar{p}^{(1)} \equiv 1/2 - v \\ 5/2 + w & \text{if } \bar{p} = \bar{p}^{(2)} \equiv \{1 - (v - w)\}/2 \\ 5/2 - w & \text{if } \bar{p} = \bar{p}^{(3)} \equiv \{1 + (v - w)\}/2 \\ 1 & \text{if } \bar{p} = \bar{p}^{(4)} \equiv 1/2 + v \end{cases}$$

For  $v - w > 0$  we have  $\bar{p}^{(1)} < \bar{p}^{(2)} < \bar{p}^{(3)} < \bar{p}^{(4)}$ , so the expected  $LOS$  is initially increasing in  $\bar{p}$ , and then decreasing: Recall that  $0 \leq w \leq 1/2$ , so  $1 < 5/2 + w > 5/2 - w > 1$ . For the service rate at  $t = 2$  we have

$$\mu_2(\bar{p}) \equiv \mathbb{P}(LOS = 2 | LOS \geq 2, \bar{p}) = \begin{cases} 1/2 - w & \text{if } \bar{p} = \{1 - (v - w)\}/2 \\ 1/2 + w & \text{if } \bar{p} = \{1 + (v - w)\}/2 \end{cases}$$

which suggests that service speed is increasing in  $\bar{p}$ .

By now one might expect that using the ending workload  $p_{end}$  can also lead to misleading relationships, and this is indeed the case:

$$\mathbb{E}(LOS|p_{end}) = \begin{cases} 5/2 - w & \text{if } p_{end} = 1/2 - w \\ 1 & \text{if } p_{end} = 1/2 - v \\ 1 & \text{if } p_{end} = 1/2 + v \\ 5/2 + w & \text{if } p_{end} = 1/2 + w \end{cases}$$

For  $w > v$  the expected  $LOS$  is initially decreasing in  $p_{end}$  and then increasing, but for  $w < v$  this relationship is reversed.<sup>1</sup>

To sum up, what we see here is that using static summaries of the inherently dynamic workload can lead to problems even in a model as simple as this. The real service process in the ED is far more complicated, so one would expect the real life situation to be far more prone.

<sup>1</sup>For service rates, it does not make sense to condition on  $p_{end}$ . This is because at a given time  $t$ , knowledge of the ending workload would require knowing when the visit will end, which requires looking into the future. For the current service rate to depend on a quantity from the future would violate the causal arrow of time.

---

## 4 Data

Next, we turn to real-world data to test whether estimating instantaneous service rate using time-varying workload matters empirically. To achieve this, we acquire a high-resolution dataset from the ED of a prominent academic hospital in the United States, covering all patient visits between April 2017 and March 2019. This dataset provides a detailed snapshot of patient interactions, tracking events in realtime throughout each visit. For every patient, the dataset captures the exact moment of arrival, the initiation of treatment, their assignment to specific healthcare professionals, and their exit from the ED.

In this section, we discuss the construction of i) Control variables; ii) Static workload variables used in the existing literature; and iii) Time-varying workload variables used in this study.

**Control variables.** In our models, we incorporate a comprehensive set of control variables that capture patient-specific, temporal, and physician shift related factors. Patient demographics and temporal controls are implemented in all analyses. The physician shift controls are inherently time-varying, and therefore only employed in the time-dependent survival analysis.

For patient demographics, we account for each patient’s gender, race, age, and associated clinical information at arrival, such as the chief complaint and the Emergency Severity Index (ESI). The ESI is a measure of the patient’s illness severity, with level 1 being the most severe and level 5 the least. We exclude ESI level 1 cases given their critical nature and the fact that they are treated in a specialized area. At the other end of the spectrum, we combine ESI levels 4 and 5 into a single category as they both reflect relatively low severity.

We then introduce the temporal controls. The exact time of a patient’s arrival, including month, day of the week (e.g., Friday), and hour (e.g., noon), which have been shown to significantly influence patient discharge timings. The month and day of arrival are both implemented as control variables without further modification. For the hour of patient arrival, we segment the control variables into four categories: midnight to 11am, 11am to 2pm, 2pm to 6pm, and 6pm to midnight. The control variables are segmented based on the feature importance from the survival machine learning tool. Leveraging the ability of ML in determining feature importance and improve the construction of control variables has become more popular and standard, we provide relevant details in later sections where we discuss the survival ML tool we employ in this paper.

Finally, we also incorporate time-varying controls that measure the time into a physician’s shift

following Batt et al. (2019), and also the time a patient has spent with their assigned physician. To illustrate, if physician  $j$  begins her shift at noon, by 2 pm she would be two hours into her shift. Similarly, if patient  $i$  is assigned to physician  $j$  at 1pm, by 2pm the patient would be one hour into his treatment with the physician. The shift controls are deployed in a continuous fashion (i.e., we do not discretize the shift controls).

### Static workload measures

Most studies on the workload effect and multitasking employ linear regression models to estimate the dependence of patient LOS on static workload measures. Given the limited availability of real-time workload data in the past, workloads are often assessed at two pivotal moments: Upon commencement of treatment, and at discharge. The average of these two roughly approximates the time-weighted average workload over a patient’s stay. We align with the literature by defining the following static workload measures for each ED visit:

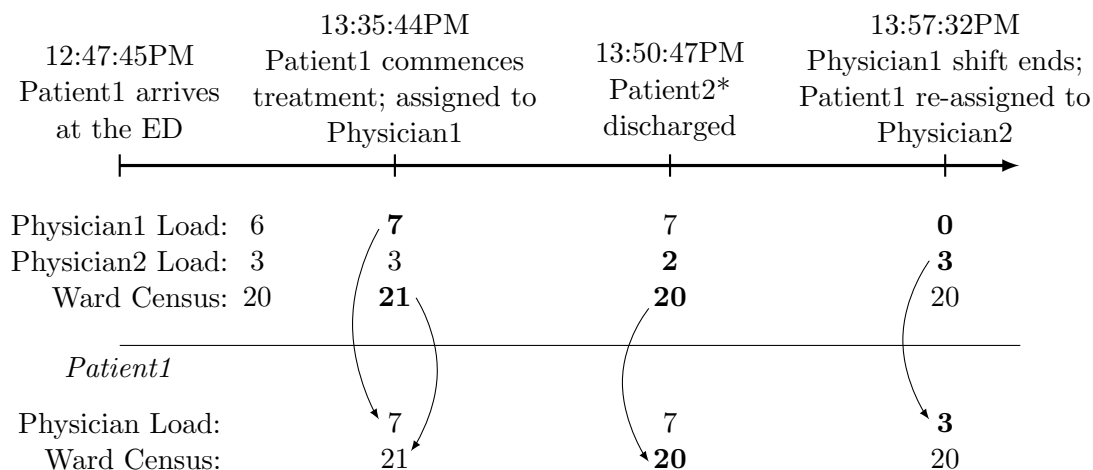
- *Arrival workload*: The workload of the patient’s assigned physician at the time when treatment commenced. This can shape treatment decisions, which can in turn influence LOS.
- *Discharge workload*: The workload of the patient’s assigned physician at the last timestamp of the patient’s visit to the ED. Along with the overall ED occupancy at that time, these can sway the decision-making process regarding discharge.
- *Average workload*: The average of the arrival and discharge workloads. This serves as a summary of the overall workload pressure experienced during the visit.

### Time-varying workload measures

For the time-dependent survival analysis setting, we leverage the granularity of our data to construct realtime workload measures for individual physicians and the aggregate ward census. In essence, the workload for physician  $j$  at time  $t$  is the number of patients assigned to  $j$  at that moment. Then the physician workload for patient  $i$  at time  $t$  is simply the workload of the physician assigned to patient  $i$  at that point in time. The ward census at time  $t$  is defined as the number of patients receiving service at that time. These workload measures can change when a new patient commences service, a patient gets discharged, or when the patient is assigned to a different physician.

Figure 2 illustrates how we construct the time-varying workload measures. Patient 1 arrives to the ED at 12:47:45PM and begins treatment at 13:35:44PM under Physician 1. At this point,

Patient 1’s physician workload is set as Physician 1’s workload of 7, and the ward census is 20. By 13:50:47PM, Patient 2, managed by Physician 2, gets discharged, altering Patient 1’s ward census but not his physician workload. When Physician 1’s shift ends at 13:57:32PM, Patient 1 is handed over to Physician 2. This changes the patient’s physician workload measure to 3, but the ward census remains unchanged.



**Figure 2: An Illustration of Workload Measure**

## 5 Empirical Analysis

In this section we describe the empirical strategies used to estimate how workload affects service speed, followed by the reporting of results. We start by replicating the approach used in the literature: Regressing LOS onto static workload measures via ordinary least squares (OLS). We then transition to a time-dependent survival analysis framework to account for the fact that in reality, the workload measures as well as the service rate itself are not static, but change over the course of a visit. Our results suggest that the instantaneous service rate depends heavily on the dynamic workload, and therefore, modeling such idiosyncrasy can more accurately capture the reality.

### 5.1 OLS Analysis

To benchmark against the empirical results from previous research, we apply the approach used in the literature to our dataset. Specifically, we employ a linear regression model to estimate the



impact of the static workload measures on patient LOS. We show that our results are consistent with a subset of findings identified in the existing literature. The results also serve as a baseline for comparison to the time-varying workload analysis in the next subsection, where the time-dynamic nature of workload is accounted for, and where the outcome of interest is the instantaneous service rate rather than LOS.

The OLS specification for patient  $i$ 's ED visit is:

$$LOS_i = \beta_0 + \beta_1 Phys_i + \beta_2 Phys_i^2 + \beta_3 Ward_i + \beta_4 Ward_i^2 + X_i' \gamma + \varepsilon, \quad (7)$$

where  $LOS_i$  is the patient's length of stay in the ED (minutes),  $Phys_i$  is the workload of the physician assigned to the patient, and  $Ward_i$  is the ward census. The vector  $X_i$  represents the patient-specific control variables described in Section 4. The quadratic workload terms serve two purposes. First, they align with the existing literature that find a non-linear relationship between workload and LOS, as discussed in Section 2. Second, the inclusion of quadratic effects can potentially reveal an inflection point of where the physician workload level is at an optimum.

We estimate (7) separately for each of the three static workload definitions in Section 4: Arrival, Discharge, and Average.

**Results.** We first present the estimation results for (7), but without including the higher order quadratic terms in the estimation. What is reported in Table 2 largely aligns with the established literature: All columns show that an increase in physician workload is associated with a decrease in patient LOS (i.e. faster service speed). However, the magnitude of this effect varies substantially across the static workload measures used, with more than a five-fold difference when the arrival workload measure is used versus the discharge one.

Relative to the physician workload effect, the ward census effect is an order of magnitude smaller, which also concurs with existing findings in the literature. Moreover, the effect varies all the way from being a slow-down effect, to being insignificant, to being a speed-up effect as the type of workload measure used is changed.

Table 3 reports the estimation results for the full model (7), including the quadratic workload terms. The results are in line with the restricted linear model, in that a higher physician workload continues to be associated with a speedup in service, as evidenced by the negative coefficients for the linear workload terms. However, the inclusion of the quadratic terms now provide a more

**Table 2: Static workload effects on patient LOS (w/o quadratic workload terms)**

Variables	Arrival (1)	Discharge (2)	Average (3)
Physician workload	−0.0043*** (0.0005)	−0.0223*** (0.0006)	−0.0155*** (0.0006)
Ward census	0.0007 (0.0006)	0.0040*** (0.0004)	−0.0030*** (0.0007)
Intercept	5.611*** (0.029)	5.764*** (0.029)	5.816*** (0.024)
<i>N</i>	42,253	42,253	42,253

This table reports the estimated coefficients and robust standard errors (in parentheses). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

refined description: The physician workload effect on LOS is convex decreasing over the applicable workload range from 1 to 15 (the observed maximum physician workload in the data). In other words, the marginal productivity gain from increasing physician workload appears to attenuate.

Similar to the results from the restricted linear model, the ward census effect is inconclusive under the full model (7). While the effect is insignificant for the arrival and discharge census measures, the average census has an inverted U-shaped relationship with LOS. In other words, increasing the average census is initially associated with a slowdown in service speed, but flips to become a speedup past a certain point.

**Table 3: Static workload effects on patient LOS (w/ quadratic workload terms)**

Variables	Arrival (1)	Discharge (2)	Average (3)
Physician workload	−0.0118*** (0.0013)	−0.0659*** (0.0019)	−0.0136*** (0.0024)
(Physician workload) <sup>2</sup>	0.0003*** (0.0000)	0.0014*** (0.0001)	0.0000 (0.0001)
Ward census	−0.0012 (0.0027)	0.0009 (0.0026)	0.0458*** (0.0021)
(Ward census) <sup>2</sup>	0.0000 (0.0000)	0.0000 (0.0000)	−0.0020*** (0.0000)
Intercept	5.692*** (0.058)	6.088*** (0.056)	5.626 (0.030)
<i>N</i>	42,253	42,253	42,253

This table reports the estimated coefficients and robust standard errors (in parentheses). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 5.2 Survival Analysis

The analysis in the previous subsection follows the predominant approach in the literature, where the outcome of interest is patient LOS. However, two patients with the same LOS in the ED

could have had experienced vastly different service histories. For example, one patient might have commenced treatment during peak ED hours, and hence had a slow start but quick service towards the end. The other patient might have experienced fast service in the beginning but slow service towards the end. It is therefore more informative to estimate the service speed at every given point in time instead of measuring overall LOS. To this end, we adopt a time-dependent survival analysis framework to estimate the instantaneous service rate function

$$\mu(t, X_{it})$$

for patient  $i$  who has already spent  $t$  units of time in treatment, with associated covariates  $X_{it}$  at that point in time. These covariates include the time-varying physician workload and ward census, physician shift controls, and the time-static controls employed in the OLS analysis. It is worth remembering here that these are covariates tied to the patient’s visit, so the patient’s physician workload refers to the workload of the physician assigned to the patient at time  $t$ , for example.

Motivated by the findings in Brown et al. (2005) and Armony et al. (2015) that suggest a log-normal distribution for service durations, we model the functional form of  $\mu(t, X_{it})$  as a log-normal hazard function:

$$\mu(t, X_{it}) = \phi\left(\frac{\log t - \theta(X_{it})}{\sigma}\right) / \left\{ \sigma t \bar{\Phi}\left(\frac{\log t - \theta(X_{it})}{\sigma}\right) \right\} \quad (8)$$

where  $\phi(\cdot)$  and  $\bar{\Phi}(\cdot)$  are the density function and complementary CDF of the standard normal distribution respectively. Specifying  $\theta(X_{it})$  linearly as  $X'_{it}\theta$  recovers the canonical accelerated failure time (AFT) regression model for the log-normal family.

To provide intuition for how to interpret the coefficients  $\theta$ , let us consider a time-static covariate version of the log-normal AFT model for a single covariate  $x$ . In this case an equivalent form of the model is

$$\log T = \theta x + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \quad (9)$$

where  $T$  is the service duration. Viewed through (9), a unit increase in  $x$  prolongs the duration by a factor of  $e^\theta$ , so the service speedup factor is  $e^{-\theta}$ . An implication is that a negative coefficient implies speedup, and a positive coefficient implies slowdown. Another implication is that a linear covariate will have a log-linear effect on the service rate  $\mu(t, X_{it})$ .

**Results.** We estimate two versions of the service rate (8), with the results presented in Table 4.

As an intermediate step between the static OLS analysis and the fully time-dependent survival analysis, we first estimate  $\mu(t, X_{i0})$  as a function of the static arrival workload.<sup>2</sup> We then perform the fully time-dependent estimation of  $\mu(t, X_{it})$  as a function of the time-varying workload, to allow the service rate to respond dynamically to changes in the workload.

For the static arrival workload, Column 1 suggests that the service rate increases with physician workload, and decreases in the ward census. These findings are qualitatively the same as what we discover from the OLS analysis. To interpret the log-normal coefficients, the service rate increases by  $e^{-(-0.0205)} - 1 = 2\%$  for every unit increase in physician workload. Whereas adding one more patient to the ward slows down the service of all patients by just  $1 - e^{-0.0016} = 0.16\%$ .

On the other hand, the fully time-dependent analysis reverses the relationships. Column 2 indicates that the instantaneous service rate actually decreases by  $1 - e^{-0.0167} = 1.7\%$  for every unit increase in the contemporaneous physician workload. For ward census, the service rate now increases by  $e^{-(-0.0028)} - 1 = 0.28\%$  for every unit increase in ward census.

**Table 4: Log-normal service rate estimation of (8)**

Variables	Static workload measures (1)	Time-varying workload measures (2)
Physician workload	−0.0205*** (0.0008)	0.0167*** (0.0006)
Ward census	0.0016*** (0.0006)	−0.0028*** (0.0006)
log $\sigma$	−0.4220*** (0.0064)	−0.4210*** (0.0064)
Intercept	6.6540*** (0.0348)	6.3640*** (0.0351)
All controls	Yes	Yes
$N$	2,322,705	2,322,705

This table reports the estimated coefficients and robust standard errors (in parentheses). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

The findings from the time-dependent analysis align with the theoretical expectations from the psychology and neuroscience literature. Moreover, the static workload survival analysis suggests that these findings are not simply an artefact of switching the estimation approach from an OLS analysis to a survival one. The material change occurs after fully accounting for the time-dependent nature of workload, holding constant the estimation procedure. As a robustness check to further ensure that the results are not driven by the parametric log-normal assumption, we re-estimate

<sup>2</sup>It does not make sense to model the service rate as a function of the discharge workload, because at any given time  $t$  before service completion, the discharge workload is unknown. Hence the service rate at  $t$  cannot possibly depend on it. Since the average workload also depends on the discharge workload, the same reasoning applies.

the service rates nonparametrically to see if the workload effects remain the same. We employ a survival machine learning technique known as BoXHED

## 6 Application to ED Staffing

Recognizing that the service process is state- and time-dependent also has prescriptive implications for managers of service systems. In this section we use a naturalistic staffing simulation to illustrate the extent to which this refinement can improve upon service process models that only account for state-dependency.

At a high level, we construct a simulation of an ED that is fed by the historical arrival of patients to the study ED during April to June 2018. Three versions of the service process are estimated from the patient visit histories from this period:

- i)  $\mu_{\text{LOS}}$ : In lieu of the service rate, we model the LOS of patient  $i$ 's visit using the regression model (7). Since only the static arrival workload is known at the start of treatment, we use this as the workload measure when simulating the LOS for a visit.
- ii)  $\mu_{\text{static}}$ : The instantaneous service rate  $\mu(t, X_{i0})$  based on the log-normal specification (8), but as a function of the static arrival workload instead of the current workload.
- iii)  $\mu_{\text{dynamic}}$ : The fully state- and time-dependent service rate  $\mu(t, X_{it})$  based on (8).

Equipping the simulation with each estimated service process, we solve for the economically optimal physician staffing level under each scenario. We then apply these staffing recommendations to the test period from April to June 2019 in the simulation, in order to assess the hospital's performance gains when staffing to the fully state- and time-dependent model  $\mu_{\text{dynamic}}$  relative to the other two.

### 6.1 Simulation Setup

We construct a virtual ED that closely mirrors our studied ED. We simulate the service time—the duration from the start of service to patient discharge—for each patient arriving at the ED during a three-month period. Patients arrive at this simulated ED following the pattern of their real-world arrival between April and June 2018, which includes 10,150 patient visits in our data, with their real-world characteristics, such as age, sex, and ESI level. We then simulate ED states and patient service time using the service rate functions derived from the survival analysis.

---

We examine two scenarios through simulations, drawing on two different service rate functions derived from the survival analysis. Specifically, the first scenario incorporates the function based on the time-varying workload analysis, while the second scenario uses the function derived from the static workload analysis. In both scenarios, we utilize the same real-world ED arrival data and patient characteristics, maintaining consistency in these variables. The only aspect we alter is the physician staffing level, to assess its impact on patient service time.

In each simulation run, upon arrival, patients are assigned to a physician using the round-robin process, a standard practice at the study ED (Traub et al., 2016; Hodgson and Traub, 2020). The round-robin approach sequentially assigns patients to physicians, disregarding their current workload. This approach is preferred by most healthcare centers as it promotes fairness in workload distribution and curtails potential free-riding behaviors. Once a patient has been assigned, treatment commences as the physician becomes available. Each physician can attend up to 10 patients simultaneously. If a physician has reached this capacity limit, subsequent patients must wait.

Once the treatment begins, our model continually estimates and updates the estimated service time throughout a patient’s stay in the ED. For example, when patient X starts treatment at time  $t_1$  with physician K, we estimate the patient’s service time for the first time using the service rate function derived from the survival analysis. At time  $t_2$ , when another patient under physician K is discharged, it changes physician K’s workload pertaining to patient X. Accordingly, we re-estimate patient X’s service time based on the updated workload level. We continue this process, updating the service time estimated for patient X every time there’s a change in the associated workload. This approach accounts for two critical components of our time-varying survival analysis: the evolving workload during a patient’s stay and the accumulated time patient X has spent in the ED up to any given moment,  $t$ .

We initiate the simulation with three physicians. For each subsequent simulation iteration, we add a physician to the ED’s busiest shift, which runs from 7am to 7pm. [Ruomeng: need to explain the staffing level set for the non-busy shift.] Given that the patient arrival rate remains consistent across different simulation runs, changes in the physician staffing level directly impact the average workload in the ED. It’s worth noting that we begin with three physicians in the simulation as this is the minimum number necessary to ensure all 10,150 patients are attended to within the three-month simulation duration. Using the average estimated service time obtained from the simulations, we can then evaluate the costs associated with each staffing level.

**Economic impact on the ED.** To assess the economic impact of physician staffing levels, we weigh the costs of hiring physicians against the potential revenue gains from shorter patient service times and subsequent improved throughput in the ED.

Hiring additional physicians represents a significant expenditure for the hospital. The cost is estimated at about \$860,000 annually, based on the median yearly salary for physicians, which stood at \$229,300 in 2022 (U.S. Bureau of Labor Statistics, 2022). To cover a 12-hour shift, it's necessary to hire the equivalent of 2.5 Full-Time Equivalents (FTEs)—a metric representing the workload of a full-time employee. However, the total expenditure isn't solely salary. When we factor in additional costs such as overhead, physician benefits, and related costs for other resources, expenses accumulate. To accommodate these extra costs, we add a 50% increment to the base salary. This calculation results in an expense of  $\$229,300 \times 2.5 \times 1.5 = \$859,875$  to cover a 12-hour shift, which we round to \$860,000 for simplicity.

Despite the hiring costs, additional physicians can enhance the efficiency of the ED by reducing the time to serve each patient. This improved efficiency can lead to significant revenue gains. As per industry findings by Becker's Hospital Review (2016), an ED with 30,000 yearly visits could see an increase of roughly \$1.4 million in additional revenue by reducing the average service time by 15 minutes. Corroborating these findings, a study by Pines et al. (2011) identifies similar financial advantages, estimating benefits in the range of \$2.7 to \$3.6 million, or equivalently \$1.8 to \$2.5 million when adjusted to the size of our ED. Collectively, these insights suggest potential revenue gains of about \$2.3 million for every 15-minute reduction in average service time, based on the dollar value for 2021. As both studies were conducted in a linear fashion, we estimate that each minute's reduction in service time corresponds to \$153,333 in revenue gains.

## 6.2 Simulation Results

In this section, we begin by discussing results from the simulation. Based on the estimated service times from the simulation, we then present their associated economic implications and identify the optimal staffing level prescribed by each service rate function. Finally, we compare the optimal staffing level identified by different models, including the OLS model, to quantify the economic implications arising from different workload measures.

**Patient service time** We begin by examining the simulation results for patients service times, as illustrated in Figure 3. The distinctive trends are evident, with the two different service functions

yielding opposite outcomes. In the context of the time-varying service function, the average time taken to serve a patient goes down as we have more physicians and less workload per doctor. On the contrary, increasing the number of doctors leads to an increased service time for each patient, suggesting the ED becomes less efficient. This finding is consistent with empirical observations from the survival analysis that link higher static workloads and faster patient service times.

**Economic Implications** We next investigate the economic implications associated with different physician staffing levels and workload. To reiterate, our economic analysis hinges on two key factors: the physician staffing cost and the revenue implications as a result of the average patient service times. Figure 4 and 3 translate the effects of physician staffing level on patient service time into financial implications.

In the first scenario based on the time-varying service function, the trade-off between the cost of hiring additional physicians and the revenue gains from higher ED throughput is shown in Figure 4. Taking the cost for staffing three physicians as the benchmark, we observe a growth in revenue, which peaks when the staffing level reaches seven physicians, and then begins to decline thereafter.

Figure 5 showcases the economic impacts based on the static service function. In this scenario, the previously mentioned trade-off is absent. This distinction is rooted in the findings in Figure 3, where having fewer physicians does not increase the time it takes to serve a patient; it actually decreases it. Therefore, the maximum revenue naturally is at the lowest physician staffing level, which is three.

In addition to determining the optimal physician staffing level from our simulation, we also compute the optimal number using the OLS results showcased in Table 3. Following Kc (2014), we calculate the optimal number using the coefficients of the linear workload term and the quadratic term. The optimal number is therefore  $0.0118 / (0.0003 * 2) = 19.67$ , which we round down to 19 physicians. In our data, the upper staffing limit is around 15 during the peak shift. While we recognize 19 as the optimal number derived from the OLS model, for practical considerations, we also regard 15 physicians as an alternative optimal level.

**The optimal optimal staffing level.** Finally, we compare the optimal staffing levels identified by different models and quantify the economic implications of employing different workload measures. To do so, we perform the same simulation, using real-world patient arrival data from the same three months period, from April to June in 2019, during which time there were 9,123 patient visits



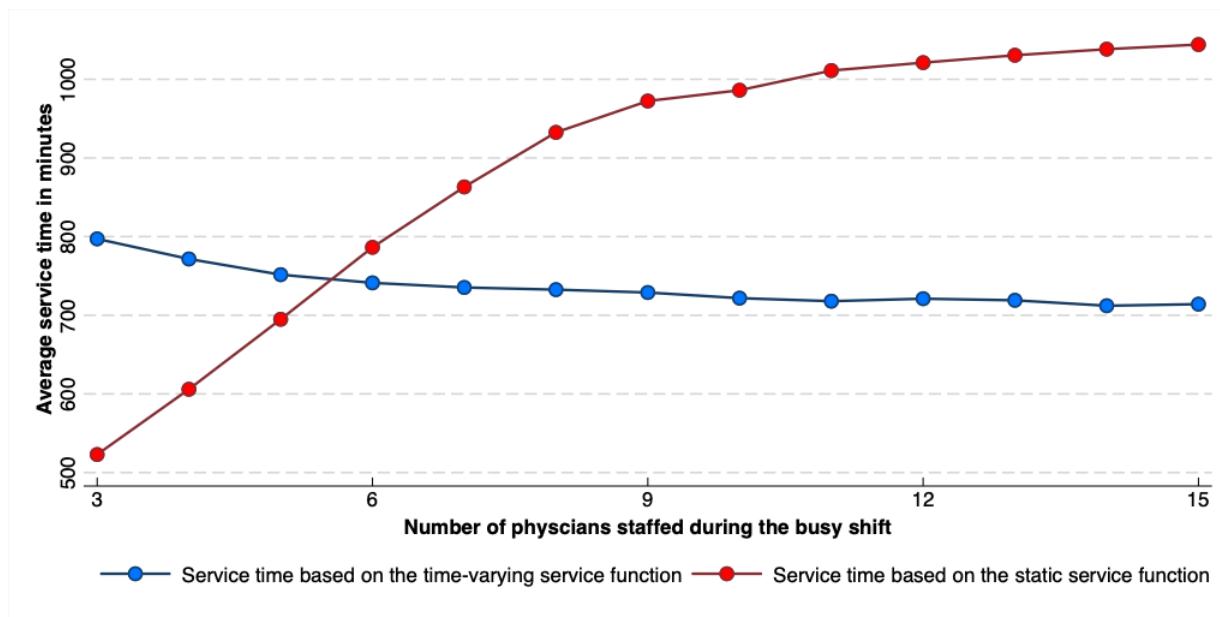


Figure 3: Curve of simulated average service time per patient

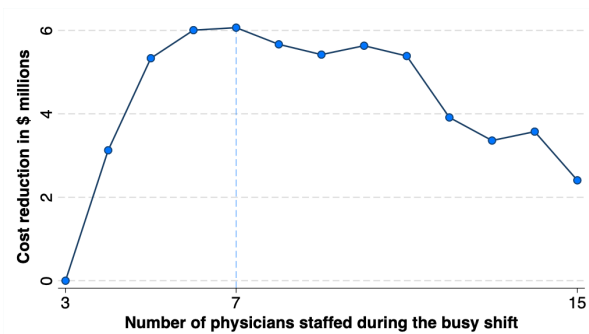


Figure 4: Time-varying revenue trends

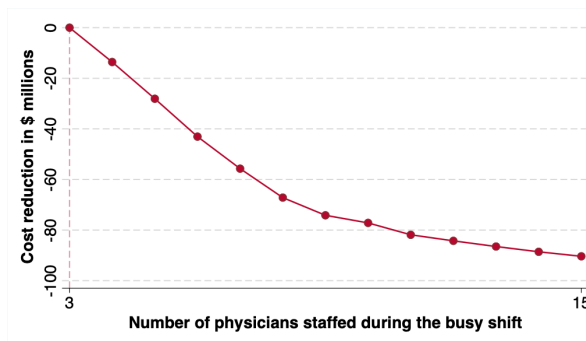


Figure 5: Static revenue trends

at the study ED. In this step, we run the simulations using the time-varying service function to estimate the patient service times at the number of physicians optimized by different models.

In particular, we quantify the differences in ED revenues based on the optimal level proposed by three different models. Namely from the simulations using the time-varying service function, and the ones using the static service function, and the optimal number of staff derived from the OLS with quadratic terms. The corresponding staffing levels are 7, 3, 15, and 19 physicians, respectively.

Table 6 reports the results. Columns (1) to (4) report the simulation results of physician staffing levels of 3, 7, 15, and 19, respectively, each corresponding to the optimal staffing level derived from the simulation using the static service function, the time-varying service function, and obtained

from the OLS with quadratic terms. Comparing the average service time across staffing levels, we can see a clear trend: as the number of physicians increases, the average service time decreases. However, in terms of the revenue, the optimal is achieved with seven physicians, which corresponds to the optimal level from the simulation, using the time-varying service function. This makes sense as we can see that the marginal gain in service speed diminishes as the number of physicians increases.

**Table 5: Costs savings comparison across different specifications**

Prescribed by Staffing level	static 3 (1)	time-varying 7 (2)	OLS 15 (3)	OLS 19 (4)
Avg. service time	777	737	715	706
ED Revenue	\$-2,693,320	\$0	\$-3,506,674	\$-5,566,677
Patients seen	7,094	9,123	9,123	9,123

This table reports the cost savings at the simulated ED using the out-of-sample data from April 2019 to June 2019. Average patient service time is in minutes. ED revenue is relative to the revenue of staffing seven physicians.

**Table 6: Costs savings comparison across different specifications**

Prescribed by Staffing level	static 3 (1)	time-varying 7 (2)	OLS 15 (3)
Avg. service time	777	737	715
ED Revenue	\$-2,693,320	baseline	\$-3,506,674
Patients seen	7,094	9,123	9,123

This table reports the cost savings at the simulated ED using the out-of-sample data from April 2019 to June 2019. Average patient service time is in minutes. ED revenue is relative to the revenue of staffing seven physicians.

In summary, the staffing level determined by the time-varying model notably outperforms other models, which aligns with our expectations. When compared against the optimal level derived from the simulation using the static service function and the two staffing levels from the OLS model, the time-varying model’s prescribed staffing level outperforms them by \$2.7MM, \$3.5MM, and \$5.6MM, respectively—all of which are economically significant.

## 7 Discussion

The need for managers of service systems to account for the state-dependency of service rates is by now well understood, thanks to the empirical evidence documented in the Operations Management literature (Kc and Terwiesch, 2009; Batt and Terwiesch, 2017; Berry Jaeker and Tucker, 2017). This paper is among the first to clarify why one also needs to account for the time-

---

varying nature of the state variables, which turns out to be of first-order concern in the estimation of the service rate function used for prescriptive planning. The fact that the sign of the estimated workload effect flips when time-dependency is ignored suggests that state-dependency and time-dependency interact, so one needs to account for both at the same time. This has important implications for the literature, future research, and for managers.

For the workload literature, our findings indicate that time-dependency could potentially be driving the seemingly contradictory results in the existing literature. We hope this work will inspire researchers to examine how workload affects service rate in a time-dependent way using our framework or other frameworks that would allow one to model workload time-dependently. From a theory perspective, our findings call for researchers to incorporate the time-dependency aspects of workload into analytical models and develop tools to estimate optimal staffing levels for service processes, such as the ED.

For the broader literature, the spirit of our approach applies to other scenarios where time-varying data is available. Our framework offers a way for researchers to estimate the effect of an independent variable that changes with time. In particular, for situations where the underlying distribution is less known, BoXHED, the non-parametric machine learning tool could be used to first estimate the distribution of service rate to determine the specification of the survival analysis. Then the parametric analysis can be performed to obtain the magnitude (i.e., coefficients) of the effect of interest, and validate the results with more robustness tools, such as the frailty model.

For managers, our findings highlight that service workers, such as physicians in the ED, may not be able to accommodate high workload levels as the literature previously suggested. When planning for future staffing levels, our results highlight the importance of reconsidering the relationship between workload and service rate. This has important implications in staffing decisions. For example, instead of relying on the assumption that physicians can speed-up service rates as workload increases and the ED becomes congested, hospital managers might want to plan ahead for higher staffing levels to ensure timely patient treatment. As demonstrated in our simulation exercise, taking the time-dependency of workload into consideration could lead to a significant improvement in net hospital revenue.

This paper has several limitations. Firstly, although our dataset is high-resolution and obtained from a large hospital with a substantial number of patients annually, it is derived from a single emergency department. While this dataset allows us to demonstrate the application of studying

---

time-varying workload, it is important to acknowledge the need for further research using data from multiple medical centers and departments to enhance the confidence in our findings. Secondly, we acknowledge that healthcare, similar to many other industries, may have less flexibility in adjusting service time per patient compared to some other sectors. Queuing theory suggests that as customers queue for service, workers have an incentive to reduce service time per customer in exchange for overall utility. However, in healthcare, such adjustments may be constrained. Despite this limitation, we believe that considering the time-varying aspects could still enhance estimation results in other service industries.

## References

- Anderson, D., Price, C., Golden, B., Jank, W., and Wasil, E. (2011). Examining the discharge practices of surgeons at a large medical center. *Health care management science*, 14(4):338–347.
- Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, 5(1):146–194.
- Batt, R. J., Kc, D., Staats, B., and Patterson, B. (2017). The effect of discrete work shifts on a nonterminating service system.
- Batt, R. J., Kc, D. S., Staats, B. R., and Patterson, B. W. (2019). The effects of discrete work shifts on a nonterminating service system. *Production and operations management*, 28(6):1528–1544.
- Batt, R. J. and Terwiesch, C. (2014). Doctors under load: An empirical study of state-dependent service times in emergency care.
- Batt, R. J. and Terwiesch, C. (2017). Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science*, 63(11):3531–3551.
- Becker’s Hospital Review (2016). Hospitals: Is your ED’s length of stay costing you millions?
- Berry Jaeker, J. A. and Tucker, A. L. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 63(4):1042–1062.
- Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.
- Dux, P. E., Ivanoff, J., Asplund, C. L., and Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fmri. *Neuron*, 52(6):1109–1120.
- Hodgson, N. R. and Traub, S. J. (2020). Patient assignment models in the emergency department. *Emergency Medicine Clinics*, 38(3):607–615.
- Kc, D. S. (2014). Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183.
- Kc, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management science*, 55(9):1486–1498.
- Kc, D. S. and Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65.
- Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., and Koch, I. (2010). Control and interference in task switching—a review. *Psychological bulletin*, 136(5):849.

- 
- Kuntz, L., Mennicken, R., and Scholtes, S. (2011). *Stress on the ward—An empirical study of the nonlinear relationship between organizational workload and service quality*. Number 277. Ruhr Economic Papers.
- Kuntz, L., Mennicken, R., and Scholtes, S. (2015). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science*, 61(4):754–771.
- Lee, D. K. K., Chen, N., and Ishwaran, H. (2021). Boosted nonparametric hazards with time-dependent covariates. *Annals of Statistics*, 49(4):2101.
- Lucas, R., Farley, H., Twanmoh, J., Urumov, A., Olsen, N., Evans, B., and Kabiri, H. (2009). Emergency department patient flow: The influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine*, 16(7):597–602.
- McCarthy, M. L., Zeger, S. L., Ding, R., Levin, S. R., Desmond, J. S., Lee, J., and Aronsky, D. (2009). Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine*, 54(4):492–503.
- Monsell, S. (2003). Task switching. *Trends in cognitive sciences*, 7(3):134–140.
- Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, 116(2):220.
- Pines, J. M., Batt, R. J., Hilton, J. A., and Terwiesch, C. (2011). The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of emergency medicine*, 58(4):331–340.
- Powell, S. G. and Schultz, K. L. (2004). Throughput in serial lines with state-dependent behavior. *Management Science*, 50(8):1095–1105.
- Rogers, R. D. and Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of experimental psychology: General*, 124(2):207.
- Schultz, K. L., Juran, D. C., Boudreau, J. W., McClain, J. O., and Thomas, L. J. (1998). Modeling and worker motivation in jit production systems. *Management Science*, 44(12-part-1):1595–1607.
- Sigman, M. and Dehaene, S. (2006). Dynamics of the central bottleneck: Dual-task and task uncertainty. *PLoS biology*, 4(7):e220.
- Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience*, 28(30):7585–7598.
- Stelzel, C., Schumacher, E. H., Schubert, T., and D’Esposito, M. (2006). The neural effect of stimulus-response modality compatibility on dual-task performance: an fmri study. *Psychological research*, 70:514–525.

- 
- Tan, T. F. and Netessine, S. (2014). When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593.
- Traub, S. J., Stewart, C. F., Didehban, R., Bartley, A. C., Saghafian, S., Smith, V. D., Silvers, S. M., LeCheminant, R., and Lipinski, C. A. (2016). Emergency department rotational patient assignment. *Annals of Emergency Medicine*, 67(2):206–215.
- U.S. Bureau of Labor Statistics (2022). Occupational outlook handbook, 2022. <https://www.bls.gov/ooh/healthcare/physicians-and-surgeons.htm#:~:text=The%20median%20annual%20wage%20for,was%20%24229%2C300%20in%20May%202022.>, accessed Sep 20, 2023.
- Worringer, B., Langner, R., Koch, I., Eickhoff, S. B., Eickhoff, C. R., and Binkofski, F. C. (2019). Common and distinct neural correlates of dual-tasking and task-switching: a meta-analytic review and a neuro-cognitive processing model of human multitasking. *Brain Structure and Function*, 224:1845–1869.
- Xu, Y., Tan, T. F., and Netessine, S. (2021). The impact of workload on operational risk: Evidence from a commercial bank. *Management Science*.

# Are Buyers Strategic in Online B2B Reviews?

Hao Ding

Goizueta Business School, Emory University, hao.ding@emory.edu

Mor Armony

Stern School of Business, New York University, marmony@stern.nyu.edu

Achal Bassamboo

Kellogg School of Management, Northwestern University, a-bassamboo@kellogg.northwestern.edu

Ruomeng Cui

Goizueta Business School, Emory University, ruomeng.cui@emory.edu

In the digital age, buyers rely heavily on online review information. Our paper studies buyers' strategic behaviors when leaving reviews in a business-to-business (B2B) context. In particular, we explore whether they are less likely to write a review when a supplier's transaction volume increases, hoping thus to curb the supplier's business growth and future bargaining power. We collect the entire review and transaction histories of 4,605 suppliers from Alibaba.com, the largest B2B global sourcing platform. Our dataset includes these suppliers' 62,529 reviews and 455,593 transactions, all timestamped, from February to November in 2017 and 2018. We use a generalized difference-in-differences method that leverages the natural experiment arising from the trade war between the US and China, which caused a sudden change in US buyers' purchasing behaviors, leading to fluctuations in sellers' transaction volumes. These changes served as natural shocks to non-US buyers, because they saw an exogenous change in some sellers' performances. We find that each additional transaction reduces the likelihood for buyers to leave reviews for the seller by 0.9 percentage point. We also find that the review numerical ratings did not significantly change after the shock. The results are mostly in line with our expectations—buyers leave fewer reviews in order to control the seller's growth and leave more when the seller's performance drops. We provide managerial insights on reviews on B2B platforms. The findings suggest that buyers are motivated to manipulate, for their own advantage, information on suppliers that is disclosed to the public on the platform. Therefore, information transparency could be a double-edged sword: disclosing too much information about sellers, such as their transaction histories, could induce buyers' strategic behaviors when leaving reviews. Platform owners should be careful in designing the level of information transparency.

*Key words:* Online reviews, B2B marketplace, strategic buyers

## 1. Introduction

Online reviews have become ubiquitous in the digital age. Buyers rely on reviews to make informed purchasing decisions, sellers use reviews to establish trust and attract buyers, and platforms depend on review information to match supply and demand. Researchers have therefore extensively examined sellers' strategic behaviors when managing their reviews; for example, anonymously manipulating



---

reviews for their products (Mayzlin et al. 2014, Luca and Zervas 2016, Xu et al. 2021). Operations researchers have studied how platforms use reviews to improve matching efficiency (Moreno and Terwiesch 2014, Cui et al. 2020, Mejia and Parker 2021, Mejia et al. 2021). However, past studies have rarely investigated the behaviors of the buyers who create and provide these reviews. In particular, few studies have explored whether and how buyers leave reviews strategically, primarily because previous studies have focused on the business-to-consumer (B2C) market, in which end consumers' behaviors as reviewers are assumed to be less strategic.

In this paper, we focus on the business-to-business (B2B) market. The global B2B e-commerce market is valued at over five times that of the B2C market<sup>1</sup> and is projected to grow more quickly than the latter (McKinsey & Company 2019). B2B buying decisions differ from those in B2C markets in terms of customer type (retailer vs. end consumer), purchase volume (large vs. small), purchase frequency from the same seller (high vs. low), and decision nature (planned and rational vs. impulsive and emotional). Consequently, B2B buyers have different motivations and behaviors when managing online reviews, making it important to study whether and how they use reviews as strategic levers.

On open platforms, buyers across the world have generated millions of online reviews—estimated to be worth \$400 billion (Weise 2017). With the rise of online B2B platforms, it becomes easier for business buyers to access, search, and share reviews in real time. Anecdotal evidence has suggested that reviews are especially helpful to business buyers (LinkedIn and G2 Crowd 2020, Shea 2020). In fact, compared to consumers in the B2C market, buyers in the B2B market are more likely to seek, trust, and base purchases on online reviews (Minsky and Quesenberry 2015). Business buyers have long been using reviews and referrals to identify, evaluate, and choose suppliers and products (Cui et al. 2021). Specifically, referrals—which play a role similar to that of reviews—have been shown to initiate 84% of B2B purchasing processes and influence 90% of B2B transactions (Minsky and Quesenberry 2016).

Reviews are critical to enhance business for sellers. They are first-hand and impartial perspectives about the product provided by existing customers, which can complement the information provided by the sellers themselves (Chevalier and Mayzlin 2006). Thus, reviews can help sellers establish credibility, drive sales, and attract future buyers (Liu 2006, Duan et al. 2008, Babić Rosario et al. 2016). Reviews are also essential operations drivers for platforms. The core function of a platform is to identify good products and match them with what customers want. When a platform hosts millions of buyers and products, matching manually becomes impossible. Review information can help platforms identify good sellers and products and has been used as a key component in advanced algorithms to search, rank, and display products (McAfee and Brynjolfsson 2017). Consequently,

<sup>1</sup> Source: <https://www.statista.com/study/44442/statista-report-b2b-e-commerce/>, accessed May 2022.

---

reviews are regarded as critical resources for platforms and sellers. For example, platforms design incentives for customers to leave reviews. Because sellers may sometimes try to manipulate their reviews by nudging customers to leave only good ones or by creating fictitious good ones, platforms also assemble teams to detect fake reviews and remove biases in reviews.

Scholars and policymakers have extensively investigated strategic behaviors in online reviews, but most studies have focused on the sellers. It has been shown that manipulation of reviews by sellers is prevalent in industries; including entertainment (Dellarocas 2006), hotels (Mayzlin et al. 2014), and restaurants (Luca and Zervas 2016). This can introduce biases in reviews, make reviews less credible, increase costs for sellers, and reduce consumer welfare (Mayzlin 2006). Online platforms, such as Amazon, have taken actions to monitor and mitigate these biases by detecting and banning fabricated reviews. Policymakers, such as the Federal Trade Commission, also recognize the tremendous value of reviews resulting in regulations that protect customers' right to leave true but negative reviews and in rules that challenge fabricated reviews (Federal Trade Commission 2019).

Buyers' strategic behaviors when leaving reviews, on the other hand, have received less attention in the literature. It has been argued that the strongest motivation for buyers, especially B2C buyers, to leave reviews is altruism; in particular, customers leave reviews primarily to share their purchasing experiences, help other customers identify good products, and promote good sellers.<sup>2</sup> However, the growing value of reviews might motivate buyers to take advantages of them, especially business buyers trying to maximize every bit of their profits. For example, knowing the influence of reviews on other buyers' purchasing decisions, business buyers might try to use this influence to indirectly affect a seller's sales.

In this paper, we study whether and how buyers in the B2B market change the way they leave reviews in response to suppliers' recent performance. Unlike consumers in the B2C market, business buyers have more sophisticated relationships with sellers and make more calculated decisions. A buyer in a B2B transaction receives capacity allocation and product supplies from a seller; the two parties negotiate wholesale prices based on factors such as their relative bargaining power (Cui et al. 2021). In addition, the buyer compete with other buyers for capacity and attractive pricing from the seller. Consequently, the buyer needs to maintain a subtle balance with its suppliers to secure capacities, valuable resources, and favorable pricing (Kalwani and Narayandas 1995).

On the one hand, if the supplier becomes too powerful or too popular among other buyers, a given buyer would lose its bargaining power over the pricing with that seller, or might not receive enough supplies because of the limited capacity allocated to it. On the other hand, if the supplier does not have enough orders to sustain a healthy business, buyers might not be able to reliably source

<sup>2</sup> Source: <https://www.spectoos.com/5-real-motivations-behind-people-write-customer-reviews/>, accessed April 2022.

---

products from it. In other words, buyers prefer their suppliers to be reasonably healthy but not too powerful. In practice, anecdotal evidence has shown that in the B2B world, business buyers, even if they are happy and satisfied with a supplier, are often reluctant to spread the word about the supplier as they may not want their competitors to know about it (Utpal M. Dholakia 2015). Online reviews can serve as powerful tools for business buyers to maintain this subtle balance with suppliers. When a supplier has capacity to spare, buyers can help the supplier grow by leaving positive reviews. When the supplier grows too popular and is short of capacity, buyers can try to limit its growth—to preserve their own bargaining power and ensure sufficient capacity—by ceasing to leave reviews.

Our empirical research context is Alibaba.com, the largest online B2B trading platform. Alibaba.com hosts buyers from more than 190 countries and sellers primarily from mainland China. It provides a typical e-commerce experience to customers, in which suppliers' and products' names, photos, number of reviews, and review ratings, are clearly displayed to buyers. The baseline number of reviews on the platform is relatively low, having one more review could therefore make a major difference for a seller.

We collect a novel dataset from Alibaba.com. On each seller's page, Alibaba publicly displays all the reviews that the seller has ever received and all transactions that it has ever had on Alibaba.com. We collect the content and timestamps of these reviews and transactions for 4,605 suppliers from February to November in 2017 and 2018. The resulting dataset includes 62,529 reviews and 455,593 transactions. Each review includes, besides the review content, three numerical ratings: supplier services, product quality, and on-time shipment. For each transaction, we collect the transaction value and the country of origin of the buyer and of the seller.

Our identification exploits a natural experiment arising from the trade war between the US and China. On May 29, 2018, the US government released its first official announcement of the trade war, in which it proposed an additional 25% in tariffs on \$50 billion of goods from China, as well as possible future tariff increases on other products.<sup>3</sup> As 97.8% of suppliers on Alibaba.com are based in mainland China, this announcement had significant implications for US buyers, who could face increased prices if they source most of their products from Chinese suppliers. In response, those US buyers might strategically change their ordering behaviors on Alibaba.com. Some might order more at the current price to hedge against the risk of price hikes, while others might stop ordering from Alibaba altogether. As a result, sellers experienced shocks—sudden increases or decreases—in their transaction volumes. At the same time, this announcement had little direct impact on buyers from other countries. But while they were unlikely to change their ordering behaviors immediately,

<sup>3</sup> Source: <https://www.piie.com/blogs/trade-investment-policy-watch/trump-trade-war-china-date-guide>, accessed April 2022.

---

they could observe this shock in sellers’ transaction volumes and in turn alter their own reviewing behaviors.

We use a generalized difference-in-differences design that leverages this policy shock. In particular, our research design uses two features of the US government’s announcement. First, it led to an immediate change in US buyers’ purchasing behaviors and thus the sales volumes of sellers with large bases of US buyers. It served as an exogenous shock to non-US buyers who also purchased from these sellers, in the form of a sudden change in sellers’ transaction volumes. Second, the change in the transaction volumes varied across sellers because they had different percentages of US customers, and different US buyers reacted to the trade war differently. We leverage these variations in transaction volumes across sellers to estimate the causal impact of sellers’ transaction volumes on non-US buyers’ likelihood of leaving reviews (i.e., whether a buyer chooses to leave a review after a transaction).

Our results suggest that buyers are less likely to leave reviews when a seller has a higher number of transactions. Specifically, having an additional transaction in a month reduces the likelihood of buyers leaving reviews for the seller by 0.9 percentage point. Yet, we also find that the review ratings do not significantly change after the shock. This suggests that when a seller’s business is getting stronger, buyers are less likely to leave reviews but they will not leave negative reviews to badmouth the seller. These results are in line with our expectations—buyers leave fewer reviews to control the growth of a seller and leave more reviews when the seller’s performance drops.

We conduct several robustness checks to validate our results. First, we include the supplier-specific time trend to provide support for the parallel trends. Second, we run a placebo test to ensure that our results are not idiosyncratic. The results show no significant effect of using the same time window in 2017. Third, we show evidence that helps rule out potential alternative explanations of our findings. Fourth, we use the transaction value as an alternative dependent variable to ensure that our results are consistent when we measure sales differently.

Feedback and reputation systems are central to the operations of online marketplaces (Tadelis 2016). It is thus critical to design incentives to encourage feedback from customers (Chen et al. 2020). To the best of our knowledge, we are the first to provide empirical evidence that buyers may strategically choose to not leave reviews in some circumstances in order to benefit themselves. Platforms need to take this into consideration when designing their review systems. For example, reviews can be manipulated by buyers and thus should be integrated carefully into the search algorithms. Furthermore, our work points out that information transparency can be a double-edged sword for online platforms (Zhu 2004, Cui et al. 2020, Mejia and Parker 2021) in that B2B buyers can exploit suppliers’ transparent transaction histories for their own benefit. In our context, information transparency in the form of sellers’ transaction histories could potentially hurt both those sellers and the platforms when buyers strategically choose to not leave reviews for good products and sellers.

---

## 2. Literature Review

Our work contributes to the empirical literature on online reviews and platform management. We also relate these two multidisciplinary streams of literature to a central topic in operations management—capacity management.

### 2.1. Online Reviews

First, our work builds on and contributes to the multidisciplinary online review literature across the fields of operations management, marketing, and information systems. Mounting research has established that both the review volume (Liu 2006, Duan et al. 2008) and review ratings (Chevalier and Mayzlin 2006, Anderson and Magruder 2012) have significant impacts on products' and services' future sales. In a meta-analysis across 96 studies, 40 platforms, and 26 product categories, Babić Rosario et al. (2016) conclude that review volume is more strongly related to sales than the rating is, both in terms of magnitude and significance. Furthermore, the literature has studied various moderating factors of reviews' impacts, such as the variance of the review rating (Sun 2012), the reviewer's identity (Forman et al. 2008), and product characteristics (Zhu and Zhang 2010).

In the operations management literature, scholars primarily focus on whether and how firms leverage online reviews to improve their operations. Most of the literature focuses on examining reviews from the perspectives of platforms and sellers. For example, online reviews have been shown to attenuate racial discrimination on sharing economy platforms (Cui et al. 2020) and reduce the negative effects of language and cultural differences on IT sourcing platforms (Hong and Pavlou 2017). Reviews have been shown to help a platform optimize its search algorithm in the context of hotels (Ghose et al. 2014). The image contents of reviews have also been shown to improve user engagement (Khernamnui et al. 2021). Sellers also use reviews to optimize their businesses. For example, reviews can help sellers optimize their pricing strategies (Moreno and Terwiesch 2014). Service providers can extract information about service quality from reviews using textual analysis in the restaurant industry (Mejia et al. 2021). Likewise, healthcare providers can derive useful operational measures from reviews, including the wait time and the friendliness of receptionists, to better understand patient choices in physicians and improve their service quality (Xu et al. 2021). User ratings have also been shown to be more effective than government ratings in influencing nursing home demands (Li et al. 2021). In contrast, there is limited evidence of whether and the extent to which buyers leverage reviews strategically. We extend the boundary of the existing literature by addressing this side of the market.

In the B2C context, buyers' strategic behaviors and their impacts on operations have been well documented. For example, Papanastasiou and Savva (2017) build a theoretical model to demonstrate that consumers may choose to strategically delay their purchasing decisions in anticipation of product

---

reviews of their peers. In addition to online reviews, B2C buyers have shown strategic behaviors in various contexts. A strategic consumer is rational and forward-looking—they compare their expected utility of making an immediate purchase against the expected utility of purchasing in the future. The literature has shown that consumers’ strategic behaviors have structural impacts on a seller’s optimal decisions, such as pricing (Aviv and Pazgal 2008), capacity choices (Liu and Van Ryzin 2008), the quick response strategy (Cachon and Swinney 2009), and discount strategy (Cachon and Feldman 2015). Our work adds to the literature by providing empirical evidence of how business buyers leverage reviews in their operations in B2B marketplaces.

## 2.2. Platform Management

Our work is also closely related to the literature on platforms, which studies the platform mechanism designs. A common theme here is to improve the matching efficiency of two sides of the market (e.g., buyers and sellers). Examples of the mechanisms that have been studied include different levels of involvement by marketplaces in facilitating communication between buyers and service providers (Allon et al. 2012), surge pricing that can improve the matching between riders and drivers as demand fluctuates (Cachon et al. 2017, Besbes et al. 2021), how to set monetary incentives to the service providers (Sinchaisri et al. 2019, Chen et al. 2022). In addition, in order to increase service capacity of drivers, ride-sharing platforms should also consider driver behavioral factors such as income-targeting behavior when optimizing incentives (Allon et al. 2018). Recent papers have also studied the negative effects of higher market thickness (Li and Netessine 2020) and potential ways to reduce matching frictions (Arnosti et al. 2021).

Our work shows the implications of revealing information about sellers’ transaction volumes. In that sense, our work is closely related to the literature on information design and operations transparency in online platforms. For example, information disclosure about product quality helps better match customers’ needs (Tadelis and Zettelmeyer 2015). Previous literature has shown that disclosing sellers’ inventory availability information can attract more sales (Allon and Bassamboo 2011, Gallino and Moreno 2014, Cui et al. 2019, Calvo et al. 2020). Disclosing sellers’ capacity information is valuable to improve platform efficiency by encouraging buyers to find sellers with higher at-the-moment capacities (Horton 2019) or reducing cannibalization among substitutable auctions during days of high supply (Bimpikis et al. 2020).

Operational transparency has been shown to be valuable. For example, Buell and Norton (2011) and Buell et al. (2017) show that transparency of service processes helps improve customer satisfaction and appreciation of the service provided. Bray (2020) demonstrates that customers have a nuanced reactions to transparency in parcel delivery; they are happier when track-package activities occur near the final delivery time. Cui et al. (2020) show that increased information transparency

---

provided through online reputation systems can help reduce racial discrimination in online service marketplaces. In some circumstances, however, transparency can backfire. It can induce unintended opportunistic behaviors. For example, revealing riders’ racial information (Ge et al. 2020) or revealing riders’ support for LGBTQ+ (Mejia and Parker 2021) can induce discrimination behaviors among passengers in ride-sharing platforms. Our paper finds that revealing suppliers’ recent transaction volumes enables buyers to observe sellers’ business performances, thus inducing buyers’ strategic behaviors when leaving reviews. They leave fewer reviews for top-performing suppliers. This may create biases in the volume of reviews, which can lead to inefficient matching. Our results highlight that platforms should carefully design the level of information transparency: sharing the performance of market participants might induce strategic behaviors and reduce market efficiency.

Our work also relates to the literature that studies B2B marketplaces. This literature has very few empirical studies primarily because of the difficulty of accessing confidential data on wholesale trading (Phillips et al. 2015). With the B2B market transitioning from private offline transactions to open online platforms, there is a greater need to understand how the B2B market works and how its operations can be improved. Recent studies in this space take advantages of data made available on such B2B trading platforms. For example, Cui et al. (2021) investigate how information strategies influence wholesale pricing and price discrimination using data from Alibaba.com. We also collect publicly available data from Alibaba.com to empirically examine the strategic role of online reviews to platforms. By analyzing this novel dataset on the largest online trading platform, our work complements this growing literature by offering insights into B2B marketplaces.

### **2.3. Capacity Management and Rationing Behaviors**

Our work also relates to a core operations management literature—capacity management (Song et al. 2020). Specifically, how should a supplier allocate its limited capacity among buyers, and how would buyers strategically respond to the limited capacity? In our research context, fighting for the suppliers’ capacity is one plausible driver of buyers’ choices in leaving reviews. Leaving more good reviews can help the seller grow its business, which might consequently tighten its capacity to some extent. Leaving fewer reviews would attract less buyers’ attention to this seller, which enables the buyer to secure its share of the supplier’s finite capacity.

This literature has extensively studied various strategic ways that buyers use to win suppliers’ capacity, especially when they anticipate a potential supply shortage. This strategic motivation to compete for a supplier’s capacity has been theoretically validated across popular allocation mechanisms, such as proportional to order quantity (Lee et al. 1997), turn-and-earn (Cachon and Lariviere 1999, Lu and Lariviere 2012), and uniform allocation (Cho and Tang 2014). Lee et al. (1997) theorize that buyers may game the system when the suppliers’ inventory is rationed. Buyers must compete for

---

the supply. They often do so by sharing inflated demand forecast information. In times of scarcity, they will request excess inventory, hoping to end up with the desired amount of product. Furthermore, buyers are even more incentivized to game the allocation when they aim to gain a larger market share for downstream consumers.

When deciding on whether to leave reviews, buyers are essentially weighing the benefits of promoting versus hiding their suppliers. Leaving reviews promotes the supplier, and the resulting competition for the supplier’s scarce capacity could cause the buyer to obtain fewer units or pay a higher price. In this sense, our paper is particularly related to Kalkanci and Plambeck (2020). They use a modeling approach to demonstrate the trade-offs between revealing buyers’ supplier lists and not. Revealing the list would impose a risk in suppliers’ scarce capacity. A key takeaway here is that controlling the visibility of suppliers could be a leading consideration for buyers, and competition for suppliers’ capacities is the primary reason in doing so. Our paper complements their work by showing empirical evidence of such a phenomenon.

The literature has limited empirical evidence of buyers’ ration gaming behaviors. The few empirical works here study offline B2B transactions. For example, Terwiesch et al. (2005) show that buyers in the semiconductor sector inflate their orders more when their sellers are historically unreliable in fulfilling orders. Bray et al. (2019) show evidence of the ration gaming behaviors of retailers—all retailers simultaneously hoard supply in anticipation of shortages in their distributors’ inventories. We add to the literature by providing empirical evidence of how buyers use online reviews to achieve the goal of preserving suppliers’ capacities. This finding supports the fundamental assumption of buyers’ rationing reactions to suppliers’ transaction volumes and capacities. Moreover, we extend the literature by showing that such a fundamental assumption remains consistent as B2B transactions move online.

### **3. Research Setting and Hypothesis Development**

In this section, we describe our research setting and theorize buyers’ motivations to leverage reviews strategically in the B2B market.

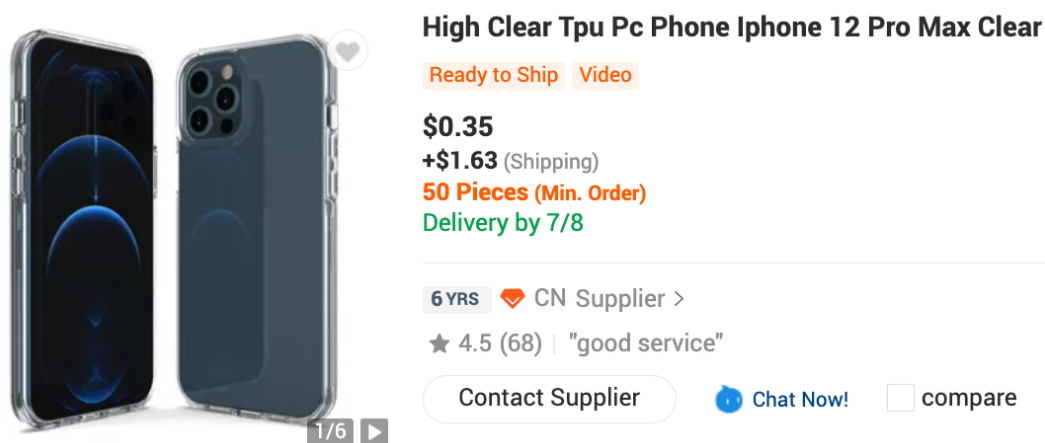
#### **3.1. Research Setting**

We study buyers’ strategic reactions in response to suppliers’ transaction volumes on an online global trading platform, Alibaba.com, where “buyers, who are located in more than 190 countries, are typically trade agents, wholesalers, retailers, manufacturers and SMEs engaged in the import and export business” (Alibaba 2018). Founded in 1999, Alibaba is the largest online B2B trading platform in the world that facilitates efficient and reliable trade between millions of buyers and suppliers. It provides us with a unique research context because of the openly available information, especially of sellers, displayed on the platform.



Alibaba.com has almost the same setting as B2C platforms, such as Amazon. Before buying a product, a buyer could first research for products or suppliers using some keywords. Based on the buyer's search terms, Alibaba.com displays the ranked results of matching products or suppliers. The review information is prominently presented on the search results page. Similar to the e-commerce experience, the buyer can immediately see the cumulative number of reviews and review ratings on the results page, in addition to other key information, including each product's name, photos, price, minimum order quantity, shipping lead time, and seller's years of experience. Figure 1 presents an example of the search results of a product.

**Figure 1** Search Result of a Product

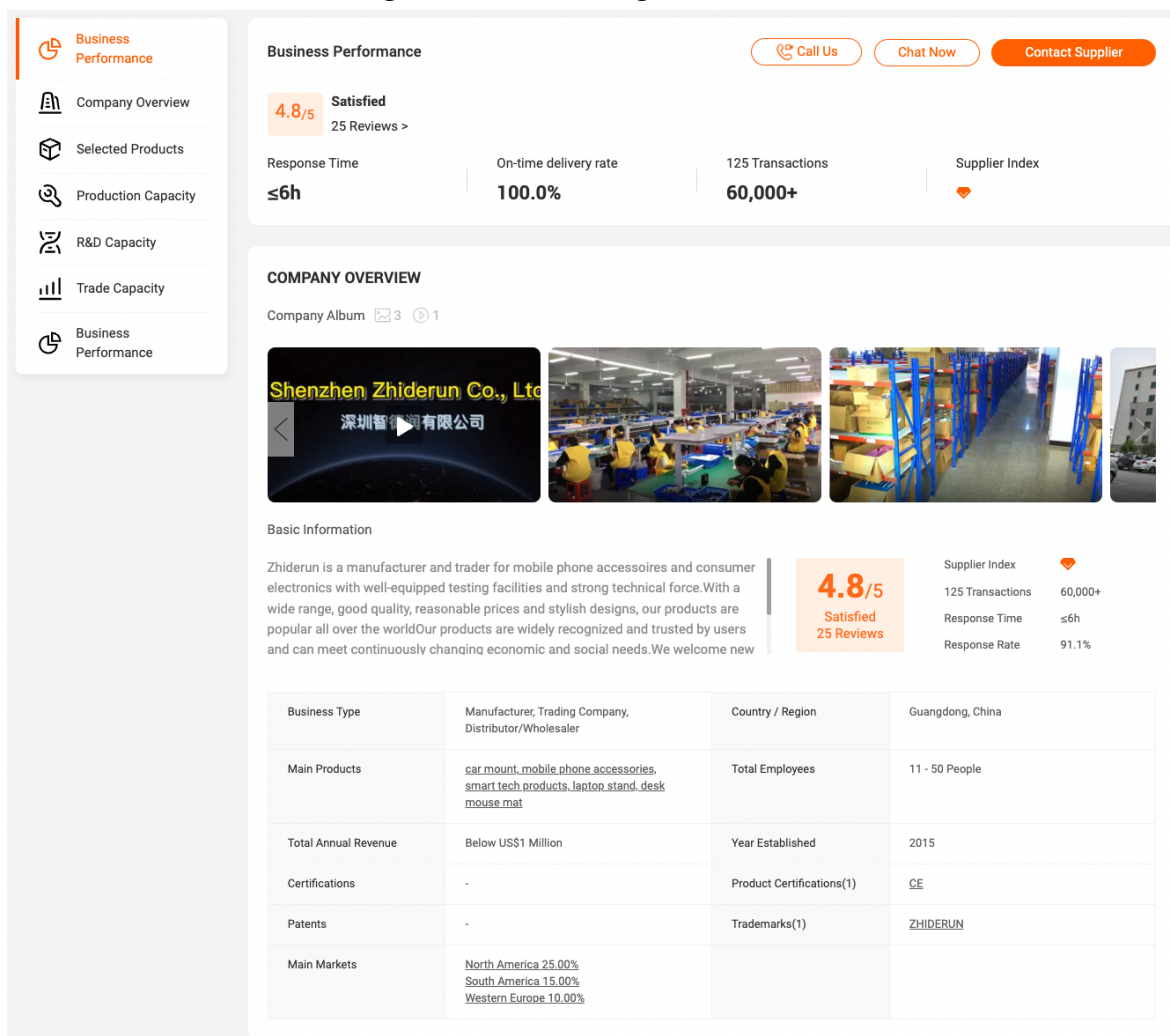


The B2B platform is designed in a supplier-oriented format, in which each seller has its own page on Alibaba.com. Each product page also provides a link that takes the buyer to the seller's page. The buyer can then visit any seller's page to view more detailed characteristics, which include the seller's past transaction volume, number of reviews, review ratings, response time, business type, and main markets. Figure 2 presents an example of the seller's homepage.

Buyers then rely on the information on sellers' pages to select suppliers and make purchasing decisions. Alibaba.com explicitly displays all historical transactions that a seller has had on the platform. It also plots the trend of a seller's transaction volume in the last 12 months. Figure 3 presents how a seller's transaction history is shown on Alibaba.com.

After a successful transaction, a buyer can choose to leave the seller a review or not. If choosing to leave a review, the buyer can give a positive review to appreciate a good transaction by the supplier or a negative review to express their dissatisfaction. Similar to B2C platforms, this is a prominent—if not the only—channel for buyers to leave public feedback for suppliers. Alibaba.com shows all historical reviews of each seller on the seller's homepage. For each review, the platform displays information, such as the name and the picture of the product being reviewed, the review's

Figure 2 A Seller's Page on Alibaba.com



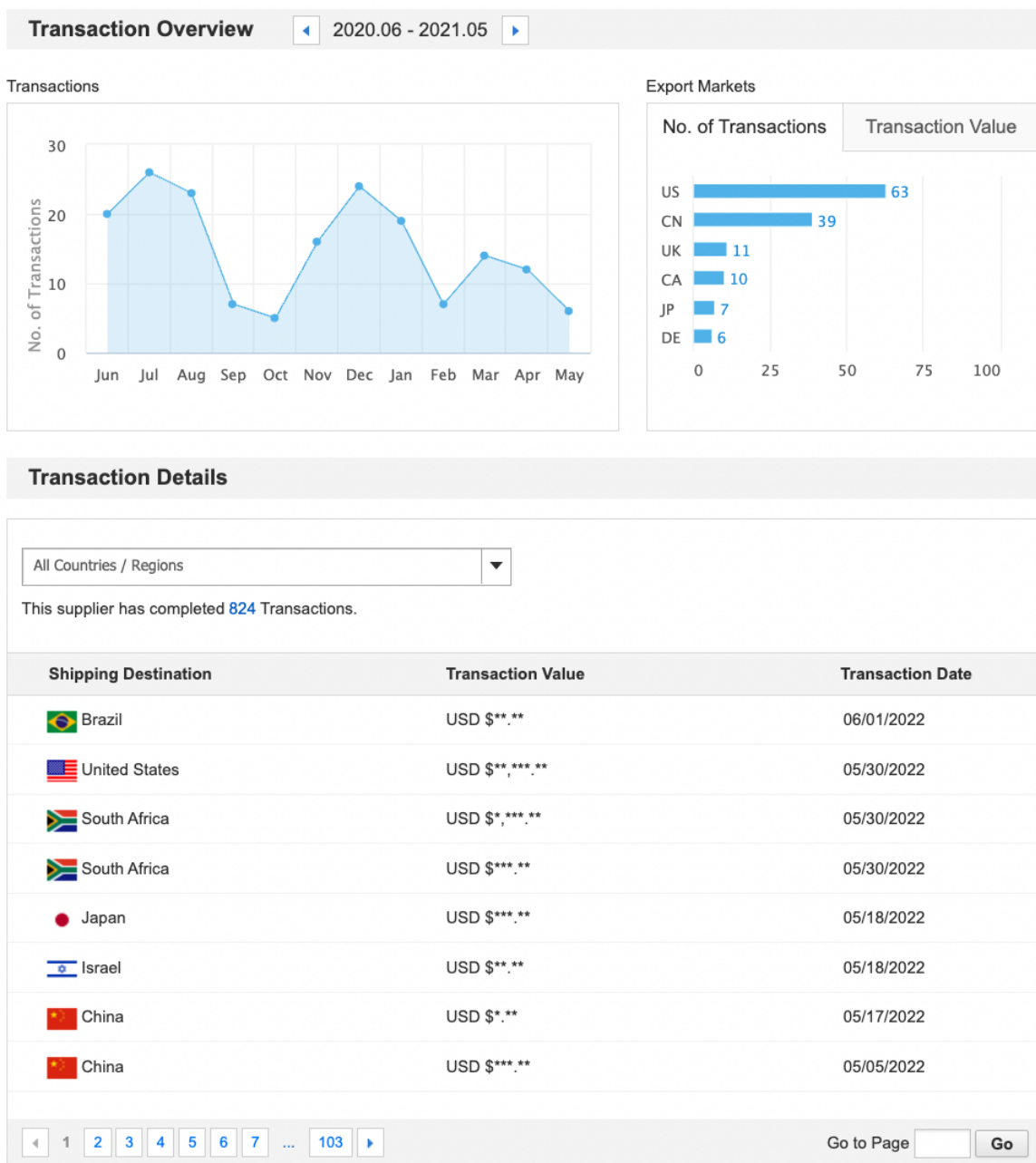
timestamp and rating, the first and last characters of the buyer's username, and the country of the buyer. Note that this provides enough information for sellers to identify which buyers left which reviews. Figure 4 presents how a seller's review history is shown on the platform.

### 3.2. Hypothesis Development

In the digital age, buyers heavily rely on online review information when making purchasing decisions (Chevalier and Mayzlin 2006). Consequently, sellers use reviews to establish credibility, drive sales, and attract buyers. In the B2C market, mounting evidence in the literature has shown that both the review volume and review ratings are strongly related to sellers' sales (Duan et al. 2008, Zhu and Zhang 2010, Luca 2016).

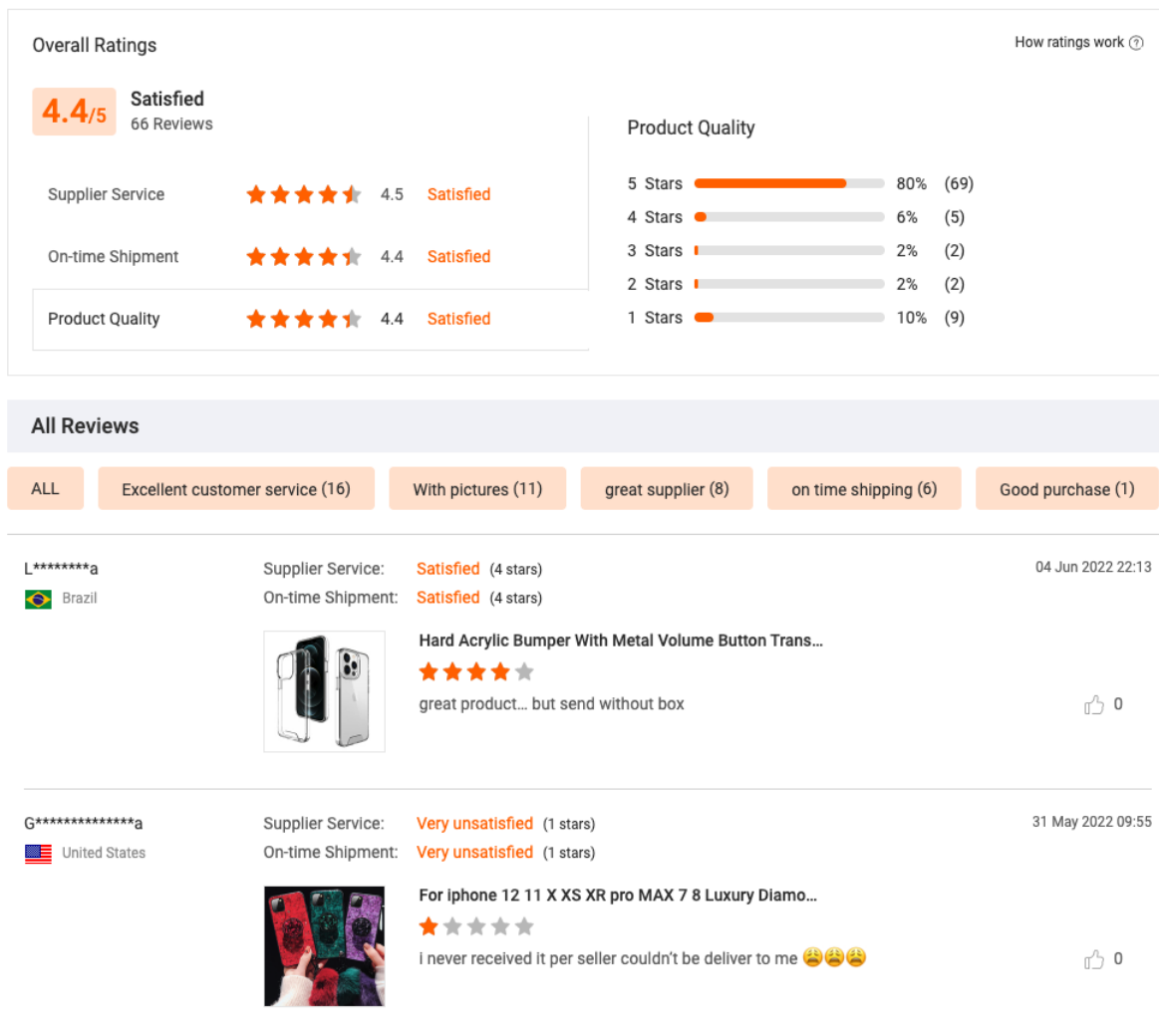
In the B2B market, the evidence of whether and how reviews affect sales is less conclusive. Anecdotal evidence suggests that business buyers also heavily rely on reviews to identify, evaluate, and choose good products and qualified suppliers. For example, a recent survey found that more than

Figure 3 Transaction History of a Seller



90% of business buyers actively seek online reviews before sourcing from new suppliers (Minsky and Quesenberry 2015). In addition, B2B trading platforms often present and use review information in a similar way compared with B2C platforms. Notably, B2B platforms prominently show the cumulative number of reviews and review ratings on the search results page. Lastly, reviews are the only channel for buyers to leave public feedback for sellers. As reviews are the primary and probably the only impartial quality indicator on B2B platforms, business buyers are likely to rely on reviews to make their purchasing decisions. We hypothesize that the review volume and review ratings can

Figure 4 Review History of a Seller



drive sellers' future sales. Note that the effect of reviews in driving sales has been extensively and empirically validated in B2C markets (Chevalier and Mayzlin 2006, Zhu and Zhang 2010, Luca 2016). Therefore, we acknowledge that the value of reviews in the B2B market is expected, and we include Hypothesis 0 as a sanity check for our data and for the sake of the completeness of our analysis.

**HYPOTHESIS 0.** *The review volume and review ratings drive sellers' future sales on the B2B platform.*

In the B2B market, business buyers have sophisticated relationships with suppliers and other buyers. A good supplier is key to the success of a buyer's business. A reliable seller can ensure a good product quality, sufficient supply, smooth shipment, and a reasonably low production cost. Therefore, buyers need to exert significant efforts to identify capable suppliers and build a professional relationship with them. Before each transaction, buyers and sellers need to negotiate on operations

details, including the wholesale price (i.e., the unit price the seller charges the buyer), the supply capacity (i.e., how much capacity the seller allocates to the buyer), and operations support (e.g., how fast to pack and ship products to the buyer's destination, see Cui et al. 2021). They would consider factors such as the seller's available capacity, the buyer's order volume, and the level of competition in the market. Based on these factors, the seller and the buyer reach an agreement on the above terms.

The seller's capacity plays a major role in the negotiation outcome (Balachander and Farquhar 1994, Liu and Van Ryzin 2008, Debo and van Ryzin 2013). In time of scarcity, it is difficult for buyers to secure stable supplies. Another critical factor that dictates the negotiation outcome is the bargaining power (Leider and Lovejoy 2016). In the face of a powerful seller, a buyer encounters difficulties in negotiating a favorable deal. Taken together, to secure a reliable, consistent, and low-cost supply of products, the B2B buyers need to maintain balanced, complex, subtle, and long-term relationships with their suppliers (Kalwani and Narayandas 1995).

Why is it necessary for buyers to keep balanced relationships with suppliers? On the one hand, if a supplier becomes too powerful or too popular, a buyer would lose the bargaining power on negotiating a contract, or the buyer would not be able to secure supplies because of the limited capacity allocation. In other words, when many buyers purchase from the same supplier, each buyer may experience negative externalities in pricing and capacity allocation. On the other hand, if the supplier does not receive enough orders, it might not be able to sustain a healthy business, and, in turn, it might not be able to consistently and reliably supply products and, in the worse case, go out of business. As a result, buyers prefer their suppliers to be reasonably healthy yet not too powerful.

The online reviews that business buyers leave for their suppliers can serve as powerful tools to maintain this subtle balance. Online reviews can help a seller stay healthy and grow its business, which ensures a healthy supply. At the same time however, it may attract too much demand from competing buyers, who would overwhelm the supplier's capacity and reduce each buyer's bargaining power. Therefore, buyers might change the way they leave reviews to manipulate the supplier's well-being. When a supplier has ample capacity to spare, buyers can leave positive reviews to help the supplier grow its business. When a supplier grows too popular or is on the verge of capacity shortages, buyers can stop leaving reviews to limit the supplier's growth in order to preserve the buyer's bargaining power in pricing and capacity. Similar phenomena have been documented by anecdotal evidence; even when business buyers are happy and satisfied with the supplier, they are often reluctant to spread the words about the supplier's great products because they do not want their competitors to know about the supplier (Utpal M. Dholakia 2015).

A key assumption in the above story is that buyers can monitor suppliers' performances, which is often made possible by B2B platforms' information transparency. Recall that Alibaba.com explicitly

---

displays suppliers' historical transaction information. For each transaction, buyers can observe the transaction date, volume, value, and time trend. In practice, buyers often use suppliers' transaction data to predict their potential inventory shortages (Kesavan et al. 2010, Cui et al. 2015). Using such information, buyers can adjust their review behaviors accordingly in response to suppliers' performances.

The core driver for buyers to manipulate reviews is ration gaming, a central operations topic in supply chain. Lee et al. (1997) define ration gaming as the strategic behaviors of buyers when supply shortage is anticipated. When a supplier is short of inventory, buyers must compete for the supply. They can game the system by demanding inflated orders from suppliers in order to obtain the limited inventory. Terwiesch et al. (2005) and Bray et al. (2019) provide empirical evidence of ration gaming behaviors; buyers make inflated orders to suppliers in anticipation of inventory shortages. Therefore, it is highly plausible for buyers to use reviews in order to game the rationed system.

In short, B2B buyers try to maintain balanced relationships with their suppliers, in which they prefer these suppliers to be reasonably healthy yet not too powerful. This balance ensures a reliable source of supplies but avoids excessive competition in capacity or the lost of bargaining powers in pricing. In online marketplaces, business buyers can exploit reviews' power to achieve this balance. By leaving positive reviews, buyers help their suppliers grow. These positive reviews, however, attract competitors who could adversely affect the focal buyers. By not leaving reviews, buyers can keep the supplier somewhat hidden from their competitors, which enhances the capacity allocated to these buyers and their relative bargaining power. We hypothesize that buyers are less (more) likely to leave reviews for sellers when sellers have a higher (lower) number of transactions.

*HYPOTHESIS 1. Buyers are less (more) likely to leave reviews for sellers when sellers have a higher (lower) number of transactions.*

When a business buyer tries to control the seller's growth by using reviews, besides not leaving any review, another action the buyer can take is to leave a fabricated negative review. We next argue that this is not very likely to happen in the B2B context. A negative review signals a bad sourcing experience, which can discourage future buyers from ordering from the supplier. However, casting aspersions on the supplier will unfairly harm its business. Doing so will also damage the partnership, on which B2B buyers have invested significant time and efforts to build and maintain. In the worse case, the buyer might lose this supplier and have to exert extra efforts to find new sellers. Therefore, even though leaving a negative review may be more effective in controlling a supplier's growth and popularity, it is an unethical and unprofessional behavior that can backfire on buyers' own reputation and trustworthiness. As a result, we hypothesize that buyers do not change their review ratings for sellers when sellers have a higher number of transactions. A natural sequel of this hypothesis is that

it is unlikely for a buyer to leave an extremely negative review when the buyer wants to keep doing business with the supplier. In other words, after a buyer leaves a negative review, it is unlikely to observe future transactions between the two parties. Thus, we have the following two hypotheses:

*HYPOTHESIS 2. Buyers do not change their review ratings for sellers when sellers have a higher number of transactions.*

*HYPOTHESIS 3. Compared to buyers who leave positive reviews, buyers who leave negative reviews are less likely to have future transactions with the supplier.*

## 4. Data

We collect the review and transaction data of suppliers on Alibaba.com. The setting of Alibaba has been demonstrated in Section 3.1. Recall that each supplier’s transaction and review histories are publicly displayed. We first identify all the 4,605 suppliers in three popular categories: USB drive, towel, and copy paper categories. We then gather data on these 4,605 suppliers from February 2017 to November 2017 and from February 2018 to November 2018. Note that we purposely exclude the holiday season in January, which is the Chinese New Year, and December, which is Christmas time. The data include a total of 455,593 transactions and 62,529 reviews.

We obtain detailed information for each transaction and review. For each transaction, we collect its approximate transaction value, transaction date, and shipping destination, as shown in Figure 3. Note that the platform shows the *scale* of the transaction value instead of the exact dollar amount. For example, \$ \* \* \* . \* \* means that the transaction value is between \$100.00 and \$999.99. For each review, we collect its review ratings, review date, the first and last characters of the buyer’s username, and the country of the buyer, as shown in Figure 4. Each review provides ratings along three dimensions: supplier service, product quality, and on-time shipment. These ratings range from 1 (unsatisfied) to 5 (satisfied). We also calculate the overall rating for each review by taking the average of the three dimensions. Note that after a buyer places an order, the seller first prepares the order and then ships it to the buyer. After the buyer receives the shipment, the buyer completes the payment to the seller, and the transaction is closed. The completed transaction will then be displayed on the seller’s transaction history page, as shown in Figure 3. It is safe to assume that the buyer will leave a review within a few days after the transaction. Therefore, it is likely that the transaction and the corresponding review occur within the same month.

With the information that we collect from Alibaba.com, we cannot match each review with the transaction that it belongs to, which prevents us from performing the analysis at the transaction level. Because a review often closely follows a transaction, we can create and analyze a panel data. We aggregate the number of transactions, the number of reviews, and the review ratings of each supplier at the monthly level. We approximate the probability of a seller receiving a review from a

transaction in a month as the the total number of reviews in that month divided by the total number of transactions in that month:

$$Review\ Probability_{it} = \frac{Number\ of\ Reviews_{it}}{Number\ of\ Transactions_{it}}, \quad (1)$$

where  $i$  denotes the seller, and  $t$  denotes the month.

Table 1 reports the split-sample summary statistics of the number of reviews, the number of transactions, the review ratings, and the review probability from February, 2018 to November, 2018 prior to and after the tariff announcement. This sample includes 4,579 suppliers, 60,826 reviews, and 297,044 transactions. In this time window, a supplier has an average of 4.45 transactions and 0.91 reviews per month.

Time Window	Variable	Mean	Std. dev	Max	Min	N
Before the announcement (02/2018—05/2018)	Number of Reviews	0.62	2.52	189	0	16,617
	Number of Transactions	3.41	6.40	183	0	91,259
	Overall Review Rating	4.72	0.64	5	1	16,617
	Supplier Service Rating	4.75	0.71	5	1	16,617
	Product Quality Rating	4.67	0.76	5	1	16,617
	On-time Shipment Rating	4.73	0.68	5	1	16,617
	Review Probability	0.17	0.59	1	0	91,259
After the announcement (06/2018—11/2018)	Number of Reviews	1.10	4.16	468	0	44,209
	Number of Transactions	5.13	9.25	497	0	205,785
	Overall Review Rating	4.78	0.56	5	1	44,209
	Supplier Service Rating	4.82	0.59	5	1	44,209
	Product Quality Rating	4.74	0.64	5	1	44,209
	On-time Shipment Rating	4.78	0.61	5	1	44,209
	Review Probability	0.20	0.69	1	0	205,785

The table summarizes the number of reviews, the number of transactions, the review ratings, and the review probability. The sample includes 4,579 suppliers and a total of 60,826 reviews.

## 5. Identification

In this section, we first introduce the identification to test Hypothesis 0, in which we estimate the extent to which sellers’ reviews affect their future sales using a fixed effect model. We then introduce the identification to test Hypothesis 1 and 2. We use the trade war announcement in June 2018 as a natural shock and adopt a generalized difference-in-differences regression. Lastly, we discuss the logistic model employed to test Hypothesis 3.

### 5.1. Fixed Effect Model

To identify the effects of sellers’ number of reviews and review ratings on driving future sales in Hypothesis 1, we follow the classic approach in the marketing literature (Liu 2006, Duan et al. 2008). We use a fixed effect model to regress the cumulative review ratings and the cumulative number of



reviews in the last month on the number of transactions. One challenge in studying the relationship between reviews and sales is to identify the direction of the effect (i.e., do more and better reviews lead to more sales or do better products that have more sales lead to more and better reviews). The following model partially solve potential concerns related to reverse causality by using lagged review variables and controlling for historical sales volume directly:

$$\begin{aligned} \text{Number of Transactions}_{it} = & c + \lambda_i + \delta_t + \alpha_1 \text{Cumulative Reviews}_{i,t-1} + \alpha_2 \text{Cumulative Ratings}_{i,t-1} \\ & + \text{Number of Transactions}_{i,t-1} + \epsilon_{it}, \end{aligned} \quad (2)$$

where  $i$  denotes the seller,  $t$  denotes the month,  $\lambda_i$  is the supplier fixed effect,  $\delta_t$  is the time fixed effect, and  $\epsilon_{it}$  is the error term.  $\text{Cumulative Reviews}_{i,t-1}$  captures the cumulative number of reviews that supplier  $i$  had up to month  $t - 1$ ,  $\text{Cumulative Ratings}_{i,t-1}$  captures the cumulative review ratings of the seller up to month  $t - 1$ , and  $\text{Number of Transactions}_{t-1}$  is the number of transactions the seller had in month  $t - 1$ . The coefficients of interest are  $\alpha_1$  and  $\alpha_2$ , where  $\alpha_1$  captures the effect of the sellers' past review volume on future sales, and  $\alpha_2$  captures the effect of the sellers' past review ratings on future sales.

## 5.2. Difference-in-Differences Identification

To identify buyers' strategic behaviors when leaving reviews in Hypothesis 1 and 2, we use a natural shock arising from the trade war between the US and China to resolve challenges in identifying the effect of sales on reviews, such as reverse causality. On May 29, 2018, the US government confirmed the plan to impose 25% additional tariffs on \$50 billion of goods from China and a potential tariff increase on more products in the future. The official announcement was then released by the US government on June 15, 2018. Because over 97.8% of sellers on Alibaba.com are based in mainland China, this announcement had significant implications for US buyers. US buyers could face increased prices when they source from Chinese suppliers. In response, US buyers might strategically change their ordering behaviors on Alibaba.com. Some buyers may increase their order quantities at the current price to hedge against the risk of price hikes, while others may stop ordering from our studied platform and turn to suppliers elsewhere. As a result, sellers experienced shocks—sudden increases or decreases—in their transaction volumes to some extent. However, this announcement had little direct impact on buyers in non-US countries. Although they were unlikely to change their ordering behaviors immediately, they could observe this shock in sellers' transaction volumes and in turn alter their review leaving behaviors.

We use a generalized difference-in-differences design to exploit this trade war. Our research design takes advantage of two features of the tariff announcement. First, this announcement led to an immediate change in US buyers' purchasing behaviors and thus a change in the sales volumes of the

sellers who had large bases of US buyers. It served as an exogenous shock to non-US buyers who also purchased from these sellers. For them, they saw a sudden change in sellers’ transaction volume, which was not driven by any inherent characteristics. Second, the change in transaction volumes varied across sellers because they had different percentages of US buyers in their customer bases, and different US buyers reacted to the trading war differently. Therefore, our treatment is the continuous variations in the transaction volumes across sellers—sellers’ number of transactions. The shock is the first official tariff announcement on June 15, 2018. The subjects in our identification are non-US buyers. Figure 5 intuitively illustrates the above key elements in our identification design.

**Figure 5 Illustration of the DID Identification**



We estimate the causal impact of sellers’ transaction volumes on non-US buyers’ likelihood of leaving reviews before and after the shock. In particular, we follow the classic generalized DID specification in Angrist and Pischke (2008, p. 229) with the supplier-level panel data,

$$Outcome_{it} = \lambda_i + \delta_t + \beta Transactions_{it} \times Announcement_t + \psi_{it} + \epsilon_{it}, \quad (3)$$

where  $i$  denotes the seller,  $t$  denotes the month,  $\lambda_i$  is the supplier fixed effect,  $\delta_t$  is the time fixed effect, and  $\epsilon_{it}$  is the error term.  $Transaction_{it}$  is the number of transactions that seller  $i$  has in month  $t$ .  $Announcement_t$  is the dummy variable that indicates whether month  $t$  is after the tariff shock.  $\psi_{it}$  captures the supplier-specific time trends. Including this term allows suppliers to follow different trends in a revealing way, which also later helps us test the parallel trends assumption (Card 1992, Angrist and Pischke 2008). The coefficient of interest is  $\beta$  of the interaction term.

$Outcome_{it}$  takes two values:  $Review Probability_{it}$  and  $Review Ratings_{it}$ .

- $Review Probability_{it}$  is the probability of non-US buyers leaving reviews for supplier  $i$  in month  $t$  after a transaction. Recall that each review and transaction shows the country that the buyer belongs to, as presented in Figures 3 and 4. This enables us to identify which reviews and transactions are from non-US buyers. For each seller  $i$  in month  $t$ , we then compute the review probability as the number of reviews from non-US buyers over the number of transactions from non-US buyers using Equation (1).

- $Review Ratings_{it}$  is the average rating of the reviews left by non-US buyers for supplier  $i$  in month  $t$ .

### 5.3. Logistic Model

To identify whether buyers are less likely to make future transactions with the supplier after leaving negative reviews in Hypothesis 3, we employ a logistic regression at the review level. Recall that the review information is presented in Figure 4. For each review, we can identify the buyer who left it by matching the first and last characters of the buyer’s username, the number of characters in the buyer’s username, and the buyer’s country. Based on the timestamps, we can identify whether a review is the last review between a buyer and a supplier. Note that we use the last review between a buyer and a supplier as the proxy for their last transaction.<sup>4</sup> Based on the review ratings, we define a review as negative if the rating is lower than 3 (which is deemed as unsatisfied by Alibaba’s standard). We then keep the reviews from the buyers who left more than one review with the same supplier, and we pool such reviews together and test whether the review being negative is associated with a higher chance that this is the last time that they do business together. Specifically, we run a logistic regression,

$$\text{logit}(\text{Last Review}_r) = c + \gamma \text{Negative}_r + \epsilon_r, \quad (4)$$

where  $r$  denotes the review.  $\text{Last Review}_r$  equals 1 if the review is the last review created by a buyer for a supplier, and 0 otherwise.  $\text{Negative}_r$  equals 1 if one of the three ratings is lower than 3, and 0 otherwise.

## 6. Empirical Results

In this section, we report the estimation results of buyers’ strategic behaviors when leaving reviews.

### 6.1. Online Reviews Drive Sales

We first analyze whether online reviews drive sellers’ future sales. Table 2 reports the estimation results using Equation (2). In column (1), we only consider the effect of the total number of reviews without any additional controls; in column (2), we add the supplier and time fixed effects; in column (3), we add the sellers’ review ratings as a predictor variable; and in column (4), we include sellers’ transaction volumes in the last month to control for the baseline popularity of sellers. In the above regressions, we include transactions and reviews from all suppliers during our main study window from February, 2018 to November, 2018.

The results show that the review volume has a positive and statistically significant effect on sellers’ future sales, and the results are highly robust after we control for the supplier and time fixed effects, review ratings, and past transaction volume. The coefficient of  $\text{CumulativeReviews}_{i,t-1}$  in column (4)

<sup>4</sup> Because Alibaba does not provide buyers’ usernames for transactions, we cannot link the buyer who left a review with the buyer who made a certain transaction. This is why we use the last review as the proxy for the last transaction. The fact that a buyer can continue to make future transactions after leaving the last review does not affect our estimation in Equation (4).

**Table 2 Impact of Online Reviews on B2B Sales**

Variables	Dependent variable: Number of Transactions			
	(1)	(2)	(3)	(4)
$CumuReviews_{i,t-1}$	0.250*** (0.021)	0.064*** (0.006)	0.057*** (0.011)	0.045*** (0.012)
$CumuRatings_{i,t-1}$			0.024 (0.197)	-0.007 (0.221)
$Number\ of\ Transactions_{i,t-1}$				0.110 (0.072)
Supplier FE	No	Yes	Yes	Yes
Time FE	No	Yes	Yes	Yes
Observations	42,469	42,207	28,590	26,046

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (2). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

suggests that each additional review leads to 0.045 additional orders per month for a supplier. In addition, the coefficients of  $CumuRatings_{i,t-1}$  in columns (3) and (4) are insignificant, indicating that review ratings have no significant effects on sellers' future sales. While this result might seem to be counter-intuitive at first, it is consistent with the findings in the marketing literature that the effect of reviews mostly comes from the review volume (Babić Rosario et al. 2016). In short, the findings provide partial support for Hypothesis 0.

## 6.2. Buyers' Strategic Behaviors when Leaving Reviews

We next examine how sellers' transaction volumes change buyers' behaviors when leaving reviews.

**6.2.1. Effect of Transaction Volume on Review Probability.** We first explore how sellers' transaction volumes affect buyers' likelihood of leaving reviews after each transaction. Table 3 reports the estimation results with  $Review\ Probability_{it}$  as the outcome variable using the DID specification in Equation (3). Column (1) reports the treatment effect without adding the supplier-specific time trends, and column (2) reports the treatment effect using the full specification. In both analyses, we use 4 months before the trade war announcement as the pre-treatment period and 6 months after the announcement as the post-treatment period.

**Table 3 Impact of Sellers' Recent Transaction Volumes on Buyers' Review Probability**

Variable	Dependent variable: Review Probability	
	(1)	(2)
$Transaction_{it} \times Announcement_t$	-0.010*** (0.002)	-0.009*** (0.003)
Supplier and Time FE	Yes	Yes
Supplier-Specific Time Trends	No	Yes
Observations	15,568	15,568
$R^2$	0.330	0.559

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (3). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

We find that buyers are indeed less likely to leave reviews when suppliers have more recent transactions, which supports Hypothesis 1. The coefficient of the interaction term in column (2) suggests that an additional transaction per month decreases the probability of buyers leaving reviews by 0.9% percentage point. This translates into a 4.5% decrease in percentage given that the average probability of buyers leaving reviews is 20% in our sample. Furthermore, both specifications yield highly consistent results with and without the supplier-specific time trends, which validates the parallel trend assumption. We discuss this further in Section 7.1.1.

The results are consistent with our conjectures. When suppliers become more popular, buyers leave fewer reviews; when suppliers become less popular, buyers leave more reviews. That is, buyers use reviews as strategic levers to control the growth of suppliers.

**6.2.2. Effect of Transaction Volume on Review Ratings.** We next study how sellers’ recent transaction volumes affect the review ratings they receive from buyers. Table 4 replicates the DID analysis using Equation (3), with  $Review\ Ratings_{it}$  as the dependent variable. We test for the treatment effect on the overall review rating, supplier service rating, product quality rating, and on-time shipment rating in columns (1)–(4), respectively.

We find that the coefficients of the interaction term are statistically insignificant for all review ratings, which means that buyers do not vary their behaviors when leaving review ratings in response to sellers’ transaction volumes. This provides support for Hypothesis 2.

**Table 4** Impact of Sellers’ Recent Transaction Volumes on Buyers’ Review Ratings

Variable	Dependent variable: Review Ratings			
	Overall (1)	Supplier Service (2)	Product Quality (3)	On-Time Shipment (4)
$Transaction_{it} \times Announcement_t$	0.0007 (0.0006)	0.0009 (0.0007)	0.0005 (0.0007)	0.0007 (0.0007)
Supplier and Time FE	Yes	Yes	Yes	Yes
Supplier-Specific Time Trends	Yes	Yes	Yes	Yes
Observations	16,305	16,305	16,305	16,305
$R^2$	0.521	0.502	0.526	0.513

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (3). The outcome variables are the overall rating, supplier service rating, product quality rating, and on-time shipment rating in columns (1)–(4), respectively. Significance at  $*p < 0.1$ ;  $**p < 0.05$ ;  $***p < 0.01$ .

**6.2.3. Negative Reviews and Repeat Purchase.** Lastly, we investigate whether buyers are less likely to make future transactions with the supplier after leaving negative reviews using Equation (4). In this analysis, we focus on reviews during our main study window from February 2018 to November 2018.

The coefficient of  $Negative\ Review_r$  in Table 5 is negative and statistically significant, which shows that compared with positive reviews, negative reviews are significantly more likely to be the

last review between a buyer and a supplier. This finding indicates that buyers are less likely to have future transactions with a supplier after leaving negative reviews. An implication of this finding is that leaving negative reviews is a serious decision in the B2B context. As a result, it is not very likely that buyers would leave a negative review just for the sake of controlling a supplier’s growth. This provides support for Hypothesis 3.

**Table 5 Negative Reviews and Repeat Purchase**

Variable	Last Review
<i>Negative Review<sub>r</sub></i>	2.310*** (0.120)
Observations	40,560
$R^2$	0.003

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (4). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 7. Robustness Checks

In this section, we report the results of several robustness tests. Specifically, we check the parallel trends assumption, conduct falsification tests to ensure that our estimated effects are not idiosyncratic, test for alternative explanations, and test an alternative measurement for the transaction volume.

### 7.1. Parallel Trend

The key identification assumption for the DID estimation is the parallel trends assumption, namely, that before the trade war announcement, buyers’ review behaviors in response to different sellers’ transaction volumes would follow the same time trend as in the absence of the announcement shock. In this study, our data tracks more than 4,000 suppliers for 10 months. This data structure allows us to directly include the supplier-specific time trends  $\psi_{it}$  in the DID regression to check for the parallel trends assumption (Card 1992, Angrist and Pischke 2008, Dasgupta and Žaldokas 2019). Adding this variable allows different suppliers to follow different trends in the estimation. Therefore, had different suppliers followed different trends, which the DID specification without the time trends cannot capture, adding this variable would significantly change the estimation results. Our estimations with and without the supplier-specific time trends in Table 3 are highly consistent, which supports that buyers’ review behaviors in our data followed the parallel trends as in the absence of the shock.

## 7.2. Falsification Test

To show that our estimated effects are not artifacts of seasonality, we test whether we can capture the same treatment effect in 2017, the year without the trade war announcement. We repeat the DID analysis specified in Equation (3) using data in 2017 for the same time window. If our results simply capture seasonality, we would be able to find significant effects in 2017 as well. Table 6 reports the falsification test results. The placebo-treated average treatment effect is insignificant, implying that buyers’ review behaviors did not change significantly in the previous year.

**Table 6 Falsification Test**

Variables	Review Probability
$Transaction_{it} \times Announcement_t$	−0.002 (0.0021)
Supplier and Time FE	Yes
Supplier-Specific Time Trends	Yes
Observations	394
$R^2$	0.774

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (3). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

## 7.3. Alternative Explanations

We discuss two potential alternative interpretations of the relationship between sellers’ transaction volumes and the probability of buyers leaving reviews.

**7.3.1. Strategic Buyers or Strategic Sellers?** One might question whether the observed effects arise from sellers’ strategic actions. When a seller experiences a sales decline, the seller may exert extra efforts in soliciting more reviews, for example, by nudging some buyers to leave good reviews. If sellers choose to do so, they tend to nudge buyers toward creating extremely positive reviews (Luca and Zervas 2016). Therefore, if sellers’ strategic actions were the primary drivers of our results, we would be able to observe significant changes in review ratings in response to changes in the transaction volume. Recall that the review ratings did not change significantly in Table 4. This indicates that it is unlikely that our observed effects are driven by sellers’ strategic actions.

**7.3.2. Temporal trend.** One might question whether the observed effects arise from a temporal trend. A temporal trend refers to the phenomenon in which buyers may perceive that new reviews make little impact when a seller already has a high number of reviews. Because leaving reviews takes time and effort, buyers would be less willing to leave reviews in the face of a higher number of existing cumulative reviews. Note that this temporal trend is unlikely to contaminate our estimation; the variations in the transaction volumes are not correlated with the existing reviews because of

the exogenous nature of our studied natural experiment. Nevertheless, we directly control for this possibility by extending Equation (3) with the total number of reviews that sellers had received up to the previous month as a control variable,

$$\begin{aligned} \text{Review Probability}_{it} = & \lambda_i + \delta_t + \beta_1 \text{Transactions}_{it} \times \text{Announcement}_t + \psi_{it} \\ & + \text{Existing Reviews}_{i,t-1} + \epsilon_{it}, \end{aligned} \quad (5)$$

where  $\text{Review Volume}_{i,t-1}$  is the total cumulative number of reviews that the supplier  $i$  had received until month  $t - 1$ .

Table 7 presents the estimation results, which show that after the existing reviews are controlled for, the coefficient of the interaction term remains almost the same. These results suggest that while buyers are indeed less willing to leave reviews when a seller already has a decent number of reviews, it does not drive our observed treatment effect.

**Table 7 Temporal Trend Not Affecting the Results**

Variables	Dependent variable: Review Probability	
	(1)	(2)
$\text{Transaction}_{it} \times \text{Announcement}_t$	-0.009*** (0.003)	-0.011*** (0.004)
$\text{Existing Reviews}_{i,t-1}$		-0.027*** (0.006)
Supplier and Time FE	Yes	Yes
Supplier-Specific Time Trends	Yes	Yes
Observations	15,568	15,568
$R^2$	0.332	0.573

This table reports the estimated coefficients and robust standard errors (in parentheses) using Equation (5). Significance at \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .

In addition to the analysis above, two settings on Alibaba.com would render the impact of temporal trend on our estimations very limited. First, sellers on Alibaba.com have limited numbers of reviews. In our sample, a seller has an average of only 25 reviews, and half of the sellers have less than 10 reviews. Thus, an additional review would always have a significant weight in affecting sellers' credibility. Therefore, buyers can use reviews to influence sellers. Second, Alibaba.com always displays the most recent reviews on the top of sellers' review histories, as shown in Figure 4, which can be immediately seen by future buyers. Therefore, it is unlikely for buyers to feel discouraged to leave reviews because of existing reviews.

#### 7.4. Alternative Measurement of Transaction Level

To ensure robustness, we test an alternative measurement of the outcome variable using sellers' monthly transaction *value* instead of their transaction *volume*. To calculate sellers' monthly transaction value, we leverage the approximate transaction value information displayed in the sellers'



transaction histories, as shown in Figure 3. For example,  $\$***.**$  means that the transaction value is between \$100.00 and \$999.99, and we use the average value (\$550) in the range to proxy the transaction value. We convert the unit from dollar to a thousand dollars by dividing the transaction value by 1,000 to make our results easier to interpret. In the above example, the final transaction value would be 0.55 thousand dollars. We then use the average transaction value per supplier per month as the dependent variable and replicate the DID analysis in Equation (3). The results in Table 8 are again highly consistent with our main findings, suggesting that our analyses are robust when we use an alternative measurement for transactions.

**Table 8** Impact of Sellers’ Recent Transaction Value on Buyers’ Review Probability

Variable	Dependent variable: Review Probability	
	(1)	(2)
$Value_{it} \times Announcement_t$	-0.0016*** (0.003)	-0.0015*** (0.003)
Supplier and Time FEs	Yes	Yes
Supplier Time Trends	No	Yes
Observations	15,568	15,568
$R^2$	0.329	0.558

This table reports the estimated coefficients and robust standard errors (in parentheses) in Equation (3). Significance at  $*p < 0.1$ ;  $**p < 0.05$ ;  $***p < 0.01$ .

## 8. Conclusion

Our paper adds to the long-standing literature on online reviews, an interdisciplinary topic in the social sciences. The two main streams of literature have extensively studied the impact of reviews on sales and sellers’ strategic behaviors in manipulating reviews. However, the point of view has rarely been studied from buyers’ perspectives. We contribute to the literature by exploring buyers’ motivations and strategic behaviors when leaving reviews.

By acquiring a novel dataset on Alibaba.com, we are among the first to provide empirical evidence of buyers’ strategic behaviors when leaving reviews. We show that buyers recognize the influence of reviews and strategically change their review behaviors in response to their sellers’ recent performances. Specifically, we find that an additional transaction per month decreases the probability of buyers leaving reviews by 0.9% percentage point. Because buyers use reviews as strategic levers to control the growth of suppliers by leaving less reviews when their sellers become more popular and leaving more reviews when their sellers become less popular.

In addition, we find that buyers do not change their review ratings in response to sellers’ performances. Fabricating a negative review for suppliers is an unethical and unprofessional behavior that can backfire on buyers’ own reputation and damage the partnership. The results are robust when

we control for the supplier-specific time trends and the temporal trends. We further validate our stories by performing a falsification test, ruling out other explanations, and testing an alternative measurement of the transaction level.

For B2B marketplaces, our findings provide two important insights. First, buyers’ strategic behaviors when leaving reviews can introduce biases in online reviews. To effectively leverage reviews, platforms should be aware of this bias when designing search and ranking algorithms. Second, information transparency could be a double-edged sword for trading platforms. In particular, revealing suppliers’ transactions in our studied context enables buyers to observe sellers’ business performances. While such information signals suppliers’ credibility and can attract more buyers, it may also induce buyers’ opportunistic behaviors when leaving reviews, such as not promoting the sellers who are doing well. One way to alleviate this issue is to reduce the level of transparency in transaction information. For example, instead of revealing detailed transaction information, platforms could consider showing aggregate information (Cui and Shin 2018). Doing so could entice new buyers but at the same time hide specific information about sellers’ business performances.

For B2B sellers, our findings point out the need for them to *actively* encourage buyers to leave reviews. Sellers should be aware that buyers can be motivated to not leave reviews regardless of how their transactions go. In addition, the findings derived in our context can be extended to offline B2B transactions, in which buyers have the same reluctance to refer suppliers. Therefore, sellers should exert effort to encourage business buyers to leave more reviews.

A limitation of our study is that we do not have enough information to match a review with the transaction it belongs to. Thus, our analyses are conducted at a coarser level. Future research could collect more fine-grained data and perform the analysis at the transaction level.

## References

- Alibaba. 2018. Alibaba.com: Leading wholesale marketplace for global trade. <https://www.alibabagroup.com/en/about/businesses>, (accessed, Dec. 13, 2021).
- Allon, Gad, Achal Bassamboo. 2011. Buying from the babbling retailer? The impact of availability information on customer behavior. *Management Science* **57**(4) 713–726.
- Allon, Gad, Achal Bassamboo, Eren B Cil. 2012. Large-scale service marketplaces: The role of the moderating firm. *Management Science* **58**(10) 1854–1872.
- Allon, Gad, Maxime C Cohen, Wichinpong Park Sinchaisri. 2018. The impact of behavioral and economic drivers on gig economy workers. *Available at SSRN 3274628* .
- Anderson, Michael, Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* **122**(563) 957–989.
- Angrist, Joshua D, Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton University Press.

- 
- Arnosti, Nick, Ramesh Johari, Yash Kanoria. 2021. Managing congestion in matching markets. *Manufacturing & Service Operations Management* **23**(3) 620–636.
- Aviv, Yossi, Amit Pazgal. 2008. Optimal pricing of seasonal products in the presence of forward-looking consumers. *Manufacturing & Service Operations Management* **10**(3) 339–359.
- Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, Tammo HA Bijmolt. 2016. The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research* **53**(3) 297–318.
- Balachander, Subramanian, Peter H Farquhar. 1994. Gaining more by stocking less: A competitive analysis of product availability. *Marketing Science* **13**(1) 3–22.
- Besbes, Omar, Francisco Castro, Ilan Lobel. 2021. Surge pricing and its spatial supply response. *Management Science* **67**(3) 1350–1367.
- Bimpikis, Kostas, Wedad J Elmaghraby, Ken Moon, Wenchang Zhang. 2020. Managing market thickness in online business-to-business markets. *Management Science* **66**(12) 5783–5822.
- Bray, Robert L. 2020. Operational transparency: Showing when work gets done. *Manufacturing & Service Operations Management* **0**(0).
- Bray, Robert L, Yuliang Yao, Yongrui Duan, Jiazhen Huo. 2019. Ration gaming and the bullwhip effect. *Operations Research* **67**(2) 453–467.
- Buell, Ryan W, Tami Kim, Chia-Jung Tsay. 2017. Creating reciprocal value through operational transparency. *Management Science* **63**(6) 1673–1695.
- Buell, Ryan W, Michael I Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Science* **57**(9) 1564–1579.
- Cachon, Gerard P, Kaitlin M Daniels, Ruben Lobel. 2017. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* **19**(3) 368–384.
- Cachon, Gérard P, Pnina Feldman. 2015. Price commitments with strategic consumers: Why it can be optimal to discount more frequently... than optimal. *Manufacturing & Service Operations Management* **17**(3) 399–410.
- Cachon, Gérard P, Martin A Lariviere. 1999. Capacity allocation using past sales: When to turn-and-earn. *Management science* **45**(5) 685–703.
- Cachon, Gérard P, Robert Swinney. 2009. Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Science* **55**(3) 497–511.
- Calvo, Eduard, Ruomeng Cui, Laura Wagner. 2020. Disclosing product availability in online retail. *Manufacturing & Service Operations Management* **0**(0).
- Card, David. 1992. Using regional variation in wages to measure the effects of the federal minimum wage. *Ilr Review* **46**(1) 22–37.

- 
- Chen, Xirong, Zheng Li, Liu Ming, Weiming Zhu. 2022. The incentive game under target effects in ridesharing: A structural econometric analysis. *Manufacturing & Service Operations Management* **24**(2) 972–992.
- Chen, Ying-Ju, Tinglong Dai, C Gizem Korpeoglu, Ersin Körpeoğlu, Ozge Sahin, Christopher S Tang, Shihong Xiao. 2020. Om forum—innovative online platforms: Research opportunities. *Manufacturing & Service Operations Management* **22**(3) 430–445.
- Chevalier, Judith A, Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* **43**(3) 345–354.
- Cho, Soo-Haeng, Christopher S Tang. 2014. Capacity allocation under retail competition: Uniform and competitive allocations. *Operations Research* **62**(1) 72–80.
- Cui, Ruomeng, Gad Allon, Achal Bassamboo, Jan A Van Mieghem. 2015. Information sharing in supply chains: An empirical and theoretical valuation. *Management Science* **61**(11) 2803–2824.
- Cui, Ruomeng, Jingyun Li, Meng Li, Lili Yu. 2021. Wholesale price discrimination in global sourcing. *Manufacturing & Service Operations Management* **23**(5) 1096–1117.
- Cui, Ruomeng, Jun Li, Dennis J Zhang. 2020. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science* **66**(3) 1071–1094.
- Cui, Ruomeng, Hyoduk Shin. 2018. Sharing aggregate inventory information with customers: Strategic cross-selling and shortage reduction. *Management Science* **64**(1) 381–400.
- Cui, Ruomeng, Dennis J Zhang, Achal Bassamboo. 2019. Learning from inventory availability information: Evidence from field experiments on amazon. *Management Science* **65**(3) 1216–1235.
- Dasgupta, Sudipto, Alminas Žaldokas. 2019. Anticollusion enforcement: Justice for consumers and equity for firms. *The Review of Financial Studies* **32**(7) 2587–2624.
- Debo, Laurens, Garrett van Ryzin. 2013. Leveraging quality information in stock-outs. *Chicago Booth Research Paper* (13-58).
- Dellarocas, Chrysanthos. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science* **52**(10) 1577–1593.
- Duan, Wenjing, Bin Gu, Andrew B Whinston. 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* **45**(4) 1007–1016.
- Federal Trade Commission. 2019. Ftc brings first case challenging fake paid reviews on an independent retail website. <https://www.ftc.gov/news-events/press-releases/2019/02/ftc-brings-first-case-challenging-fake-paid-reviews-independent>.
- Forman, Chris, Anindya Ghose, Batia Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* **19**(3) 291–313.
- Gallino, Santiago, Antonio Moreno. 2014. Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science* **60**(6) 1434–1451.

- 
- Ge, Yanbo, Christopher R Knittel, Don MacKenzie, Stephen Zoepf. 2020. Racial discrimination in transportation network companies. *Journal of Public Economics* **190** 104205.
- Ghose, Anindya, Panagiotis G Ipeirotis, Beibei Li. 2014. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* **60**(7) 1632–1654.
- Hong, Yili, Paul A Pavlou. 2017. On buyer selection of service providers in online outsourcing platforms for it services. *Information Systems Research* **28**(3) 547–562.
- Horton, John J. 2019. Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution. *Management Science* **65**(8) 3518–3540.
- Kalkanci, Basak, Erica L Plambeck. 2020. Reveal the supplier list? A trade-off in capacity vs. responsibility. *Manufacturing & Service Operations Management* **22**(6) 1251–1267.
- Kalwani, Manohar U, Narakesari Narayandas. 1995. Long-term manufacturer-supplier relationships: Do they pay off for supplier firms? *Journal of Marketing* **59**(1) 1–16.
- Kesavan, Saravanan, Vishal Gaur, Ananth Raman. 2010. Do inventory and gross margin data improve sales forecasts for us public retailers? *Management Science* **56**(9) 1519–1533.
- Khernamnuai, Warut, Hyunji So, Maxime C Cohen, Yossiri Adulyasak. 2021. Selecting cover images for restaurant reviews: Ai vs. wisdom of the crowd. *Social Science Research Network* .
- Lee, Hau L, Venkata Padmanabhan, Seungjin Whang. 1997. Information distortion in a supply chain: The bullwhip effect. *Management Science* **43**(4) 546–558.
- Leider, Stephen, William S Lovejoy. 2016. Bargaining in supply chains. *Management Science* **62**(10) 3039–3058.
- Li, Jun, Serguei Netessine. 2020. Higher market thickness reduces matching rate in online platforms: Evidence from a quasiexperiment. *Management Science* **66**(1) 271–289.
- Li, Yuanchen, Lauren Xiaoyuan Lu, Susan F Lu. 2021. Do social media dominate government report cards in influencing nursing home demand? *Available at SSRN 3531964* .
- LinkedIn, G2 Crowd. 2020. The rise of b2b product reviews. <https://business.linkedin.com/content/dam/me/business/en-us/marketing-solutions/resources/pdfs/linkedin-crowd-b2b-product-review-book.pdf>.
- Liu, Qian, Garrett J Van Ryzin. 2008. Strategic capacity rationing to induce early purchases. *Management Science* **54**(6) 1115–1131.
- Liu, Yong. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing* **70**(3) 74–89.
- Lu, Lauren Xiaoyuan, Martin A Lariviere. 2012. Capacity allocation over a long horizon: The return on turn-and-earn. *Manufacturing & Service Operations Management* **14**(1) 24–41.

- 
- Luca, Michael. 2016. Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* (12-016).
- Luca, Michael, Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science* **62**(12) 3412–3427.
- Mayzlin, Dina. 2006. Promotional chat on the internet. *Marketing Science* **25**(2) 155–163.
- Mayzlin, Dina, Yaniv Dover, Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104**(8) 2421–55.
- McAfee, Andrew, Erik Brynjolfsson. 2017. *Machine, platform, crowd: Harnessing our digital future*. WW Norton & Company.
- McKinsey & Company. 2019. How b2b online marketplaces could transform indirect procurement <https://www.mckinsey.com/business-functions/operations/our-insights/how-b2b-online-marketplaces-could-transform-indirect-procurement>.
- Mejia, Jorge, Shawn Mankad, Anandasivam Gopal. 2021. Service quality using text mining: Measurement and consequences. *Manufacturing & Service Operations Management* **23**(6) 1354–1372.
- Mejia, Jorge, Chris Parker. 2021. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* **67**(1) 166–184.
- Minsky, Laurence, Keith A Quesenberry. 2015. How b2b marketers can get started with social media. <https://hbr.org/2015/12/how-b2b-marketers-can-get-started-with-social-media>.
- Minsky, Laurence, Keith A Quesenberry. 2016. How b2b sales can benefit from social selling. <https://hbr.org/2016/11/84-of-b2b-sales-start-with-a-referral-not-a-salesperson>.
- Moreno, Antonio, Christian Terwiesch. 2014. Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research* **25**(4) 865–886.
- Papanastasiou, Yiangos, Nicos Savva. 2017. Dynamic pricing in the presence of social learning and strategic consumers. *Management Science* **63**(4) 919–939.
- Phillips, Robert, A Serdar Şimşek, Garrett Van Ryzin. 2015. The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* **61**(8) 1741–1759.
- Shea, Elizabeth. 2020. Why b2b reviews should be an essential part of your marketing mix. <https://www.forbes.com/sites/forbescommunicationscouncil/2020/09/22/why-b2b-reviews-should-be-an-essential-part-of-your-marketing-mix/?sh=7c1059b533e0>.
- Sinchaisri, Park, Gad Allon, Maxime Cohen. 2019. The impact of behavioral and economic drivers on gig economy workers. *Academy of Management Proceedings*, vol. 2019. Academy of Management Briarcliff Manor, NY 10510, 10216.
- Song, Jing-Sheng, Geert-Jan van Houtum, Jan A Van Mieghem. 2020. Capacity and inventory management: Review, trends, and projections. *Manufacturing & Service Operations Management* **22**(1) 36–46.

- 
- Sun, Monic. 2012. How does the variance of product ratings matter? *Management Science* **58**(4) 696–707.
- Tadelis, Steven. 2016. Reputation and feedback systems in online platform markets. *Annual Review of Economics* **8** 321–340.
- Tadelis, Steven, Florian Zettelmeyer. 2015. Information disclosure as a matching mechanism: Theory and evidence from a field experiment. *American Economic Review* **105**(2) 886–905.
- Terwiesch, Christian, Z Justin Ren, Teck H Ho, Morris A Cohen. 2005. An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Management Science* **51**(2) 208–220.
- Utpal M. Dholakia. 2015. What to do when satisfied b2b customers refuse to recommend you. <https://hbr.org/2015/08/what-to-do-when-satisfied-b2b-customers-refuse-to-recommend-you>.
- Weise, Elizabeth. 2017. That review you wrote on amazon? priceless. <https://www.usatoday.com/story/tech/news/2017/03/20/review-you-wrote-amazon-priceless/99332602/>.
- Xu, Yuqian, Mor Armony, Anindya Ghose. 2021. The interplay between online reviews and physician demand: An empirical investigation. *Management Science* **67**(12) 7344–7361.
- Zhu, Feng, Xiaoquan Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing* **74**(2) 133–148.
- Zhu, Kevin. 2004. Information transparency of business-to-business electronic markets: A game-theoretic analysis. *Management Science* **50**(5) 670–685.