

**TOWARDS A BETTER FUTURE OF WORK: ENHANCING EFFICIENCY,
PROMOTING FAIRNESS, AND UPHOLDING LABOR STANDARDS IN THE
CONTEMPORARY WORKPLACE**

A Dissertation
Presented to
The Academic Faculty

By

Li Ding

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Scheller College of Business
Operations Management

Georgia Institute of Technology

May 2024

© Li Ding 2024

**TOWARDS A BETTER FUTURE OF WORK: ENHANCING EFFICIENCY,
PROMOTING FAIRNESS, AND UPHOLDING LABOR STANDARDS IN THE
CONTEMPORARY WORKPLACE**

Thesis committee:

Dr. Basak Kalkanci
Scheller College of Business
Georgia Institute of Technology

Dr. Vinod Singhal
Scheller College of Business
Georgia Institute of Technology

Dr. Seyed Emadi
Kenan-Flagler Business School
University of North Carolina - Chapel Hill

Dr. D.J. Wu
Scheller College of Business
Georgia Institute of Technology

Dr. Manpreet Hora
Scheller College of Business
Georgia Institute of Technology

Date approved: April 18, 2024

To my grandfather Jinling Zhou

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation committee members for their support throughout this journey. First, I am extremely grateful to my advisor, Dr. Basak Kalkanci. Your guidance and mentorship have been instrumental in shaping this work. You are always there to provide thoughtful feedback, encourage me to dig deeper, and refine my ideas through multiple rounds of brainstorming and revisions. You have demonstrated to me what it means to be a curious, dedicated, and truth-seeking researcher, whom I aspire to be. I am also deeply thankful to Dr. Seyed Emadi, Dr. Manpreet Hora, Dr. Vinod Singhal, and Dr. D.J. Wu. Seyed, I am very fortunate to have you as my coauthor. You always bring insightful perspectives to our discussions. I appreciate those sessions where we went through every sentence in the paper. Your advice about maintaining intellectual rigor in writing and presentation is invaluable to me. Manpreet, I cannot express how thankful I am for your support during my job market. You are the most enthusiastic professor I have encountered in the classroom, and I am honored to have worked as your TA. Your dedication to teaching has been an inspiration, and I have learned immensely from your approach. Vinod, thank you for always believing in me, encouraging me, and supporting me throughout this process. I have enjoyed our conversations about research, life, family, nice restaurants, etc. Your warmth and mentorship have made Atlanta feel like a second home. Last but not least, I am forever grateful to D.J. for his support and for seeing the potential in me. You have demonstrated to me the qualities of an open-minded, knowledgeable, encouraging, and humble researcher. I always learn new things from our discussions, and your willingness to provide guidance and feedback has been invaluable.

I also would like to thank the faculty and fellow PhD students in the Operations Management Area. The supportive and collegial environment at Scheller has made my PhD journey much smoother. I am fortunate enough to have a close and friendly PhD cohort. Chris, my officemate, no one is more dedicated, persistent, and self-disciplined than you. I

will always cherish those days we spent finding classrooms, studying for exams, and solving equations. Swanand and Zeya, we have learned and grown together throughout this program. Your companionship, friendship, and support have meant a lot to me. Abhinav, Shikha, Daniel, and Cagri, this journey would have been far less enjoyable without you being around. Shikha, you have always said that I encouraged and supported you, but the sentiment is entirely mutual. You have been a source of strength and inspiration for me.

Special thanks go to my family members. Neo, my husband and my closest friend, thank you for your unconditional love, support, and understanding. You have always believed in me and encouraged me to aim higher. I am deeply grateful to my grandparents, Jinling and Lanzhen, for raising me and being my first teachers. Without your support and dedication, this journey would not have been possible. I would like to thank my parents, Qing and Yihua. Thank you for always encouraging me to pursue my dreams, even thousands of miles away from home. Finally, I would like to thank my aunt and uncle, Lan and Ziqi, for being my family here. Thank you for making me feel always supported.

TABLE OF CONTENTS

Acknowledgments	iv
List of Tables	x
List of Figures	xii
Summary	xiv
Chapter 1: Trips for Tips? Implications of Tips on Drivers' Search Behavior . .	1
1.1 Introduction	1
1.2 Literature Review	7
1.2.1 Incentive Design in Service Operations	7
1.2.2 Tipping in Service Operations	8
1.2.3 Bayesian Learning	9
1.3 Data	10
1.3.1 Research Context	10
1.3.2 Dataset	11
1.4 Impact of Past Observations and Experiences on Relocation Decisions . . .	13
1.4.1 Evidence for History-Dependent Relocation Behavior	15
1.4.2 Motivation for the Structural Model	20

1.5	Structural Model	21
1.5.1	Preliminaries	21
1.5.2	Bayesian Learning Process	22
1.5.3	Relocation Decision Making	25
1.6	Estimation	26
1.6.1	Estimation Procedure	26
1.6.2	Identification	27
1.6.3	Data Processing	28
1.6.4	Estimation Results	31
1.6.5	Comparison of Model Fit	33
1.7	Policy Simulations	33
1.7.1	Tipping Policies	34
1.7.2	Simulation Details	35
1.7.3	Simulation Results	39
1.8	Concluding Remarks	44
 Chapter 2: Reducing Gender Disparities in Online Labor Platforms: The Role of Performance Feedback and Task Exposure		 47
2.1	Introduction	47
2.2	Literature Review	50
2.2.1	Gender Disparities in Online Platforms and Mitigation Strategies	50
2.2.2	Gender Gap in Performance Self-Evaluation Biases	54
2.3	Method	56
2.3.1	Empirical Context	56

2.3.2	Experiment Design	57
2.3.3	Sample and Balance Check	62
2.4	Results	65
2.4.1	Empirical Approaches	65
2.4.2	Is There a Gender Earnings Gap in Phase 1?	65
2.4.3	Mitigation Strategies: Performance Feedback and Early Exposure to An HH Task	67
2.4.4	Mechanisms: The Role of Performance Self-evaluation Biases	73
2.5	Concluding Remarks	77
	Chapter 3: Beyond Health Risks: The Impact of COVID-19 on Labor Viola- tions in Global Supply Chains	82
3.1	Introduction	82
3.2	Literature Review	85
3.2.1	Antecedents of Labor Violations	85
3.2.2	The Impact of COVID-19 on Operations	87
3.3	Method	88
3.3.1	Data	88
3.3.2	Model	89
3.4	Results	91
3.4.1	Parallel Trend Test	92
3.4.2	Effect of COVID-19 on Labor Violations	93
3.4.3	Heterogeneity across Characteristics of Suppliers	93
3.5	Concluding Remarks	95

3.6 Future Steps	97
Appendices	99
Appendix A: Trips for Tips? Implications of Tips on Drivers' Search Behavior .	100
Appendix B: Reducing Gender Disparities in Online Labor Platforms: The Role of Performance Feedback and Task Exposure	118
Appendix C: Beyond Health Risks: The Impact of COVID-19 on Labor Viola- tions in Global Supply Chains	127
References	133
Vita	143

LIST OF TABLES

1.1	Descriptive statistics	12
1.2	Multinomial logit regression results	19
1.3	Estimates of the utility function parameters (as described in subsection 1.5.3)	31
1.4	Model comparison table	33
1.5	Simulation results	39
2.1	Control and treatment conditions	62
2.2	Worker characteristics by gender groups (balance check)	64
2.3	Piece Rate Pay and Hourly Pay in Phase 2	69
2.4	Comparison among treatment groups	71
2.5	Mediation Analysis	76
3.1	Parallel trend test (US buying firms)	94
3.2	DID analysis results (US buying firms)	95
3.3	Heterogeneity Analysis over 1-year post period (US buying firms)	96
A.1	Definitions of features	102
A.2	Out-of-sample predictive performance	103
A.3	Multinomial logit regression results - moving averages over 5 periods	107
A.4	Estimates of the location-specific dummy variable coefficients	111

A.5	Comparison of the imputed prior belief parameters and the actual value of the attribute mean	112
B.1	Mediation effect of risk aversion	125
B.2	The moderation effect of positive and negative feedback	126
C.1	Heterogeneity Analysis over 2-year post period (US buying firms)	128
C.2	Heterogeneity Analysis over 3-year post period (US buying firms)	129
C.3	Parallel trend test (Chinese buying firms)	130
C.4	DID analysis results (Chinese buying firms)	131
C.5	Heterogeneity analysis over 1-year post period (Chinese buying firms)	132

LIST OF FIGURES

1.1	Relocation frequencies in the two biggest and two smallest clusters (the respective percentages of drivers in each cluster are given in parentheses) . .	17
1.2	The generation process of attribute realizations	37
1.3	Changes in relocation probability in different locations and times (Hinder learning - Baseline)	40
1.4	Unmet demand in different locations and times under the Baseline scenario	42
1.5	Changes in relocation probabilities in different locations (Service charge - Baseline)	43
2.1	Field experiment design flow	58
2.2	Average accuracy in Task 1 of workers who self-selected in Phase 1	66
2.3	Average performance self-evaluation bias by gender	66
2.4	Average pay (\$) and hourly pay (\$/hour) of workers who self-selected (left) and assigned to HH (right) in Phase 1	67
2.5	Gender differences in the percentage of choosing HH for Task 1	68
2.6	Average hourly pay by gender and treatment	71
2.7	Full path mediation analysis	78
3.1	Control and Treatment Groups (the US)	90
3.2	Parallel trend (companies with HQs in the US)	92
A.1	Temporal and spatial variations in tip percentages	106

A.2	The distribution of Box-Cox transformed attribute realizations	109
B.1	User interface on Upwork	118
B.2	User interface on Prolific	119
B.3	Math proofreading task on Upwork	120
B.4	Math grading example question (EL)	121
B.5	Math grading example question (HH)	121
B.6	Recruitment ad on Prolific	122
B.7	Random quality check on Mturk	123
B.8	Performance feedback and control messages in Prolific	124

SUMMARY

The three essays in this dissertation examine critical issues at the intersection of the future of work and labor practices. As the nature of employment and the workplace continues to evolve, it is essential to consider how these changes might affect organizational efficiency, worker welfare, and broader societal implications. In the dissertation, I focus on three important contexts that shape the future of work, which are digital service platforms, gig economy platforms, and global supply chains, in each essay. I adopt empirical methodologies including structural estimation, policy simulations, online experiments, and econometric analyses to offer actionable insights for organizations and policymakers to navigate the complexities of the changing work landscape and work towards a better future of work.

Essay 1: Trips for Tips? Implications of Tips on Drivers' Search Behavior

Motivated by the increasing adoption of tipping in ride-sharing services, we examine how tips influence drivers' relocation decisions and subsequent outcomes such as driver earnings and system efficiency in an empirical setting. We first show that drivers' relocation decisions are not only driven by their inherent heterogeneity, but are also affected by their learning from past experiences. We propose a discrete-choice structural model that captures drivers' learning from their past experiences in a Bayesian fashion and estimate the model using a large-scale and granular taxi trip dataset. We find that drivers' sensitivity to tip amount is greater than that of the fare amount, suggesting that the source of income is an important leading factor that shapes service providers' relocation decisions. We perform a series of counterfactual analyses to find the impact of policies such as hindering learning by delaying the tip information, and a service charge that replaces the tip with a fixed percentage of the fare on driver earnings and system performance. We find that while tipping is increasingly adopted to supplement driver income, it can create an unintended consequence of lower system efficiency. Although hindering drivers from learning about

tips reduces their hourly income by 12%, it improves the system efficiency, indicated by a 1% decrease in wait time for drivers and 10% decrease in the probability of unmet demand for customers. Replacing tips with a service charge not only increases drivers' hourly income by 4%, but also leads to a small but positive improvement in system efficiency.

Essay 2: Reducing Gender Disparities in Online Labor Platforms: The Role of Performance Feedback and Task Exposure

Given the persistence of gender disparities in task selection and earnings in online labor platforms, we investigate the underlying behavioral mechanisms that drive these disparities and the interventions to mitigate them. We conduct a two-phase experiment on Prolific, where online workers are recruited to grade elementary school-level math problems of two difficulty and payment levels. We find that women earn significantly less than men because they are less likely to select the more difficult task compared to men of comparable capabilities, despite a higher payment for the same expected task completion time. This difference in task selection is partially driven by women's negatively biased self-evaluation of their performance, particularly for more difficult tasks. In terms of mitigation strategies, we show that performance feedback alone may not be sufficient to motivate women to pursue the more difficult tasks later on, because women often choose to work on the easier task in the first phase, for which they have relatively unbiased self-evaluation. Supplementing performance feedback with early exposure to challenging tasks can effectively reduce the gender disparities in task selections and significantly boost women's hourly earnings without adversely affecting men's earnings. Our study provides actionable insights for online labor platforms to tackle gender disparities and build a more inclusive and equitable online work environment.

Essay 3: Beyond Health Risks: The Impact of COVID-19 on Labor Violations in Global Supply Chains

Many global organizations such as the International Labor Organization have called for an in-depth understanding of the impacts of COVID-19 on labor practices globally

beyond the immediate health risks. Responding to these calls, we examine the impacts of the pandemic on the prevalence of worst forms of labor violations in global supply chains and whether the impacts (if any) persist over time. While business shutdowns, production slowdowns, and reduced demand during the pandemic can lead to fewer opportunities of labor violations, the supply chain disruptions induced by unexpected changes in orders as well as diverted compliance resources and attention might aggravate the risks of such violations. Motivated by the empirical ambiguity, we perform Year-to-Year difference-in-difference analyses to examine the impact of the pandemic outbreak on labor violation incidents in global supply chains. Our analysis covers 2385 globally public-listed firms and their supply chains in two different industries (apparel and electronics) and across 8 years (from 2015 to 2022). We find that the pandemic increases labor violations in global supply chains, and the impact persists over a 2- to 3-year post period. Moreover, heterogeneity analyses show that the impact is exacerbated for buying firms with more suppliers, more diversified geographic exposure of suppliers, but is mitigated by longer-term relationship with suppliers. This study provides practical insights for managers, policymakers, and industry regulators on strategies to assess and mitigate the risk of labor violations induced by the pandemic.

CHAPTER 1

TRIPS FOR TIPS? IMPLICATIONS OF TIPS ON DRIVERS' SEARCH BEHAVIOR

1.1 Introduction

In traditional service settings such as the taxicab industry and on-demand service settings such as ride sharing, food and grocery delivery, and household services, service providers have discretion over the hours and/or the locations in which to work. The autonomy of service providers has rendered the match between demand and supply difficult. For example, according to data from five large US cities, taxi drivers have a passenger in the car an average of 30 to 50% of the time they are working, and the low capacity utilization is partially explained by their tendency to drive in congested times of the day (Cramer & Krueger, 2016). In the same set of cities, Uber drivers are estimated to have a passenger in the car only about half the time their app is turned on, despite the driver-passenger matching technologies utilized (Cramer & Krueger, 2016). While service providers cannot always be scheduled centrally to work at the right place and time, monetary incentives or nudges can be implemented to align service providers' behavior with system-wide service requirements. Designing such incentives or nudges has attracted attention from academics, industry practitioners, and policy makers alike (Taylor 2018, Buchholz 2019 and references therein on taxi regulation).

An overlooked incentive mechanism likely to influence independent service provider decisions is tips from customers. Tips make up a significant portion of service providers' overall income in traditional service sectors such as taxi industries, restaurants, and barber shops (Azar, 2003), and have also been increasingly adopted as additional compensation to service providers in on-demand platforms such as Uber, Lyft, Grubhub, and Instacart

(Rapier, 2019). Many of these workers are low-income earners who rely heavily on tips to reach or surpass the minimum wage¹. While tips help to supplement service provider income, they may also distort independent service provider incentives and create inefficiencies. Service providers may rely on past experiences with tips to learn about the earning potential of different regions and decide where to serve next. If service providers are driven by the expectation of higher tips rather than demand, imbalance between supply and demand can be exacerbated. Such imbalances can contribute to societal externalities such as limited access to services in the areas where tips are expected to be low. This was indeed cited by Uber as a rationale for not including a tipping option in 2016 (which has since been reversed). Uber has claimed that *“tipping could give drivers an incentive to spend more time where tips are likely to be the highest — typically the wealthiest neighborhoods”* (Uber, 2020a). Anecdotal evidence from Uber Eats suggests that drivers’ income heavily relies on tips, with many of them cherry-picking lucrative areas which are often crowded with drivers (Abramson, 2023). Despite these claims, prior research has not investigated how tips influence service providers’ decisions regarding where to provide service and subsequent market outcomes such as provider income and system efficiency, which is the objective of this paper.

Examining the role of tips in drivers’ customer search decision-making processes is important. As ridesharing continues to expand, ride-hailing apps for licensed taxis have emerged to modernize the taxi industry, such as Curb² in North America and Grab in Southeast Asia³. Since 2022, Uber has extended its reach into the New York taxi industry, with plans to include every taxi driver on its platform by 2025⁴. In addition to these global companies, local taxi companies in major cities in the US, such as Los Angeles⁵

¹<https://points.datasociety.net/racing-for-tips-4816da5b5096>

²<https://www.gocurb.com/business>

³<https://www.grab.com/sg/about/>

⁴<https://www.reuters.com/business/autos-transportation/uber-reaches-deal-list-all-new-york-city-taxis-app-wsj-2022-03-24/>

⁵<https://rideyellow.com/cities/los-angeles-area/taxi-los-angeles/>

and Atlanta⁶, have also introduced app-based taxi-hailing services. The taxi apps offer new functionalities including contactless in-app payments, which present opportunities to implement changes to tip policies. In a similar vein, tip policies that vary in the granularity and timing of tip-related information shown to drivers are being tested by many on-demand service platforms. For example, Doordash shows a detailed breakdown of earnings (including the base pay, promotions, and customer tip) as soon as a delivery is completed (Xu, 2019), consistent with Curb’s current tip policy. On Uber, drivers can review tip earnings corresponding to each past trip (including timestamps and street names) after they rate a passenger (Uber, 2020a). In contrast, Postmates allows drivers to review tip earnings corresponding to a specific past delivery (without the customer address) at least 24 hours after their delivery is completed (Postmates, 2021). Instacart announced plans to replace tip with a service charge distributed evenly among the entire fleet of drivers to create “more consistent pay with fewer variables”. However, there is limited evidence on which of these policies are most effective. Since the aforementioned tipping policies vary the salience of information on tips, insights into whether and how service providers learn about tips and how tips would affect their relocation decisions will help inform tipping policies and strategies for information sharing with service providers.

Our paper examines the following questions: (1) Do drivers learn about earning opportunities available in different locations and times based on the tip-related information they are exposed to as they complete more trips over time, and if so, how? (2) How do the policy interventions that vary the salience of information on tips (such as those outlined above leading to revealed or withheld tip information) affect drivers’ customer search behavior, income, and system efficiency?

To answer these questions, we analyze individual drivers’ relocation decisions, which refer to their movement in search of new customers after a drop-off. We first provide empirical evidence that drivers’ relocation decisions are dependent on their previous obser-

⁶<https://atlantacheckercab.com/get-a-cab/>

vations and experiences, after controlling for the effect of driver heterogeneity and location. These past observations and experiences include past earnings from fare and tips, supply and demand conditions (as captured by pickups and drop-offs observed nearby), and speed and wait time between trips, which reflect the inconvenience of customer search, at different locations and times. We refer to these six features reflecting past observations and experiences as trip *attributes*. We then explicitly model the behavioral process by which drivers' history-dependence in relocation decisions materializes. Specifically, we propose a discrete-choice structural model that captures drivers' learning from their past experiences in a Bayesian fashion. In our Bayesian learning framework, as drivers complete more trips, they learn about the aforementioned attributes in different locations and times, and make relocation decisions based on their updated beliefs. Finally, using our framework, we examine how interventions on tipping policies affect drivers' relocation decisions, income, and system efficiency.

We utilize a large scale and granular data set on taxi trip records of the licensed New York City (NYC) yellow taxicab drivers to estimate our structural model. The data provides information including anonymized car identifier, driver identifier, pick-up and drop-off time stamps and GPS coordinates, ride time duration, ride distance, number of passengers, fare amount, tip amount, toll amount, and surcharge. We complement this data set with the geographic data from the NYC City Planning and the weather history data from the National Oceanic Atmospheric Administration (NOAA) to control for exogenous factors that might affect drivers' relocation decisions such as location characteristics and weather conditions. Overall, our approach focuses on individual service providers' relocation decisions, which became possible only recently due to increased data availability (Chandar et al., 2019; Zhang et al., 2020). This approach is in contrast with previous research on tips, which was predominantly based on surveys (Azar, 2020).

Our results show that drivers learn about the tip distribution of different locations and times, in addition to the distribution of other conventional attributes expected to influence

drivers' expectations of earnings (including fare, number of nearby pickups and drop-offs, speed, and wait time between trips). All else being equal, drivers prefer to relocate to locations and work during times wherein their beliefs about the average tip percentage are higher. Furthermore, drivers' sensitivity to the tip amount is greater than that of the fare amount, suggesting that the source of income (tips vs. fare) is an important leading factor that shapes service providers' relocation decisions. We show that our Bayesian learning model outperforms the rational expectations model, wherein drivers' beliefs match actual attribute distributions perfectly, in terms of both in-sample and out-of-sample model fit. This provides support for the existence of driver learning within the data.

Our policy simulation results reveal both benefits and costs induced by drivers' learning from tips. Specifically, providing detailed tip information in a timely manner helps drivers to increase their hourly income through identifying locations and times wherein higher earnings are expected. However, it also reduces the system efficiency because drivers experience longer wait times due to intensified competition in lucrative regions and unmet demand is higher in those less lucrative regions which can be underserved. On the other hand, withholding information about tips decreases drivers' hourly income by about 12% but improves the system efficiency, indicated by a 1% decrease in wait times for drivers, a 10% decrease in the magnitude of unmet demand, and a 2% decrease in the probability of unmet demand. This is because, guided by (possibly less accurate) prior beliefs about tips, some drivers may end up relocating to the locations wherein the actual tips are low. Since low-tip regions tend to be underserved under the baseline scenario, withholding information about tips mitigates the shortage of supply and reduces unmet demand in these locations. Finally, replacing tips with an equivalent service charge (a fixed percentage of the base fare paid by customers to the drivers) not only increases drivers' hourly tip income and total income by about 4%, but also improves system efficiency, albeit modestly. Under the service charge policy, the magnitude of the efficiency improvement is limited because drivers are not sufficiently motivated to move to low-tip locations and times, which coin-

cidentally also have low demand and longer expected wait times for drivers. Therefore, the service charge policy leads to a more modest mitigation of system inefficiency than withholding information about tips.

Our paper makes several contributions. First, we contribute to the growing body of knowledge on information and incentive design for transportation service systems (Kabra et al., 2020; Erkan et al., 2022; Feldman et al., 2022; Allon et al., 2023). In particular, we complement the extant literature on incentive design in service operations by identifying the source of income (tips vs. fare) as an important leading factor that shapes service providers' relocation decisions. While previous research has focused on levers such as surge pricing (Cachon et al., 2017; M. K. Chen, 2016; Hall et al., 2019; Bimpikis et al., 2019; Guda & Subramanian, 2019; Ozkan & Ward, 2020) and information sharing (Zhang et al., 2020; Cui et al., 2020; Karacaoglu et al., 2018; Kulkarni & Kalkanici, 2021), tips are largely overlooked in the literature. However, increased prevalence of tipping as an incentive scheme in technology-facilitated service contexts highlights the need for analyzing the effect of customer tips. Furthermore, our structural approach allows us to perform policy simulations, which provide practical implications for ride-dispatching systems on the impact of tipping policies on drivers' income and system performance. In doing so, we reveal a previously overlooked trade-off: While interventions that enhance the salience of tips help to increase drivers' income through more informed relocation decisions, they can also lead to unintended consequences such as higher wait times for drivers and greater unmet demand for customers, especially in some underprivileged areas where tips may be low.

The rest of the paper is structured as follows. In section 1.2, we review the existing literature related to our work. section 1.3 describes the research context and the data used in our analysis. section 1.4 provides empirical evidence that demonstrates drivers' learning behavior. section 1.5 presents the structural model developed to capture drivers' relocation decision-making processes. section 1.6 introduces the estimation procedure, identification,

data processing steps and the estimation results. In section 1.7, we examine the impact of different tipping policies on drivers' hourly income and system efficiency. Finally, section 1.8 concludes the paper with a discussion of interesting avenues for future research.

1.2 Literature Review

Our paper contributes to the literature on incentive design in service operations, the role of tips in service operations, and Bayesian learning. Below, we review each of these streams of literature and explain how we contribute to them.

1.2.1 Incentive Design in Service Operations

Tips are important sources of income for many workers in the service industry (Azar, 2009). Incentive manipulations are crucial for exerting indirect control over discretionary service providers, as their decisions cannot be dictated through central scheduling. As such, there is growing interest in Operations Management (OM) on the design of effective levers to coordinate supply and demand in these settings, thereby improving operational performance and social welfare (Y.-J. Chen et al., 2020).

A growing stream of research has focused on service providers' relocation decisions and the incentive mechanisms to influence them. These incentive mechanisms include spatial surge pricing for balancing supply and demand across locations (Guda & Subramanian, 2019; Ozkan & Ward, 2020; Buchholz, 2018) and/or subsidies for rides from low-demand regions (Bimpikis et al., 2019). Spatial real-time information sharing on demand and/or supply conditions is found to be highly influential on drivers' relocation decisions and earnings (Zhang et al., 2020; Karacaoglu et al., 2018; Kulkarni & Kalkanci, 2021; Jiang et al., 2020), consistent with the growing recognition of the importance of information design in service settings (Küçükgül et al., 2022; Blair et al., 2022). Some papers have also identified driver experience as crucial for shaping relocation behavior. For example, drivers may utilize information with varying levels of success in searching for customers depend-

ing on their experience levels (Zhang et al., 2020). Haggag et al. (2017) demonstrate that a driver’s cumulative experience in a neighborhood facilitates his search for customers in that neighborhood.

Our work is unique in uncovering whether and how providers learn from customer-generated monetary incentives (i.e., tips) in discretionary service settings. That is, we demonstrate that providers’ past experiences with tips can influence their relocation decisions by affecting their expectation of earnings in different regions and times. By doing so, we establish the source of income (tips vs. fare) as an important factor in discretionary service providers’ relocation decisions.

1.2.2 Tipping in Service Operations

Since tips account for a significant portion of service providers’ total income (Azar, 2003), tipping has attracted attention from multiple disciplines including anthropology, economics, tourism management, human resource management, marketing, and social psychology (Alexander et al., 2021). We refer the reader to Azar (2020) and Lynn et al. (1993) for excellent reviews of the literature on consumer tipping behavior.

A stream in this literature explores the antecedents of tipping. Lynn et al. (1993) and references therein suggest that factors such as server friendliness and attractiveness, and customer race, gender and frequency of visit are strong predictors of customers’ tipping behavior. Interestingly, previous literature, which mainly focuses on restaurant industry, finds only a modest sensitivity of tips to perceived service quality (Azar, 2020; Lynn et al., 1993). Recently, analyzing more than 40 million Uber trips, Chandar et al. (2019) explore explanatory factors of variations in riders’ tipping behavior. They observe that neighborhood-specific characteristics such as household income, race composition, and education are important in explaining variations in tipping, and demand-side factors explain roughly three times more of the observed tipping variation than the supply-side variables or features of the trip. This suggests that tipping outcomes tend to reflect personal charac-

teristics of the consumer, more so than those of the provider or even the service quality.

Previous research has also explored how tipping influences service providers' behavior. In discretionary service settings, service providers' expectations of tips presumably influence where and when they choose to work; however, this relationship is not yet well studied. In addition, technology-facilitated service interactions enable consumer tipping behavior to be manipulated (e.g., through tip recommendations). Chandar et al. (2019) find a moderate impact of suggested tip amount on riders' tipping behavior through field experiments. Through field experiments on a laundry service platform, Alexander et al. (2021) find that tip recommendations affect the tip amount, but not customer satisfaction, frequency, and spending. These papers primarily focus on how *consumers* respond to such technology-facilitated interventions. In contrast, our work is unique in studying the effects of tipping-related interventions on *service provider behavior* through policy simulations.

1.2.3 Bayesian Learning

Our Bayesian learning model builds upon the seminal work by Erdem and Keane (1996) that applies Bayesian learning to explicitly model how consumers' brand choice behavior evolves over time. The proposed framework has been widely applied to analyze individuals' learning processes under uncertainty and discrete choice behavior in various contexts (Zhang et al., 2020; Narayanan & Manchanda, 2009; Shin et al., 2012; Zhao et al., 2013; Y. Huang et al., 2014; Wu et al., 2015; Sriram et al., 2015; Emadi & Swaminathan, 2018). For instance, Y. Huang et al. (2014) model contributors' learning about the potential of their ideas and implementation costs from peer voting on an idea crowdsourcing platform. Emadi and Swaminathan (2018) explore callers' learning about the waiting time from previous waiting experiences in call centers. Zhang et al. (2020) consider how drivers learn from conventional signals (such as the observed pickups and drop-offs, and fare income) in an Asian city where tips are rarely expected or offered. In their model, drivers update their beliefs about the unobserved earning potentials of different location-time sets after

experiencing signals, and the information content of signals is unobserved and treated to be random. Through simulations, the paper shows that aggregating noisy individual-level signals across drivers temporally or spatially through information sharing helps to improve drivers' income. Different from Zhang et al. (2020), we assume that drivers use their past experiences and observations directly in updating their beliefs about the average levels of attributes affecting their relocation decisions (such as tip percentage, fare income, average speed, number of nearby pickups and drop-offs, and wait time). As such, we disentangle the impact of drivers' intrinsic preferences for each of these attributes from that of their prior beliefs and the learning process, which enables us to elicit the role of tips and examine the impact of different tipping policies. Our approach is consistent with Sriram et al. (2015) which investigates how customers learn about the quality of a video on demand (VOD) service, wherein the actual service quality experienced by customers in the VOD setting is observable through the number of new movies received by each household.

1.3 Data

1.3.1 Research Context

Our study is based on NYC's yellow taxicab market. We utilize a data set of trip records by licensed NYC yellow taxi drivers in 2009 between January and July, available through an online repository of social, behavioral, and health sciences research data⁷. The anonymized driver identifier which allows us to track each driver's trajectory of pick-ups and drop-offs is only available for 2009 but not future years.

According to the NYC Taxi & Limousine Commission (TLC) 2009 Annual report⁸, there were 13,237 yellow taxi medallion licenses and 48,524 licensed medallion drivers at the end of 2009. A TLC 2011 survey⁹ shows that about 24.6% (41.1%) of the population took a taxi daily (weekly). The NYC yellow taxi market presents an ideal setting to study

⁷<https://www.openicpsr.org/openicpsr/>

⁸https://www1.nyc.gov/assets/tlc/downloads/pdf/tlc_annual_report_2009.pdf

⁹https://www1.nyc.gov/assets/tlc/downloads/pdf/tot_survey_results_02_10_11.pdf

drivers' relocation decisions for the following reasons. First, there is a market entry restriction that limits the total number of medallions available in the market (Buchholz, 2018), making the supply side relatively stable in 2009. Second, drivers largely rely on their own experiences for customer searching (Haggag et al., 2017), and make decentralized relocation decisions independently. In 2009, the taxi industry was offline and there was no centralized dispatching platforms such as Curb or Uber for real-time information sharing or coordination (Buchholz, 2018), meaning that all drivers acted on their own and there were no confounding effects of surge pricing on drivers' relocation decisions. Third, drivers have the discretion to choose when and how long to drive with an objective to maximize their earnings. Lease-drivers lease a car from a fleet company or owners and pay a leasing fee per day (about \$100) or a flat fee per week. Owner-drivers pay annual maintenance and repair, insurance, and license fees. All drivers pay for gas, annual TLC fees (\$100), and fines (Haggag et al., 2017, p. 73). All drivers keep all fares and tips¹⁰. Fourth, according to TLC regulations¹¹, only yellow taxicabs are allowed to serve street-hailing customers anywhere in NYC. Yellow taxicabs are not allowed to serve prearranged pickups via telephone or the Internet (Buchholz, 2018; Haggag et al., 2017). This regulation ensures that all trips in our dataset occurred as a result of drivers searching rather than central dispatching (Buchholz, 2018).

1.3.2 Dataset

Our taxi trip dataset provides information including anonymized car identifier, driver identifier, pick-up and drop-off time stamps and GPS coordinates, ride time duration, ride distance, number of passengers, fare amount, tip amount, toll amount, and surcharge. The NYC taxi dataset has been used to study the labor supply decisions of taxi drivers in other contexts (Haggag et al., 2017; Camerer et al., 1997; Farber, 2015).

In our study, we focus on the subset of taxi trips both picked up and dropped off within

¹⁰<https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page>

¹¹<https://www1.nyc.gov/site/tlc/vehicles/get-a-vehicle-license.page>

Manhattan for the following two reasons. First, intra-Manhattan trips are representative as they account for about 75% of the total revenue and 90% of the total trips (Buchholz, 2018). Second, intra-Manhattan trips exclude airport trips. Hence, the confounding effect of large fares is mitigated, and all fares are calculated using a consistent pricing scheme. Standard metered fares include a \$2.5 initial charge and additional fares are accumulated based on the distance traveled and low-speed time. In contrast, trips between Manhattan and the JFK Airport charge a flat fee of \$52¹². We provide the descriptive statistics of all trips picked up and dropped off in Manhattan in Table 1.1.

Table 1.1: Descriptive statistics

Variable	Mean	Median	Std. Dev.	Min	Max
<i>Trip-level characteristics</i>					
Average trip speed (MPH)	12.455	11.600	5.004	2.000	70.000
Trip fare (\$)	9.268	8.500	3.521	2.500	208.100
Tip percentage	19.1%	18.2%	9.0%	0	199.7%
Time gap between trips (minutes)	19.142	8.000	90.893	0.017	42,085
<i>Temporal characteristics</i>					
Weekday indicator	0.734	1	0.442	0	1
Precipitation (inches/hour)	0.159	0	0.358	0	2.300
Fog/haze level indicator	0.609	0	0.737	0	2

Note: Descriptive statistics are based on the data in 2009.

In addition to NYC taxi trip dataset, two other datasets were combined to obtain the geographic and demographic information of Manhattan. First, we rely on the 2010 TIGER/Line Shapefiles Census Tracts to map the GPS coordinates corresponding to each pick-up and drop-off in our dataset to the 2010 NYC Census tracts. We then rely on the geographic files on the NYC City Planning website¹³ to map the census tracts to Public Use Microdata Areas (PUMAs). The five boroughs in NYC (Manhattan, Brooklyn, Queens, The Bronx, and Staten Island) are divided into 2,168 census tracts. The 2,168 census tracts are nested within 55 PUMAs. For tractability of this study, service zones are identified using PUMA boundaries, which approximate community districts and also correspond to the conventional views of NYC neighborhoods. Specifically, the Manhattan borough is divided into

¹²<https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page>

¹³<https://www.nyc.gov/site/planning/data-maps/open-data.page>

10 PUMAs such as Central Harlem, Upper East Side, Chelsea, Clinton & Midtown Business District, and each PUMA has a minimum population of 100,000 people. Second, we collect the weather history in Manhattan from the NOAA website¹⁴, because weather conditions have been shown to affect drivers’ driving behavior in the previous literature (Camerer et al., 1997; Farber, 2015). Before variable construction, we took several steps to preprocess the data including removing trips of a single driver with overlapping time stamps, which could be due to a system error (details in Appendix section A.1). In subsection 1.6.3, we elaborate on the construction process of all variables in our study.

1.4 Impact of Past Observations and Experiences on Relocation Decisions

In this section, we conduct reduced-form regression analysis to show that drivers’ previous observations and experiences affect their current relocation decisions. Consistent with the previous literature (Emadi & Swaminathan, 2018; Hathaway et al., 2022), this history-dependent decision-making can be driven by drivers’ learning about the earning potentials and search costs in different locations, which we formally examine in section 1.5 and show further evidence for through a model comparison in subsection 1.6.5.

Prior literature has documented that earnings as well as the supply and demand imbalance and search costs in different locations as perceived by drivers can influence their relocation decisions (Zhang et al., 2020; Karacaoglu et al., 2018; Jiang et al., 2020). This motivated us to incorporate two factors related to earnings (*tip percentage, base fare*), two factors associated with the supply and demand conditions (number of nearby *pickups, drop-offs*), and two factors related to drivers’ search costs (*speed, drivers’ wait time* between trips) as explanatory variables in our analysis.

Tip percentage is the percentage of tip over the base fare (excluding tip) of a trip originating from a specific location¹⁵. In low-wage service industries such as the taxi industry,

¹⁴<https://www.ncdc.noaa.gov/data-access>

¹⁵Only tips paid with credit cards are observable. To address the common challenge of missing values of cash tips (Alexander et al., 2021; Haggag & Paci, 2014), we adopt a machine learning approach to predict

customer tips account for a fair portion of service providers' total income (Alexander et al., 2021; Brewster & Mallinson, 2009). Taxi drivers often view tips as the “windfall money” (Karacaoglu et al., 2018), and a higher tip percentage leads to a larger compensation with the same fuel expenditures, vehicle depreciation, and driving time. *Fare income* captures the base fare amount of a trip originating from a specific location. The effect of fare income is nuanced. While some drivers may prefer to work fewer trips with larger fares to reduce search efforts, others may favor more trips with smaller fares to take advantage of large fixed upfront charges. In terms of the supply and demand conditions, we take into account the number of nearby *pickups* and *drop-offs* observed by drivers. They are measured as the number of pickups and drop-offs within 490 feet in spatial distance (approximately half-block length in Manhattan¹⁶) and 1 minute in temporal gap from each pickup and drop-off time point. The rationale for accounting for nearby pickups is that when drivers observe more nearby pickups in a specific location and time in previous trips, they may expect higher demand in the same location and time in the long run, or anticipate lower demand in that location in the near future because most of the demand has been satisfied. Similarly, more drop-offs could indicate higher competition in the same location and time in the long run as the vacant taxis become competitors of nearby taxis, or higher potential demand in the same location in the near future. For example, drop-offs at an arena at the beginning of a sports event indicate potential future pick-ups when the event ends. *Trip speed* is measured as the average speed during a trip, which is calculated as trip distance divided by trip duration. *Trip speed* captures the real-time traffic congestion level, which can affect the search cost in a particular location, although the direction is nuanced. In a heavily congested area, drivers may face steeper search costs due to limited mobility or lower search cost due to high pedestrian intensity. *Wait time* is measured as the time gap between a drop-off in a specific location and the subsequent pickup. A longer wait time indicates more difficulty in locating the next customer and thus a higher search cost.

the missing values of cash tips. Details about the machine learning approach are in Appendix section A.2.

¹⁶<https://bindleyhardwareco.com/what-is-one-block-distance>

1.4.1 Evidence for History-Dependent Relocation Behavior

Assuming that drivers' intrinsic preferences (e.g., their sensitivity to wait time) do not change over time, any changes in their decision making process can be attributed to changes in their beliefs about the aforementioned factors. Since drivers' beliefs about the distribution of these factors are shaped by their own unique history of trips, we expect drivers' relocation behavior to depend on their individual history. For example, a driver who received a high tip percentage from a passenger picked up in an area can be more likely to relocate to that area in future trips because the driver may anticipate passengers from the same area to have similar tipping habits. Yet, each realization of tips is a noisy signal of its true distribution. Therefore, given the variations in tip percentages within a location during a certain time window (which are demonstrated in Appendix section A.3), as the driver completes more trips and observes more realizations, his belief about the tip distribution will be updated, which is then likely to affect his subsequent relocation decisions.

Importantly, history-dependent relocation behavior does not necessarily indicate learning if drivers' heterogeneity is not controlled. For example, a driver may have intrinsic preferences for a certain area and thus is more likely to relocate to that area irrespective of his/her previous observations and experiences. If the area happens to be associated with generous tips, we may observe a positive correlation between the driver's relocation probability and the tip percentage previously earned in that area even in the absence of learning.

To control for drivers' heterogeneity, we use a two-step grouped-fixed effect (GFE) approach (Emadi & Swaminathan, 2018; Hahn & Moon, 2010; Bonhomme et al., 2017). In our context, it is computationally infeasible to estimate the fixed effects given hundreds of trips of each individual driver due to the curse of dimensionality¹⁷. Therefore, we implement two-step GFE as a dimension reduction tool. First, we classify drivers based on their location preferences developed before our analysis period between June and July. We

¹⁷The fixed-effects estimator is feasible with shorter panels where the number of repeated observations usually does not exceed 9 or 10 periods, depending on the number of alternatives, data size, and possibly other features of the data (<https://www.stata.com/manuals/xtxtmlogit.pdf>).

use June to July data for consistency of the reduced-form analysis with our structural estimation, where we use January to May data to preestimate drivers' prior beliefs, and June to July data to estimate the other parameters in the structural model (the details are explained in subsection 1.6.1). To capture drivers' location preferences, we focus on drivers' relocation behavior in their first shift between June and July, when the relocation decisions are mainly driven by drivers' intrinsic preferences and learning (if any) before June. Specifically, we first use K-means clustering to classify the drivers into groups based on their imputed relocation probabilities in different locations in the first shift¹⁸. We basically assume that drivers with similar patterns in relocation probabilities in the first shift have similar preferences. Hence, by controlling for drivers' cluster dummies, we control for driver preference heterogeneity. The drivers are classified into 26 groups, which is the optimal number of clusters suggested by the Wards' method with the gap procedure (Tibshirani et al., 2001; Teixeira et al., 2012; Corritore et al., 2020). Second, we include the driver cluster dummies as independent variables in our regressions to control for driver heterogeneity in location preferences. We exclude drivers' trips from the first shift in the analysis period to avoid the contamination of trips over which driver clusters are defined. Figure 1.1 illustrates the two biggest and two smallest clusters of drivers with respect to their average relocation probabilities towards each of the ten locations in the first shift. Different location preferences for drivers across different clusters are evident from the figure. For example, while the drivers in Cluster 3 (upper left panel) prefer Location 7, 5, and 8, the drivers in Cluster 16 (lower left panel) prefer Location 2 and 7.

To show that drivers' relocation decisions are indeed affected by their previous observations and experiences, we develop a multinomial logistic model (Guadagni & Little, 1983), that includes variables on drivers' cumulative experiences and observations, alternative-specific intercepts and temporal variables, and controls for drivers' location preferences developed before the analysis period. As such, we suppose that a driver i 's latent utility of

¹⁸A shift is identified as consecutive trips with a gap time duration no more than 5 hours (Haggag et al., 2017).

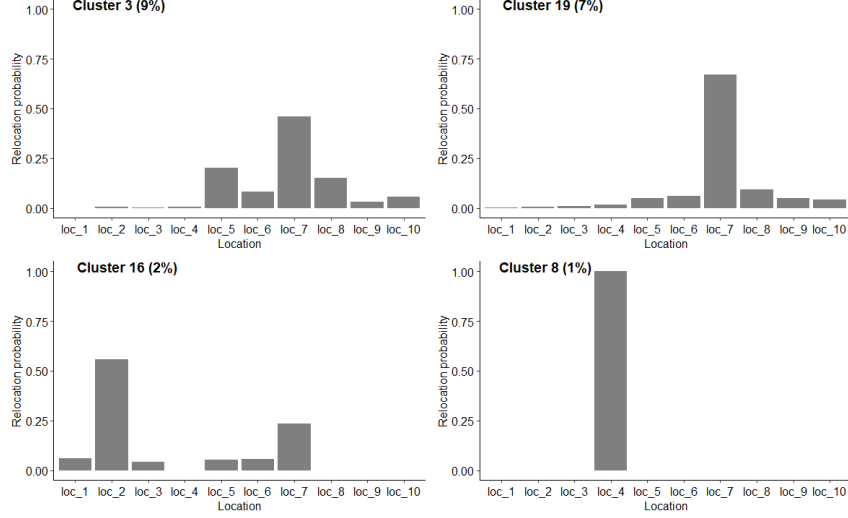


Figure 1.1: Relocation frequencies in the two biggest and two smallest clusters (the respective percentages of drivers in each cluster are given in parentheses)

relocating to location l at time t is

$$U_{ilt} = \alpha_l^r + \bar{X}_{il}\beta^r + Z_t\gamma_l^r + c_i\delta_l^r + \epsilon_{ilt}, \quad (1.1)$$

where i denotes driver, $l \in (1, \dots, 10)$ denotes 10 spatial PUMAs in Manhattan, $t \in T$ denotes the time when drivers make relocation decisions at the end of each trip, and \bar{X}_{il} is the moving average of driver i 's most recent observations at location l prior to the relocation decision at t , including tip percentage, fare income, average speed, number of nearby pickups and drop-offs, and wait time. The term \bar{X}_{il} is imputed as $\frac{1}{m} \times \sum_{\tau=t-m+1}^t X_{il\tau}$, which is the moving average over m periods prior to the relocation at t that measures driver i 's prior experience. The superscript r of the utility coefficients denotes that they are in the reduced-form regression, which are differentiated from those in the structural model to be introduced later. To control for unobserved location-specific confounding factors, we include the location-specific intercept α_l^r . The term Z_t is a vector that contains temporal variables including weekday indicator and rush hour indicator, c_i is drivers' cluster indicator, and ϵ_{ilt} is an observation-level error term following a type-I extreme value distribution.

Let y_{iwt} be a binary variable that indicates whether driver i chooses alternative w , $w \in \{0, 1, \dots, 10\}$ which includes 10 locations and the outside option indexed by $\{0\}$, at time t . For identifiability, we set the outside option as the reference outcome and restrict the respective utility coefficients to be 0, such that $U_{i0t} = \epsilon_{i0t}$. Given the distributional assumption for the error terms, the probability of choosing alternative w is given by

$$Pr(y_{iwt} = 1) = \frac{e^{\alpha_w^r + \bar{X}_{iw}\beta^r + Z_t\gamma_t^r + c_i\delta_t^r}}{1 + \sum_{s \in \{1, \dots, 10\}} e^{\alpha_s^r + \bar{X}_{is}\beta^r + Z_t\gamma_s^r + c_i\delta_s^r}}. \quad (1.2)$$

Table 1.2 shows the logistics regression analysis results. In the full model (Column (1)), we include all six factors that may affect drivers' relocation decisions and assume that moving averages of past observations are calculated over $m = 10$ periods (we report consistent results in the robustness analysis based on $m = 5$ periods in Appendix section A.4).

Overall, the results show that drivers' previous observations and experiences affect their current relocation decisions. Importantly, such history-dependent relocation behavior is significant after controlling for drivers' heterogeneity in location-preference cluster dummies. Specifically, as indicated by the positive coefficients of tip percentage (3.843, significant at 1%) and fare (0.026, significant at 1%), drivers are more likely to relocate to locations wherein they received higher tip percentages and base fare in most recent trips. The negative coefficient of speed (-0.019, significant at 1%) suggests that drivers also tend to relocate to locations wherein the speed is low, which can indicate greater pedestrian density and potential demand. While the coefficient for the number of nearby pickups (-0.666, significant at 1%) is negative, it is positive for the number of drop-offs (0.264, significant at 1%), consistent with our expectations about the short-term association. The negative coefficient of wait time (-0.001, significant at 1%) shows that drivers are less likely to move to locations with longer wait time, which can signal intensified competition and lower demand.

In addition, we demonstrate the importance of tip percentage as an explanatory factor by

Table 1.2: Multinomial logit regression results

Dependent variable: relocation decisions	Full model (1)	Partial model (2)
Moving average of tip percent	3.843*** (0.093)	
Moving average of fare	0.026*** (0.001)	0.023*** (0.001)
Moving average of speed	-0.019*** (0.001)	-0.034*** (0.001)
Moving average of number of pickups	-0.666*** (0.004)	-0.681*** (0.004)
Moving average of number of dropoffs	0.264*** (0.003)	0.242*** (0.003)
Moving average of wait time	-0.001*** (0.000)	-0.001*** (0.000)
Location-specific intercepts	Included	Included
Weekday indicator	Included	Included
Rush hour indicator	Included	Included
Driver cluster indicator	Included	Included
-LL	682,308	683,126
AIC	1,365,209	1,366,843
BIC	1,368,448	1,370,072
Likelihood ratio test Chisq		1,636***

Note: Independent variables are the moving averages calculated over 10 periods prior to the relocation. Standard errors are in parentheses. *** indicates significance at 1%.

comparing the model fit of the full model (Column (1)) with a nested model that excludes the tip percentage (Column (2)). The coefficients of these models are consistent. Compared to the nested model, the full model has lower AIC and BIC, indicating a better model fit. The significance of the Likelihood Ratio Test statistic (1636, significant at 1% level) supports that tip percentage is an important factor on drivers' relocation decisions and should be included in the model.

As a robustness analysis, we perform the above analyses based on the new drivers who did not drive between January and March, but started to drive after March. As such, we are able to identify these drivers' first shift in their lifetime, when the relocation decisions are mainly driven by intrinsic preferences. We find a similar history-dependence pattern in new drivers' relocation behavior, and the results are reported in Appendix section A.4.

1.4.2 Motivation for the Structural Model

Although the multinomial logistic regression results establish the dependence of drivers' current relocation decisions on their past observations and experiences, it does not model the behavioral process through which such dependence materializes. Specifically, the reduced-form regression does not separate the effect of drivers' intrinsic preferences such as their sensitivity to expected tip percentage from that of their prior beliefs and the learning process. Such a separation not only provides a deeper understanding of how the drivers' behavior evolves, but also enables us to examine the impact of hypothetical interventions on the tipping policy. For example, different tip information disclosure policies facilitated by new technologies may affect drivers' abilities to learn about tips, but they should not change drivers' intrinsic preferences for tips. Using reduced-form analysis to make predictions is problematic because the estimated coefficients of the regressors in Table 1.2 are endogenous to the current policy and thus may change under counterfactual tipping policies (Erdem & Keane, 1996; Y. Huang et al., 2014). Therefore, we develop a structural model of drivers' relocation decision-making process, which enables us to estimate the underlying

parameters of drivers that are invariant to tipping policies.

1.5 Structural Model

In our structural modeling framework, we assume that after each drop-off, drivers decide where to search for the next passengers. Such decisions are governed by drivers' expected utility of relocating to different locations at different times, which can be approximated by an additive multi-attribute utility model (Erdem & Keane, 1996; Sriram et al., 2015). To form the expected utility, drivers rely on their beliefs about the averages of attributes associated with a location-time set including the tip percentage, fare income, speed, number of nearby pickups and drop-offs, and wait time, which reflect the earning potential and the inconvenience of operating in that specific location-time set. As drivers complete more trips, they accumulate personal experiences and observations which provide new information regarding the attribute levels in different locations and times. Hence, drivers update their beliefs with the new information and use the updated beliefs to guide their subsequent relocation decisions. We model drivers' belief updating process in a Bayesian manner, consistent with Erdem and Keane (1996) and Y. Huang et al. (2014) and Zhang et al. (2020).

In this section, we first introduce the model preliminaries. We then lay out a Bayesian framework for drivers' learning of earning potentials and search costs across locations and times. Finally, we embed the Bayesian framework into a discrete choice model to capture drivers' relocation decisions.

1.5.1 Preliminaries

Suppose that drivers are indexed by $i \in N$, drivers' potential relocation choices are indexed by location-time set $j \in J$, which is a combination of both spatial and temporal dimensions including 10 spatial PUMAs in Manhattan and 3 time slots in a day, namely midnight (1-6 a.m.), rush hours (7-10 a.m. and 4-8 p.m.), and non-rush hours (6-7 a.m., 10 a.m.-4 p.m., and 8p.m.-1 a.m.). Hence, J contains 30 location-time sets. The term $t \in T$ denotes the

time when drivers make relocation decisions (i.e., the drop-off time at the end of each trip).

1.5.2 Bayesian Learning Process

In the data, we observe significant heterogeneity across location-time sets in terms of the average levels of the six attributes of interest. Moreover, within a location-time set, temporal variations exist in attribute realizations (details shown in Appendix section A.3). Given these spatial and temporal variations and following the standard practice in the Bayesian learning literature (Erdem & Keane, 1996; Sriram et al., 2015), we consider a Normal distribution¹⁹ for attribute realizations. Let A_{ijt} denote the attribute realizations²⁰ observed or experienced by driver i in the pickup and drop-off location-time set j 's²¹ associated with the trip that ended at t . We assume A_{ijt} follows a Normal distribution

$$A_{ijt} \sim N(\mu_{A_j}, \sigma_{A_j}^2), \quad (1.3)$$

where μ_{A_j} is the average of the attribute realizations in location-time set j , and the variance $\sigma_{A_j}^2$ captures temporal variations of the attribute realizations in location-time set j .

We assume that drivers are uncertain about the mean of the attribute levels (i.e., μ_{A_j}), but know the variance (i.e., $\sigma_{A_j}^2$), which is a standard assumption in the Bayesian learning literature (Zhang et al., 2020; Erdem & Keane, 1996; Y. Huang et al., 2014; Sriram et al., 2015). Hence, the attribute realizations inform drivers about the mean attribute levels of different location-time sets. Drivers then combine these attribute realizations and their prior beliefs to construct posterior beliefs.

¹⁹Among the six attributes, tip percentage, fare amount, and speed are approximately normally distributed, but the number of nearby pickups, drop-offs, and wait time are nonnormal due to the concentration of mass at zero. Since it is difficult to implement Bayesian updating rules if some attributes are nonnormal, we apply the Box-Cox transformation to bring all attributes closer to normality, consistent with Erdem et al. (2008) (details are described in Appendix section A.5).

²⁰For ease of notation, A_{ijt} is further decomposed into $A_{tip_{ijt}}$, $A_{fare_{ijt}}$, $A_{speed_{ijt}}$, $A_{pick_{ijt}}$, $A_{drop_{ijt}}$, $A_{wait_{ijt}}$.

²¹While observations along a driver's trajectory can potentially influence driver decisions as well, we focus exclusively on pickups and drop-offs due to the unavailability of trajectory data. We expect observations at pickup and drop-off locations to be more salient to the drivers' decision-making due to their ability to act on them directly.

Let B_{ij0} denote driver i 's prior beliefs about the mean attribute levels (i.e., μ_{A_j}) in the location-time set j at the beginning of the estimation period, which is given by

$$B_{ij0} \sim N(\mu_{B_0}, \sigma_{B_0}^2), \quad (1.4)$$

where μ_{B_0} is the mean of the prior belief, and $\sigma_{B_0}^2$ is the variance of the prior belief. The distribution of the prior belief is assumed to be the same across all location-time sets j for computational tractability (Zhang et al., 2020; Erdem & Keane, 1996; Y. Huang et al., 2014).

Let B_{ijt} denote a driver's posterior belief about the average attribute levels (i.e., μ_{A_j}) in location-time j at time t , $t \geq 1$:

$$B_{ijt} \sim N(\mu_{B_{ijt}}, \sigma_{B_{ijt}}^2), \quad (1.5)$$

where $\mu_{B_{ijt}}$ is the mean of the posterior belief, and $\sigma_{B_{ijt}}^2$ is the variance of the posterior belief. At the end of each trip, conditional on the cumulative information received during the trip, the driver updates his beliefs associated with both the pick-up and drop-off location-time sets in a Bayesian manner. Because prior beliefs and attribute realizations are assumed to be Normally distributed, given the conjugacy of the prior belief and attribute realizations, the posterior mean is given by

$$\mu_{B_{ijt}} = \mu_{B_{ij,t-1}} + (A_{ijt} - \mu_{B_{ij,t-1}}) \times \frac{\sigma_{B_{ij,t-1}}^2}{\sigma_{B_{ij,t-1}}^2 + \sigma_{A_j}^2}, \quad (1.6)$$

where A_{ijt} is the attribute realization experienced or observed by driver i in location-time set j during the trip ending at time t , and the posterior variance is

$$\sigma_{B_{ijt}}^2 = \frac{1}{\frac{1}{\sigma_{B_{ij,t-1}}^2} + \frac{d_{A_{ijt}}}{\sigma_{A_j}^2}} = \frac{1}{\frac{1}{\sigma_{B_0}^2} + \frac{\sum_{k=1}^t d_{A_{ijk}}}{\sigma_{A_j}^2}}, \quad (1.7)$$

where $d_{A_{ijt}}$ is a binary indicator which equals 1 if driver i has experienced or observed a certain attribute associated with location-time set j (either at pick-up or drop-off) during a trip that ended at time t , and is 0 otherwise. As such, the indicator variable captures each driver's unique relocation trajectory up to time t .

Equation 1.6 implies that a driver's posterior mean at t is a weighted sum of his last period's mean belief at $t - 1$ (i.e., $\mu_{B_{ij,t-1}}$) and the information received from the attribute realizations at t , which is its deviation from the last period's mean belief (i.e., $A_{ijt} - \mu_{B_{ij,t-1}}$). The driver's learning rate in Equation 1.6, $\frac{\sigma_{B_{ij,t-1}}^2}{\sigma_{B_{ij,t-1}}^2 + \sigma_{A_j}^2}$, is higher if the attribute is inherently less variable (i.e., smaller $\sigma_{A_j}^2$), or if the driver has greater uncertainty about the mean attribute level (i.e., greater $\sigma_{B_{ij,t-1}}^2$) especially in early stages. As the driver accumulates knowledge from personal experiences and observations over time (i.e., greater $\sum_{k=1}^t d_{A_{ijk}}$), his learning rate decreases due to reduced uncertainty, as captured by Equation 1.7. Taken together, the updating process characterized by Equation 1.6 and Equation 1.7 implies that drivers' posterior belief distribution about the average attribute level converges to a distribution with mean equal to μ_{A_j} and variance equal to zero.

Let $Info(i, t)$ denote the information about the attribute levels received by driver i during the Bayesian learning process. Let $A_{ijt}^{pos} | Info(i, t)$ denote driver i 's posterior predictive belief about future attribute realizations in location-time set j given the information received by driver i through his observations and experiences up to time t , and it is given by (Gelman et al., 1995)

$$A_{ijt}^{pos} | Info(i, t) \sim N(\mu_{B_{ijt}}, \sigma_{B_{ijt}}^2 + \sigma_{A_j}^2). \quad (1.8)$$

Here, $\sigma_{B_{ijt}}^2 + \sigma_{A_j}^2$ contains two sources of variations: (1) The uncertainty in the driver's belief about the average level of the attributes in the location-time set j at t , which is assumed to be distributed according to $N(0, \sigma_{B_{ijt}}^2)$. (2) Inherent temporal variations of attribute realizations in the location-time set j , which is assumed to be distributed according

to $N(0, \sigma_{A_j}^2)$. Over time, as the driver completes more trips and accumulates experience, the mean $\mu_{B_{ijt}}$ converges to the true average of the attribute realizations μ_{A_j} , the uncertainty in the driver's belief $\sigma_{B_{ijt}}^2$ converges to 0, but the temporal variation $\sigma_{A_j}^2$ persists.

1.5.3 Relocation Decision Making

Given the spatial and temporal variations in attribute realizations, when a driver makes a relocation decision at the end of a trip at time t , he is likely to be uncertain about the attribute realizations in the next trip. Therefore, the driver's decision is based on his expected utility from relocating to a location-time set. Let $E(U_{ijt}|Info(i, t))$ be the expected utility of driver i if he chooses to relocate to j at time t , conditional on the information $Info(i, t)$ that driver i has up to time t , and it is given by:

$$E(U_{ijt}|Info(i, t)) = \alpha E(A_{ijt}^{pos}|Info(i, t)) + \beta T_t + \gamma Loc_l + \epsilon_{ijt}, \quad (1.9)$$

where $E(A_{ijt}^{pos}|Info(i, t)) = \mu_{B_{ijt}}$ is the mean of the posterior predictive distribution of the attribute realizations in relocation choice j at time t conditional on the information that driver i has up to t , T_t is a vector of temporal factors including a dummy variable indicating weekday or weekends, daily precipitation volume, and an ordinal variable of fog and haze intensity, Loc_l is the location-specific dummy variable, and ϵ_{ijt} is the idiosyncratic shock, independent and identically distributed and follows the Type I Extreme Value distribution. The term ϵ_{ijt} appears in the expected utility function because drivers' utility function is stochastic from the researcher's point of view.

Let Y_{ikt} be a binary variable indicating whether driver i chooses alternative $k \in \{0, 1, \dots, 30\}$, which includes 30 location-time sets and the outside option (i.e., end the shift) indexed by $\{0\}$, at time t . As is typical in discrete choice models, we normalize the nominal utility of the outside option, which is the deterministic part of the expected utility that does not depend on ϵ_{ijt} , to be zero. Given the distributional assumption for the error terms, the

probability that driver i chooses alternative k after a drop-off at time t has the multinomial logistic form and is given by

$$Pr(Y_{ikt} = 1) = \frac{e^{\alpha\mu_{B_{ijt}} + \beta T_i + \gamma Loc_l}}{1 + \sum_{j \in J_{sub}} e^{\alpha\mu_{B_{ijt}} + \beta T_i + \gamma Loc_l}}, \quad (1.10)$$

where J_{sub} is driver i 's choice set at t , which is a subset of J and contains only the location-time sets that are reachable by driver i at time t . The construction details of drivers' choice set and relocation decision variables are further explained in subsection 1.6.3.

1.6 Estimation

We estimate the structural model using observations in our combined dataset as described in subsection 1.3.2. In subsection 1.6.1, we explain how the model is estimated. In subsection 1.6.2, we describe the identification strategies. In subsection 1.6.3, we describe the construction of variables such as drivers' relocation decisions and choice set. In subsection 1.6.4, we present the estimation results.

1.6.1 Estimation Procedure

We estimate our model in two stages. First, recall from our Bayesian learning framework that drivers know the variance of attribute realizations ($\sigma_{A_j}^2$). Thus, in Stage 1, we estimate the variance of attribute realizations in different location-time sets from the data. In Stage 2, given the attribute variances estimated in Stage 1, we estimate the structural model parameters including the utility weight coefficients (α , β , and γ), and drivers' prior mean (μ_{B_0}) and variance ($\sigma_{B_0}^2$) using the maximum likelihood approach. We maximize the likelihood of observing driver i 's decision outcome at time t , which is given by

$$L_{it} = \prod_{k \in \{0,1,\dots,30\}} Pr(Y_{ikt} = 1)^{Y_{ikt}}. \quad (1.11)$$

In our estimation, we focus on full-time experienced drivers, who account for the majority of the driver population in the data. Specifically, we identify 21,291 full-time experienced drivers as those who completed at least 300 trips between January and March, and at least 400 trips between April and July (corresponding to 93% of the driver population). For tractability, our estimation is based on a randomly sampled 1,044 (about 5%) experienced drivers and their respective trips from April 1st to July 31st (1,065,591 trips in total).

Since the drivers may have accumulated some experiences depending on their unique trajectories, their prior beliefs about the attribute distributions can be different both across location-time sets and among drivers. To capture the drivers' heterogeneous prior beliefs, we conduct the Stage-2 estimation in two steps, similar to (Zhang et al., 2020; Erdem & Keane, 1996; Shin et al., 2012). We (1) pre-estimate the model with a common mean (μ_{B_0}) and variance ($\sigma_{B_0}^2$) based on the trips between April and May, and then (2) impute each driver's posterior mean and variance at the end of May as their priors at the beginning of June, and estimate the rest of the structural model using June-July data.

1.6.2 Identification

In this section, we describe the identification of the structural parameters, including the utility weight coefficients (α , β , and γ), and drivers' prior mean (μ_{B_0}) and variance ($\sigma_{B_0}^2$).

In Equation 1.10, the utility weight coefficients (α , β , and γ) are identified by the variation in the relocation behavior of drivers across different time periods in their later shifts. Specifically, the intrinsic preferences for attributes (α) can be identified based on drivers' steady-state behavior. Consider a subset of drivers who have accumulated experiences over a sufficiently long time period. From Equation 1.8, their posterior belief mean ($\mu_{B_{ijt}}$) evolves to the actual average level of the attribute (μ_{A_j}) in their late shifts. As mentioned, we directly observe attribute realizations and thus can identify their distribution parameters (μ_{A_j} and $\sigma_{A_j}^2$) from the data. Therefore, for these experienced drivers in their late shifts,

we can write Equation 1.10 as $Pr(Y_{ikt} = 1) = \frac{e^{\alpha\mu_{A_j} + \beta T_t + \gamma Loc_l}}{1 + \sum_{j \in J_{sub}} e^{\alpha\mu_{A_j} + \beta T_t + \gamma Loc_l}}$. Since drivers'

relocation decisions on the left-hand side of this equation are observed, and the average of attribute realizations μ_{A_j} , temporal variables T_t , and location-specific dummies Loc_l are identified or observed from the data, the variations in the left-hand side would identify the respective utility coefficients α , β , and γ . Intuitively, the identification process is similar to regressing drivers' relocation decisions on observed μ_{A_j} , T_t , Loc_l , which would give us the slope estimators.

Next, we discuss the identification of the parameters of drivers' prior belief distribution. From Equation 1.6 and Equation 1.7, we can precisely characterize the evolution of the posterior belief conditional on the prior mean μ_{B_0} and variance $\sigma_{B_0}^2$. From Equation 1.10, drivers' relocation behavior is driven by the prior mean μ_{B_0} and variance $\sigma_{B_0}^2$, beyond the utility coefficients and attribute realizations which are identified. Given the identification of the utility coefficients and the distribution parameters of attribute realizations (μ_{A_j} and $\sigma_{A_j}^2$), we can further identify the prior belief through the variations in drivers' relocation behavior after being exposed to the attribute realizations. Specifically, drivers' prior belief mean (μ_{B_0}) is identified through the direction of change in population-level average behaviors before and after being exposed to attribute realizations. For example, for a particular attribute, after receiving the actual realizations, if drivers' relocation probabilities increase (decrease) on average, it implies that their initial belief about mean attribute level is pessimistic (optimistic). Drivers' initial belief variance ($\sigma_{B_0}^2$) is identified through the variation in population-level average behaviors before and after being exposed to the attribute realizations. If drivers' behaviors do not change significantly before and after receiving the realizations, then it implies that the drivers' initial belief is quite precise, and the prior belief variance should be low.

1.6.3 Data Processing

In this section, we explain how we process the data to operationalize the variables in the structural model, including location-time sets (j), timing of updating ($d_{A_{ijk}}$), temporal (T_t)

and spatial variables (Loc_t), and relocation decision (Y_{ikt}).

Location-time Set

According to our model in subsection 1.5.2, drivers update their beliefs and make relocation decisions at the location-time level. As mentioned, the Manhattan borough is divided into 10 PUMAs corresponding to the conventional views of NYC neighborhoods. We further divide 24 hours in a day into 3 time slots including midnight (1-6 a.m.), rush hours (7-10 a.m. and 4-8 p.m.), and non-rush hours (6-7 a.m., 10 a.m.-4 p.m., and 8p.m.-1 a.m.). For tractability, we control for the effect of weekday or weekend in the utility function by adding weekday dummies (further explained in subsection 1.6.3), instead of distinguishing them in our definition of time slots. In total, there are 30 location-time sets.

Timing of Updating

For *tip percentage*, *fare income*, *trip speed*, *pickup*, and *drop-off*, drivers update their beliefs about the averages of attribute values related to both the pickup and drop-off j 's of a trip after they complete the trip at t . In contrast, each driver associates the *wait time* incurred after the completion of a trip with the drop-off location-time set of that trip, and hence, drivers update their beliefs about the average *wait times* only in the drop-off location-time sets. Moreover, because drivers' wait time after the current drop-off will not be realized until the next pickup occurs, we assume that when drivers make relocation decisions at the end of the current trip at t , their observations on the wait time in j are from previous trips that ended by $t - 1$.

Temporal and Spatial Variables

To control for confounding variables that may affect drivers' relocation decisions, three temporal variables (*weekday/weekend indicator*, *daily precipitation*, and *fog/haze indicator*) and one spatial variable (location-specific dummies) are included in the utility function.

Specifically, the weekday indicator is included since the demand pattern can be different for weekdays and weekends, and the variables on the daily precipitation, fog, and haze are included to capture the daily shocks to demand in the utility function. By including location-specific dummies, we control for drivers' intrinsic preferences for certain locations.

Relocation Decision Variables

We capture drivers' relocation decisions by identifying the subsequent pickup location-time set within a pre-specified time window after the last completed trip. Specifically, if the subsequent pickup occurs within 1 minute (corresponding to the 10th percentile of the wait time distribution) after a drop-off, the driver may have picked up the next passenger shortly after a drop-off without deliberate search. If the subsequent pickup occurs beyond the 32-minute threshold (corresponding to the 90th percentile of the wait time distribution) but within 300 minutes after a drop-off, the driver may have chosen to take a short break. If the subsequent pickup occurs beyond the 300-minute threshold (about 9% of the trips), the driver may have ended the shift. We treat this as the driver choosing an outside option of not working. Therefore, only if the subsequent pickup occurs within the window of (1, 32) minutes or beyond 300-minute threshold, it is treated as a valid relocation decision (about 86% of the trips).

To construct a driver's choice set, we identify his reachable location-time sets after a drop-off. For example, if a driver dropped off a passenger at Midtown South at 5 p.m. on January 15th, 2009, his choice set at that time only includes those location-time sets reachable within a reasonable search time window after 5 p.m. Specifically, we set the lower bound of the search time window to be the minimum traveling time between the two locations observed from the data. The upper bound of the search time window is set to be 32 minutes, corresponding to the 90th percentile of the distribution of time gap between subsequent trips.

1.6.4 Estimation Results

Table 1.3 presents the estimates of the parameters in the utility function, including drivers' intrinsic preferences for attributes and temporal variables. We use these estimates to establish the consistency of our findings with the previous literature as well as to generate new insights. The coefficient of tip is significantly positive (5.645, significant at 1% level), indicating that drivers prefer to relocate to location-time sets wherein they expect to receive higher tip percentages. Additionally, drivers prefer to relocate to location-time sets wherein they expect to receive higher fare income, indicated by the significant positive coefficient of fare (0.186, significant at 1% level).

Table 1.3: Estimates of the utility function parameters (as described in subsection 1.5.3)

	Utility function weight
α _tip percent	5.645*** (0.318)
α _fare	0.186*** (0.065)
α _speed	-1.831*** (0.02)
α _number of pickups	2.794*** (0.031)
α _number of dropoffs	-0.398*** (0.023)
α _wait time	-0.452*** (0.014)
β _weekday	-0.094*** (0.012)
β _rain	0.015 (0.011)
β _fog	0.033*** (0.008)
Log-likelihood	846,400

Note: Standard errors are in parentheses. *** indicates significance at 1%. The coefficients for location-specific dummy variables are omitted here for brevity and shown in Table A.4.

Comparing the magnitude of coefficients of tip and fare, we find that drivers' extrinsic preference for tips is greater than that for fare. For example, given one dollar increase in the tip amount in Location 8²² (corresponding to an increase in tip percentage from 19.1%

²²Location 8 is chosen because the average tip percentage in this location is 19.3%, closest to the mean of location-specific tip percentage of 19.1%.

to 30.8%), the odds of relocating to Location 8 relative to choosing the outside option increases by about 40%. However, one dollar increase in the fare amount (from the average fare amount of \$8.5 to \$9.5) only increases the odds by 0.4% (calculation details are in Appendix section A.6)²³. A possible explanation is that higher tip percentages imply larger compensation with the same costs including fuel expenditures, vehicle depreciation, and driving time. However, a higher fare income is accompanied with proportional increases in the aforementioned costs. Therefore, the results suggest that the source of income (tips vs. fare) is an important leading factor in shaping service providers' relocation decisions.

The negative coefficient of speed (-1.831, significant at 1%) indicates that drivers prefer to relocate to location-time sets wherein they expect greater pedestrian density and potential demand, which coincides with a lower speed. The positive coefficient of pickups (2.794, significant at 1%) suggests that, drivers perceive the number of pickups as an indication of potential demand in the long run, and thus are more likely to relocate to location-time sets wherein higher number of pickups are expected. On the other hand, drivers are less likely to relocate to location-time sets wherein higher drop-offs are expected, indicated by the negative coefficient of drop-offs (-0.398, significant at 1%). This can be explained by drivers' expectations of higher competition among vacant taxis following a large number of drop-offs. Finally, the negative coefficient of wait time (-0.452, significant at 1%) shows that drivers are less likely to relocate to locations wherein longer wait times are expected, which usually indicates a lower demand and a higher search cost.

The estimated coefficients for location-specific dummy variables and drivers' prior belief parameters are shown in Appendix section A.7. Consistent with our expectation, drivers prefer to relocate to more popular areas with higher earning opportunities such as the business districts in lower Manhattan, but are less likely to relocate to the residential areas in upper Manhattan which can have lower earning opportunities.

²³We verify the contrast of the effect size by conducting two policy simulations comparing one-dollar increase in tip amount versus one-dollar increase in fare amount in Location 8. The simulation results show that increasing \$1 in tip amount increases the relocation probability by 20%, and increasing \$1 in fare amount increases the relocation probability by 0.1%. The simulation details are available upon request.

1.6.5 Comparison of Model Fit

To illustrate the importance of modeling driver learning, we compare the model fit of the learning model (described in section 1.5 with a rational expectations model (implementation details are in Appendix section A.8). The latter disregards the learning process, assuming that all drivers possess common knowledge about the true distributions of every attribute. As demonstrated in Table 1.4, both the in-sample and out-of-sample fits reveal that the learning model substantially outperforms the rational expectations model with respect to AIC and BIC values. The model comparison further supports the learning behavior in our data. In particular, drivers’ beliefs about attribute means evolve in line with a Bayesian learning manner, rather than conforming to the actual attribute distribution.

Table 1.4: Model comparison table

	In-sample fit			Out-of-sample fit		
	-LL	AIC	BIC	-LL	AIC	BIC
Rational expectations model	603,489	1,207,013	1,207,207	254,501	509,037	509,216
Bayesian learning model	596,123	1,192,282	1,192,477	249,965	499,965	500,144

1.7 Policy Simulations

Tips have been widely adopted in service industries to subsidize workers through earning additional income (Donovan, 2016; Rosenbloom, 2016). However, anecdotal evidence has also shown that tips can backfire and induce inefficiencies by incentivizing discretionary service providers to cherry-pick where to serve (Matarese, 2021). In view of such a trade-off, should app-based taxi-hailing companies adopt the in-app tipping option or consider alternative compensation mechanisms such as implementing a default fixed service charge rate? If tipping is adopted, how should companies manage tip-related information sharing with service providers? For example, should companies disclose the tip amount immediately after it is available, or should they withhold such information?

To shed light on these questions, we conduct policy simulations on how different tipping

policies influence key performance outcomes, including drivers' hourly income and system efficiency, by accounting for drivers' learning about tips in their relocation decision-making processes. Below, we first introduce different tipping policies observed in practice and how we implement them in the simulation. We then provide a detailed description of the procedures used for simulating the system dynamics. Finally, we describe the performance measures and discuss the results.

1.7.1 Tipping Policies

In practice, a variety of tipping policies that vary the salience of tips in drivers' learning processes have been adopted. For example, Curb²⁴ and Doordash show a detailed breakdown of earnings (including the base pay, promotions if any, and customer tip) as soon as a delivery is completed (Xu, 2019). In contrast, Postmates withhold information about tips by allowing drivers to review tip earnings corresponding to each past delivery (without the customer address) at least 24 hours after an order is completed (Postmates, 2021). In 2016, Instacart announced plans to replace tips with an optional service charge and to be distributed evenly among the entire fleet of drivers to create "more consistent pay with fewer variables". Providing detailed information about tips in a timely fashion may enable drivers to spontaneously relate their earnings (including tips) to different pickup and drop-off conditions and become more selective in deciding where and/or who to serve. In contrast, withholding tip-related information makes it rather difficult, if not impossible, for drivers to relate the tip to a particular address, because customer addresses are not accessible for later review due to privacy considerations²⁵. Therefore, withholding tip information may attenuate or hinder drivers' learning about tips in different location-time sets. Consistently, replacing tips with a service charge (a fixed percentage of the base fare paid by customers

²⁴https://assets-global.website-files.com/611e89c803b9cf81b119072e/629a328bb054685a7a6bc489_ChicagoDriver2021June21_compressed.pdf

²⁵"Once an order is delivered, the address is removed from the driver's app." as noted in the link: <https://www.shopfood.com/online-shopping/what-personal-information-do-food-delivery-drivers-see-about-customers/>

to the drivers) is expected to remove tip learning completely by eliminating variations in tips.

In this section, we simulate drivers' relocation decisions under different tipping policies representative of those mentioned above and thus determine how these policies would affect drivers' income and system efficiency. Our simulations build on drivers' underlying utility weights and prior beliefs about the attributes estimated in the previous section.

To approximate the scenario where tip details are provided immediately after they are available (which we call *the baseline policy*), we start all drivers with their initial beliefs about the six attributes and allow them to update their beliefs after each trip in the Bayesian fashion as described in subsection 1.5.2. In the alternative scenario wherein tip information sharing is withheld (which we call *the hinder learning policy*), we start all drivers with their initial beliefs about the six attributes, set drivers' learning rate of the tip (i.e., $\frac{\sigma_{B_{ij,t-1}}^2}{\sigma_{B_{ij,t-1}}^2 + \sigma_{A_j}^2}$) to be 0, but allow them to learn about the rest of the five attributes. That is, drivers update their beliefs about all attributes except the tip. Finally, in the scenario where tip is replaced with a fixed service charge (which we call *the service charge policy*), we start all drivers with a belief of a fixed service charge percentage (as a substitute of their tip prior belief), which is equivalent to the average tip percentage received by drivers under the baseline policy, and their initial beliefs about the other five attributes. Drivers' beliefs about the service charge remain fixed but their beliefs about the other five attributes are updated in the Bayesian fashion over the entire simulation period. We expect that the differences in the evolution of drivers' beliefs about tips will lead to different relocation decisions and will ultimately be reflected in drivers' income and system efficiency.

1.7.2 Simulation Details

Below, we explain how we set up the system, generate the attribute realizations, and measure key performance outcomes. More details on the simulation (including the generation of drivers' prior beliefs, attribute realizations and variances, and relocation decisions) are

in Appendix section A.9.

System Setup

We simulate the system as a series of discrete pickup and drop-off events conducted by 1,044 drivers over 62 trips, corresponding to the actual number of drivers in the data and the average number of trips completed in a 7-day week²⁶, respectively. We draw drivers' first pickup and drop-off location-time sets ($j \in J$) and timestamps ($t \in T$) from the frequency distribution of pickup and drop-off pairs corresponding to the first trips in a shift in the data.

Attribute Realizations

At the end of each trip and before relocation, drivers update their beliefs using the attribute realizations observed or experienced at the pickup and drop-off location-time sets. To generate attribute realizations in the simulated system, we follow the procedures in Figure 1.2 and explain them as follows. The red area indicates observed data and the blue area indicates the simulated system. We first categorize the attributes to be generated in the simulated system into two groups: customer-driven attributes (i.e., tip percentage and fare in block 5 in the green area) and system-driven attributes (i.e., speed, wait time in block 6 and the number of nearby pickups and drop-offs in block 7 in the yellow area). For the customer-driven attributes, we draw the attribute realizations from the fitted empirical distribution of tip and fare in each location-time set as observed in the data (block 1 \rightarrow 2 \rightarrow 5)²⁷.

For the system-driven attributes, a key challenge is that they are endogenous to drivers' relocation decisions in the simulated system. Therefore, we cannot simply rely on the

²⁶The simulation horizon is set as a 7-day week window due to computational tractability.

²⁷We assume that the customer preferences for locations to travel (which determine the fare amount) will not change over the course of our simulation, which covers a 7-day week window. Customer tips are assumed to be exogenous based on the empirical evidence that tipping outcomes reflect the personal characteristics of the customers more so than those of the service providers (Chandar et al., 2019).

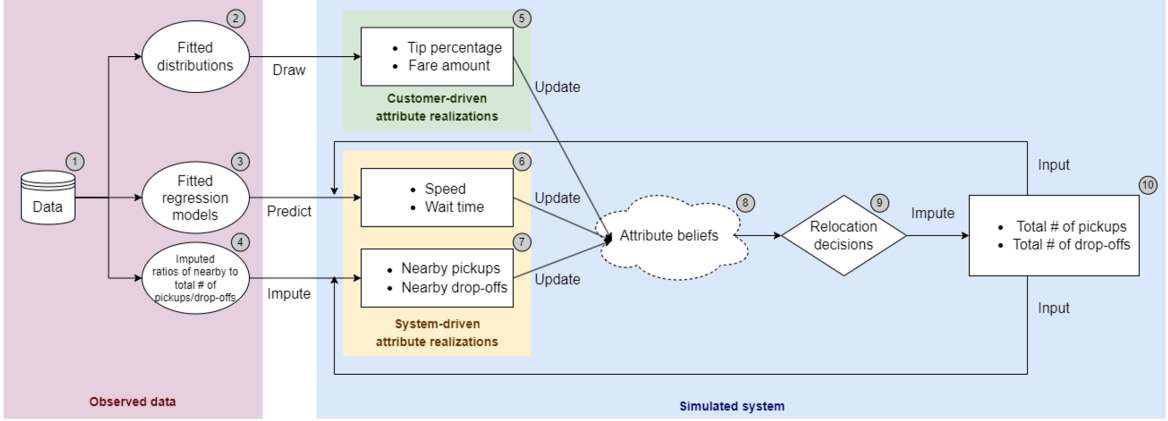


Figure 1.2: The generation process of attribute realizations

empirical distribution observed in the data, but we need to build a feedback loop such that these attributes reflect the simulated supply and demand conditions. To achieve this, for speed and wait time attributes (block 6), we model attribute realizations as a function of supply conditions, demand conditions (measured by the total number of pickups and drop-offs), location and time dummies. We then estimate the model coefficients based on the observed data (block 1 \rightarrow 3). Next, we predict the attribute realizations in the simulation with inputs from the simulated system (i.e., simulated total number of pickups and dropoffs in the blue area) (block (3 & 10) \rightarrow 6).

For the number of nearby pickup and dropoff attributes (block 7), we assume that they are proportional to the total number of pickups and dropoffs in the location. We first impute the ratios of the number of nearby pickups to the total number of pickups in each location from the data (block 1 \rightarrow 4), and then impute the simulated number of nearby pickups by multiplying the ratio with the inputs from the simulated system (i.e., simulated total number of pickups in the blue area) (block (4 & 10) \rightarrow 7). The drop-off attribute is generated in a similar manner. More details about the generation of attribute realizations are in Appendix section A.9.

With the simulated attribute realizations, drivers update their attribute beliefs and make relocation decisions as in subsection 1.5.3 (block (5 & 6 & 7) \rightarrow 8 \rightarrow 9). Based on drivers' simulated relocation decisions, we impute the total number of simulated pickups

and dropoffs as the input to generate system-driven attribute realizations, which complete the feedback loops (block 9 → 10 → (6 or 7) → 8 → 9).

Outcome Measures

We construct the following measures to determine the impact of the three tipping policies on drivers' welfare and system efficiency:

- *Average hourly tip income* We calculate each driver's hourly tip income as his tip income received across all trips divided by the total driving time (including the trip duration and wait time after the drop-off) across all trips. We then average the hourly tip income across all drivers.
- *Average hourly total income* This is calculated in a similar way to average hourly tip income, but includes drivers' income from both base fare and tip.
- *Average wait time per trip* It is measured as the sum of wait time (in minutes) experienced by all drivers and across all trips divided by the total number of drivers and trips.
- *Probability of unmet demand* We equate the total number of demand to the total number of possible pickups (i.e., 1044 drivers × 62 trips), and allocate the demand across different location-time sets based on the frequency distribution of pickups observed in the data. We then count the number of occurrences when demand is greater than the simulated driver supply in each location-time set during a one-hour period, and calculate the probability of unmet demand for each location-time set as the number of occurrences with demand greater than supply divided by the number of hours over the entire simulation horizon. We report the minimum, maximum, and the average of the probabilities across all location-time sets.
- *Total number of unmet demand* We count the lost demand (if any) in each one-hour period in a location-time set and sum it up over all unique combinations of

the location-time sets and hours.

1.7.3 Simulation Results

We run each of our 3 policy scenarios for 100 independent replications. We report the simulation results averaged across the 100 replications in Table 1.5. In the sections below, we compare the two aforementioned policies with the baseline policy, and discuss their impact on the outcome measures including drivers' hourly tip income, hourly total income, wait time, and unmet demand.

Table 1.5: Simulation results

	Baseline	Hinder learning	Service charge	Hinder - Baseline	Service - Baseline
			Drivers' welfare		
Hourly tip income (\$/hr)	1.916 (0.037)	1.690 (0.031)	1.992 (0.029)	-0.226*** (0.005)	0.077*** (0.005)
Hourly total income (\$/hr)	13.702 (0.219)	12.065 (0.205)	14.231 (0.208)	-1.637*** (0.030)	0.529*** (0.030)
			System efficiency		
Wait time per trip (minute)	10.663 (0.062)	10.519 (0.094)	10.554 (0.059)	-0.144*** (0.011)	-0.109*** (0.009)
Prob(unmet demand)_min	0.040 (0.007)	0.029 (0.009)	0.038 (0.009)	-0.011*** (0.001)	-0.002** (0.001)
Prob(unmet demand)_max	0.369 (0.017)	0.332 (0.014)	0.371 (0.017)	-0.038*** (0.002)	0.002 (0.002)
Prob(unmet demand)_mean	0.172 (0.003)	0.155 (0.003)	0.171 (0.003)	-0.018*** (4e-04)	-0.002*** (4e-04)
Unmet demand	5029.530 (85.230)	4911.198 (100.147)	4990.495 (91.956)	-118.333*** (13.151)	-39.035*** (12.538)

Note: Standard deviations across 100 replications are in parentheses. *** indicates significance at 1%; ** indicates significance at 5%.

Result 1 *Hindering drivers from learning about tips reduces their hourly tip income and total income, but improves the system efficiency, indicated by a shorter wait time for drivers and reduced unmet demand for customers.*

As shown in Table 1.5, when drivers are hindered from learning about tips, their average hourly tip income decreases by about \$0.23 (12%), and their hourly total income (including fare and tips) decreases by about \$1.64 (12%) from the baseline policy. Despite the loss in drivers' income, the system efficiency is improved. Specifically, drivers' wait time is

lower by about 0.14 minutes (1%) per trip, the average probability of unmet demand across location-time sets decreases by about 2% points (10%), and the total unmet demand decreases by about 118 customers (2%) compared with the baseline policy. Interestingly, the maximum probability of unmet demand across location-time sets decreases by about 4% points with the hinder learning policy, more so than the decrease in the average probability of unmet demand, indicating a more equitable distribution of supply across location-time sets.

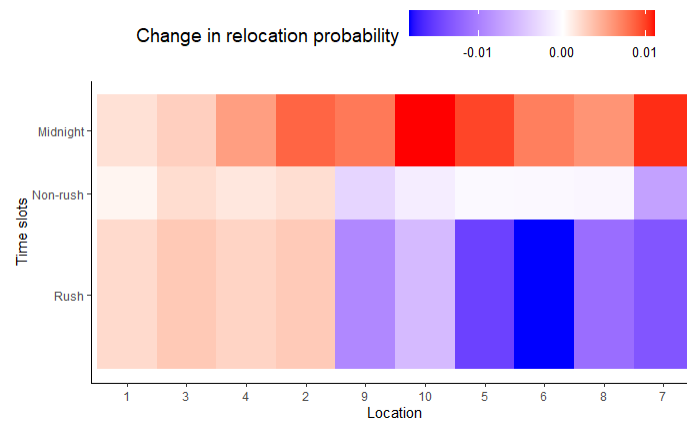


Figure 1.3: Changes in relocation probability in different locations and times (Hinder learning - Baseline)

Note: X- and Y-axes are ordered by the average tip percentages in the location-time set. Rush hours: (7-10 a.m. and 4-8 p.m.); Non-rush hours: (6-7 a.m., 10 a.m.-4 p.m., and 8p.m.-1 a.m.); midnight: (1-6 a.m.). The height of the blocks corresponding to different time slots are proportional to the frequency with which trips under the hinder learning policy belong to these time slots in the simulation.

To understand the rationale behind these results, let us consider Figure 1.3, which illustrates how the drivers’ frequency of relocation to a particular location-time set changes with the hinder learning policy compared to the baseline. Here, x- and y-axes capture locations and times respectively, which are ordered in their average tip percentages observed in our data. First, we consider drivers’ relocation behavior during the rush- and non-rush-hour shifts, which correspond to the majority of trips in the simulations. With the hinder learning policy, drivers are more likely to relocate to areas associated with lower average tips. This is because it is more difficult for drivers to develop a good understanding of the tip distribution across location-time sets over time under the hinder learning policy. While individual drivers may have different prior beliefs about tips at different locations, the dis-

tribution of these beliefs across drivers do not vary as much for different locations as in the baseline policy which allows for tip learning. Overall, that induces drivers to explore areas associated with lower tips more.

Next, we observe from Figure 1.3 that trips occurring in the midnight shift are more common with the hinder learning policy than the baseline policy. The rationale is that drivers tend to have limited prior experiences with the midnight shifts and hence less accurate prior beliefs about tips incurred during midnight. Hence, once a driver with optimistic prior beliefs starts a midnight shift, he may be more inclined to continue the shift, as opposed to choosing the outside option. In fact, our simulations indicate that drivers' probability of continuing a midnight shift increases by about 1% point with the hinder learning policy than with the baseline.

Overall, drivers' income is lower with the hinder learning policy due to a reduction in their tip income particularly during rush and non-rush hour shifts, which dominates any potential increase in their tip income from the midnight shifts. In contrast, the hinder learning policy mitigates the shortage of supply and reduces unmet demand in location-time sets which can be underserved under the baseline. To see this point, consider Figure 1.4, which illustrates the probabilities of unmet demand in different locations and times under the baseline scenario. As evident from the figure, within each of the three time slots, the probability of unmet demand is consistently greater in low-tip areas (indicated by the darker blue area to the left of the x-axis), which coincide with underprivileged neighborhoods where household income is relatively low²⁸. Moreover, the probability of unmet demand is slightly higher during midnight than rush hour, and is the lowest in high-tip locations during the rush hour (indicated by the white area in the bottom right of the figure). Therefore, the system efficiency improvement with the hinder learning policy described above can be attributed to two changes: drivers' increasing frequency of relocation to low-tip areas and

²⁸Location 1, 3, 4 correspond to Washington Heights, Central Harlem, and East Harlem, respectively. The median household income in these three locations is relatively low, based on data from <https://data.ccnyc.org/data/map/66/median-incomes#66/39/3/107/127/a/a>.

working longer shifts during midnight. Overall, our results show that the uncertainty in drivers’ prior beliefs about tips incentivize them to explore the locations and times which can be underserved in the presence of learning.

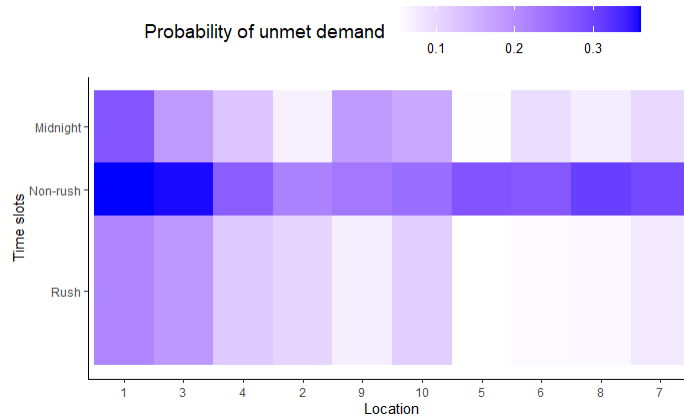


Figure 1.4: Unmet demand in different locations and times under the Baseline scenario
 Note: X- and Y-axes are ordered by the average tip percentages in the location-time set; Rush hours: (7-10 a.m. and 4-8 p.m.); Non-rush hours: (6-7 a.m., 10 a.m.-4 p.m., and 8p.m.-1 a.m.); midnight: (1-6 a.m.).

Result 2 *Replacing tips with an equivalent service charge not only increases drivers’ hourly tip income and total income, but also slightly improves system efficiency, indicated by shorter wait time for drivers and less unmet demand for customers.*

Table 1.5 shows that when tips are replaced by a service charge equivalent to the average tip percentage under the baseline policy, drivers’ hourly tip income increases by about \$0.08 (4%), and their hourly total income increases by about \$0.53 (4%) from the baseline. Moreover, the average probability of unmet demand under the service charge policy is slightly (0.16% points) lower than the baseline, and its magnitude reduces by about 39 customers (1%) from the baseline.

Why would drivers’ hourly income increase when tip is replaced by an equivalent service charge? To explain the rationale, we consider Figure 1.5, which shows how drivers’ relocation frequency to different location-time sets change with the service charge policy compared to the baseline. With the service charge policy, drivers no longer rely on their beliefs about tips for relocation, but exploit the rest of the attributes including the base fare, number of nearby pickups and drop-offs, speed, and wait time to identify where to drive

next. For example, Figure 1.5 shows that the largest increase in frequency of relocation is in Location 10 (Battery Park City, Greenwich Village & Soho) during rush hours (indicated by the darkest red block), where the wait time is the lowest among all location-time sets, yet the tip percentage is below the median. On the other hand, the largest decrease in relocation frequency is in Location 8 (Murray Hill, Gramercy & Stuyvesant Town) during midnight, wherein the tip percentage is the highest among all location-time sets, but the wait time is longer than the median. Overall, since drivers' relocation decisions are not distorted by tips, they are more likely to relocate to locations with higher base fares and shorter wait times, consistent with the positive coefficient of fare and negative coefficient of wait time in our estimation results. Ultimately, both higher base fares and lower wait times contribute to a higher hourly income.

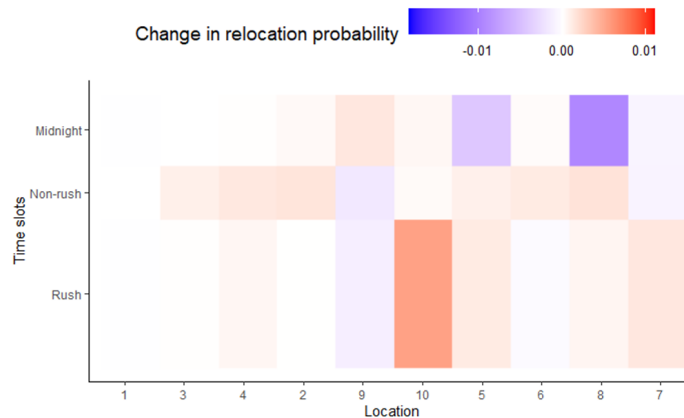


Figure 1.5: Changes in relocation probabilities in different locations (Service charge - Baseline)

Note: X- and Y-axes are ordered by the average tip percentages in the location-time set. Rush hours: (7-10 a.m. and 4-8 p.m.); Non-rush hours: (6-7 a.m., 10 a.m.-4 p.m., and 8p.m.-1 a.m.); midnight: (1-6 a.m.).

Additionally, replacing tips with a service charge improves the system efficiency, but the magnitude of the improvement is limited. On the one hand, the service charge policy helps to eliminate the inefficiency due to tips, by prompting drivers to relocate more frequently to location-time sets with low expected wait times, such as Location 10 or the non-rush hour time sets, which are underserved under the baseline. On the other hand, the service charge policy does not sufficiently motivate drivers to relocate to areas associated

with low tips in the baseline, despite the fact that these areas tend to be underserved as shown in Figure 1.4. This is because these areas happen to coincide with those locations with low demand, presumably due to the low incomes of customers (Chandar et al., 2019), and are also subject to long expected wait times for drivers, especially during the midnight shifts. Therefore, the service charge policy leads to a more modest improvement in the system efficiency compared to the hinder learning policy.

1.8 Concluding Remarks

The service industry is characterized by low base wages (Semuels & Burnley, 2019; Sainato, 2019). Tips have been adopted in some service settings such as the taxi industry, partly motivated by consumers' social preferences (Azar, 2010), and as a way to complement service provider income. This paper evaluates how tips affect drivers' decisions on where to serve and subsequent market outcomes such as driver income and demand-supply match efficiency.

Using a large-scale and granular NYC yellow taxi trip dataset and a logit regression, we first show that drivers' relocation decisions depend on their past experiences related to earnings (from tips and fare), supply and demand conditions, and search costs in different locations and times. This history-dependent relocation behavior can be considered as an evidence for drivers' learning. To disentangle the impact of drivers' intrinsic preferences from that of their prior beliefs and the learning process, we propose a discrete-choice structural model that captures drivers' learning from their past experiences in a Bayesian fashion. In our Bayesian learning framework, as drivers complete more trips, they learn about the attributes that reflect their earnings (such as tip and fare), as well as supply and demand conditions (such as nearby pickups and drop-offs) and search costs (such as trip speed and wait time) in different locations and times, and make relocation decisions based on their updated beliefs. Through model comparison, we further show that our Bayesian learning model outperforms the rational expectations model in terms of in-sample and out-

of-sample model fit, supporting the existence of driver learning within the data.

Our results reveal a previously overlooked consequence of presenting tipping options in settings with independent service providers. By relying on tip information, drivers can increase their income through more informed spatial customer search and relocation decisions, but these decisions may create an imbalance between demand and supply, especially in some underprivileged areas where tips tend to be low. Furthermore, drivers' sensitivity to tip amount is greater than that to fare amount, suggesting that the source of income (tips vs. fare) is an important leading factor in service providers' relocation decisions.

Our framework provides practitioners such as app-based taxi-hailing companies with a tool to manage tip-related information sharing with drivers. Our counterfactual results show that enabling learning about tips (e.g., by providing detailed tip information to drivers immediately after it is available) improves drivers' hourly income, but at the expense of some customers. Customers in some low-tip neighborhoods may have to pay extra or experience longer wait time due to reduced supply of drivers, which can increase customer churn rate in the long run. On the other hand, drivers in high-tip neighborhoods may experience longer wait times as a result of intensified competition.

Hindering learning from tips (for example, by withholding tip information to be shown to drivers) decreases drivers' hourly income, but improves the system demand-supply matching efficiency. When learning is hindered, drivers have to rely on their (possibly less accurate) prior beliefs about tips for relocation, and they may end up relocating to some location-time sets where the actual tips are low. Although drivers' overall income can be reduced due to lower tips, increased relocation to these low-tip regions mitigates the shortage of supply and reduces unmet demand, benefiting the customer population as a whole.

In contrast to the previous two policies that reveal a trade-off between drivers' welfare and the benefits of customers, replacing tips with an equivalent service charge not only increases drivers' hourly income, but also improves the system efficiency. With the service charge policy, drivers' relocation decisions are not distorted by their beliefs about tips, but

are guided by the demand and supply conditions, thus reversing the negative impact of tips on drivers' income and system efficiency. However, the efficiency improvement is modest because drivers are less likely to relocate to location-time sets wherein the wait time is longer, which happen to coincide with some location-time sets with lower tips, presumably due to lower incomes of customers. For a more meaningful improvement in the system efficiency, the service charge policy may need to be complemented with other changes (such as bonuses or subsidies).

Our study has certain limitations that can be addressed in future research. Our data comes from the taxi industry, potentially limiting the generalizability of our insights to platforms. It is important to note that tipping is a well-established norm in the taxi industry, whereas it is still relatively new in platforms, and it may take more time to cultivate consumers' tipping habits. Nevertheless, tipping has been commonly adopted as an important source of compensation for low-paid gig workers, and we are observing a momentum towards customer tips in some platforms. For example, Uber drivers and delivery partners have earned more than \$600 million in tips in one year (Uber, 2020b) after the tipping option was first implemented. Tips are making up a significant portion of Instacart drivers' overall income (Donovan, 2016). Future research-based platform data could help enhance the generalizability of our conclusions to platforms. Additionally, our analysis builds on the assumption that drivers' experiences and observations during pickups and drop-offs have the biggest impact on their relocation decisions. However, as more data becomes available on the full trajectory of drivers in the NYC, observations in the entire trajectory can be incorporated to update drivers' beliefs in location-times along the way. Our study represents the first attempt to examine the role of tips in drivers' relocation decisions and the impact of popular tipping policies, and future research based on more granular trajectory data could help verify the conclusions drawn in our research.

CHAPTER 2

REDUCING GENDER DISPARITIES IN ONLINE LABOR PLATFORMS: THE ROLE OF PERFORMANCE FEEDBACK AND TASK EXPOSURE

2.1 Introduction

The gender earnings gap is a prevalent issue in the traditional labor market (Hoisl & Mariani, 2017), with studies showing that women in 2022 still earned 18% less than men on average¹, despite the efforts to reduce gender disparities in education and work experience. Beyond these conventional human-capital variables, factors such as gender stereotypes (Chan & Wang, 2018), discrimination (Goldin & Rouse, 2000), gender-based occupational segregation (Das & Kotikula, 2019), demand for schedule flexibility (Goldin, 2014), and willingness to compete and negotiate (Niederle & Vesterlund, 2007; Flory et al., 2015; Leibbrandt & List, 2015), are increasingly being recognized as contributing to the gap. For a comprehensive review of this literature, we refer the reader to Blau and Kahn (2017).

Platform-based gig economy has emerged and opened new livelihood pathways, particularly for marginalized or disadvantaged groups. In particular, online labor platforms are believed to narrow the gender earnings gap as they differ from their traditional counterparts in several aspects, such as allowing workers greater flexibility in terms of work schedules and locations (Foong et al., 2018), and the relatively homogeneous nature of tasks presented at these platforms (Litman et al., 2020). However, they are increasingly scrutinized for failing to deliver on that promise (Daniels & Grinstein-Weiss, 2019; Kasliwal, 2020), as evident from calls for regulatory interventions to address disparities among gig workers².

¹<https://www.weforum.org/agenda/2023/03/charted-the-gender-pay-gap-in-the-us-hasnt-changed-much-in-decades/>

²<https://fair.work/wp-content/uploads/sites/17/2023/07/Fairwork-Gender-Report-2023-FINAL-red.pdf>

Recent research has shown that the longstanding gender disparities can carry over to online platforms (Kricheli-Katz & Regev, 2016). These disparities are often attributed to demand-side factors such as employers' gender biases or discrimination (Chan & Wang, 2018; Hannák et al., 2017). On the supply side, there is empirical evidence of gender differences in workers' job application strategies and earnings in online labor platforms (Foong et al., 2018; Fernandez-Mateo & Fernandez, 2016; Liang et al., 2018). However, anticipation of employers' gender biases or discrimination does not fully explain such gender differences, as differences in task selection and earnings remain evident even in gender-blind online labor platforms such as MTurk and Prolific (Litman et al., 2020; Hara et al., 2018), wherein information about an individual's gender is invisible during the recruitment process. This necessitates a further look into the behavioral drivers of the gender disparities in task selection and earnings in online platforms, which is the focus of this paper. Specifically, we focus on gender differences in task selection as a contributor to the gender earnings gap, and drawing from sociology and psychology literatures, performance self-evaluation as a driver of these differences. We then evaluate interventions such as performance feedback and task exposure to mitigate the effects of self-evaluation and gender disparities in online labor platforms.

In this study, we conduct a two-phase field experiment on Prolific, an online platform that presents work opportunities in the form of micro-tasks and is gender-blind in recruitment. In our experiments, workers are asked to grade elementary school-level math tests in exchange for compensation and are presented with two tasks of varying difficulty levels (low corresponding to grade 3 and high corresponding to grade 5, respectively). We find that although there is no gender gap in actual performance in terms of the accuracy rate of the graded submissions by workers, women earn significantly less than their male counterparts, resulting in a gap of 22% in average pay. These differences are driven by women's tendency to select easier and lower-paying (EL) tasks than men in our experiments, even after accounting for other sources of heterogeneity among workers such as demographics,

online work experiences, and actual performance in the task. Furthermore, we find that the gender-based differences in task selection can be attributed to women’s negatively biased evaluation of their performance, while other potential explanations such as risk aversion can be ruled out.

In terms of the mitigation strategies, we find that early exposure to harder and higher-paying (HH) tasks and performance feedback combined are effective in increasing women’s hourly pay and likelihood of selecting HH tasks, but do not significantly affect men. Performance feedback (task exposure) alone helps to narrow the gender earnings gap, but may not fully eliminate differences in task selection behavior, as indicated by 18% (16%) points fewer women select HH tasks, significant at 0.05 (0.10) level. Only when workers have previously been exposed to hard tasks and received performance feedback on them, the gender differences in task selection substantially reduce to 9% points and become insignificant, while the gender gap in average pay remains insignificant. This is partly because performance feedback helps to alleviate the effect of workers’ self-evaluation biases on task selections, especially for women who have negative evaluation biases on their performance on HH tasks.

Our study makes the following contributions to the research on gender disparities in online labor platforms. First, this study is among the first few attempts (e.g., Coffman et al. 2023) to examine whether and how *supply-side* factors can contribute to the gender disparities in online labor platforms. We contribute to this nascent literature by examining gender differences in performance self-evaluation as additional potential behavioral mechanisms that drive online workers’ task selection. Our findings suggest that eliminating demand-side discrimination alone does not ensure an equitable and inclusive marketplace due to gender disparities in workers’ task self-selection behavior. Exposing workers to experiment with challenging tasks early on and providing performance feedback on them is likely to be successful in further closing the gap.

Second, our study enriches the discussions over human-algorithm interactions in online

platforms. While previous research has emphasized the role of performance feedback in improving matching efficiency between employers and workers by signaling worker quality (Tadelis, 2016), its impact on worker welfare has been relatively underexplored. Our research contributes to this discussion by highlighting the benefits of unambiguous performance feedback in helping to bridge gender disparities in task selection among equally capable online workers. Furthermore, it has been recognized that algorithmic job matching, based on the historical patterns of job-seeking behavior, might inadvertently exacerbate the gender gap in the future (Wall & Schellmann, 2021). Our study corroborates these insights and underscores the importance of scrutinizing and adjusting these algorithms to proactively expose workers to potentially more challenging jobs that they might not consider otherwise. We find that such task exposure encourages women to break free from the pattern of picking easier tasks, and instead to pursue difficult and higher-paying tasks later on, thereby contributing to more equitable and inclusive task choices between genders in online labor platforms.

2.2 Literature Review

Our paper contributes to the literature on gender disparities in online platforms and the strategies to mitigate them, as well as the gender gap in performance self-evaluation biases. Below, we review each of these streams of literature and explain how we contribute to them.

2.2.1 Gender Disparities in Online Platforms and Mitigation Strategies

The phenomenon of gender-based employment segregation, which refers to the unequal representation of female and male workers in certain job categories, has long been identified as a leading contributor to the gender earnings gap in traditional labor markets (Das & Kotikula, 2019). This imbalance not only persists but has been found to intensify over time (Blau & Kahn, 2017). Despite the critical differences of online platforms such as

increased flexibility and autonomy, research has shown that task-based differences in the representation of men and women persists in online platforms as well. Both demand-side and supply-side factors are demonstrated to contribute to these disparities. For example, employers may exhibit statistical discrimination and biased beliefs against specific gender groups, influencing their hiring decisions (Chan & Wang, 2018; Coffman et al., 2021). On the supply-side, analyzing the hourly earnings of over 22,000 workers across nearly 5 million tasks on MTurk (a gender-blind online labor platform), Litman et al. (2020) finds that women tend to opt for tasks with lower pay rates, contributing to a gender earnings gap, with women earning on average 10.5% lower than men, even after controlling for a large number of co-variables such as race, age, income, or family composition. These findings highlight the complexity of the gender earnings gap in online labor platforms.

Analysis of demand-side factors points to employers' gender biases and discrimination as significant contributors to gender disparities in workers' job applications and earnings. For instance, studies have highlighted how racial and gender biases on platforms like TaskRabbit and Fiverr influence the evaluations received by workers and, consequently, their employment prospects (Hannák et al., 2017). Additionally, employers are found to exhibit preferences for hiring female workers for traditionally feminine tasks, such as administrative and customer support, which leads to task segregation and wage differences (Chan and Wang (2018)). Mitigation strategies to address the demand-side factors predominantly aim to adjust how information is shared with potential employers and customers. Examples include detailed information sharing about workers' performance indicators to better signal worker credibility to potential employers (Chan & Wang, 2018), enhancing review visibility to mitigate the differences in the number of reviews received (Hannák et al., 2017), and anonymizing worker identity to minimize possible biases from the demand side (Chan & Wang, 2018).

In contrast, the drivers of supply-side factors and the mitigation strategies targeting them remain relatively underexplored in the literature (Liang et al., 2018; Coffman et al.,

2023). Most existing studies provide empirical evidence that the differences in how men and women approach job applications and selections play a crucial role in shaping the gendered landscape of online employment (Foong et al., 2018; Fernandez-Mateo & Fernandez, 2016; Liang et al., 2018; Cook et al., 2021). For example, on Freelancer.com, women's hourly earnings were found to be roughly 81.4% of men's, attributed to factors like workers' bid timing, job selection based on hourly wage budget, and monitoring avoidance (Liang et al., 2018). Similarly, studies on Upwork and Uber have demonstrated significant gender earnings gaps, not fully explained by differences in workers' qualifications such as experience or education, but also by workers' preferences related to job pay rate setting (Foong et al., 2018), customer search preferences, and working hours (Cook et al., 2021).

The aforementioned empirical evidence calls for a more formal examination of the behavioral drivers of gender differences in job application and selection observed in online platforms, as well as the strategies to mitigate them. Within the limited body of literature that attempts to do so, a notable example is (Coffman et al., 2023), which shows that the ambiguity of required qualifications (i.e., where the bar is) can contribute to gender differences in job application decisions in online labor platforms, while ambiguity reduction has limited success on reducing gender gaps in a controlled experiment setting. Our study complements Coffman et al. (2023) and contributes to this growing body of literature by focusing on gender differences in performance self-evaluation biases as potential underlying behavioral mechanisms (further discussed in subsection 2.2.2), and by exploring performance feedback and task exposure as possible mitigation strategies even when the ambiguity of required qualification is low.

In online platforms, performance feedback such as reviews and ratings are vital mechanisms to reduce frictions caused by information asymmetry (Tadelis, 2016), acting as quality signals to users (Cui et al., 2020) and facilitating platforms' internal assessment of workers (Rahman, 2021). Much of the existing literature on feedback mechanisms has concentrated on the role of feedback for enhancing the matching efficiency between employers

and workers, by facilitating the identification of competent workers from employers' perspectives (Tadelis, 2016; Kokkodis & Ipeirotis, 2016). However, the impact of feedback on workers has received relatively less attention (Paul, 2018). Among the recent body of work in this area, Rahman (2021) demonstrates that opaque third-party evaluation algorithms, such as private performance feedback that is invisible to workers, can urge them to limit their activities in online labor platforms (Rahman, 2021). Our study aims to contribute to this nascent area by examining how performance feedback can mitigate the effects of self-evaluation biases, particularly in the context of gender disparities in task selections among online workers of comparable ability.

Another potential mitigating strategy examined in this paper is to expose workers to certain tasks early on, such as those they might initially perceive to be difficult. Gneezy et al. (2003) argue that participants' task choices not only depend on their estimated ability in completing different types of tasks and the difficulty levels of each, but also on participants' risk and ambiguity aversion. Exposure to a particular type of task early on can reduce the ambiguity associated with working on such a task and might encourage subsequent selection of similar tasks, with a potentially disproportional effect on women (Coffman et al., 2023). In practice, workers' exposure to tasks can be influenced through task assignments in online platforms. For example, similar to the automatic matching algorithm on rideshare platforms, TaskRabbit had a "Quick Assign" feature where customers can simply request that a task be completed within a given timeframe, at a given price, by any available worker (Hannák et al., 2017). Such a feature allows platforms to tailor workers' exposure to various tasks, thereby potentially influencing their future task choices. Additionally, the presentation of tasks to workers through search result rankings or personalized job recommendations (Upwork, 2024a) can further shape their task exposure.

2.2.2 Gender Gap in Performance Self-Evaluation Biases

Research in sociology and psychology has consistently identified a gender gap in performance self-evaluations and decision-makings in traditional offline contexts. For instance, drawing from longitudinal survey data on high-school students, Correll (2001) examines how students' perceptions of their abilities in academic subjects are shaped by societal gender norms and influence their choices of career paths. Specifically, men make higher assessments of their own math competency than their otherwise equal female counterparts, and their self-assessment of task competency partly contributes to the decision to pursue a career in quantitative professions. Furthermore, the gender gap in performance self-evaluations is contingent on situational factors such as the nature of the task (masculine or feminine) (Lenney, 1977; Exley & Kessler, 2022) and its difficulty level (Beyer, 1990; Krawczyk & Wilamowski, 2019). Specifically, women's self-evaluations are found to be more dependent on the situational factors, while men's evaluations tend to be more stable and independent. For instance, Beyer (1990) find that women perceived masculine-typed tasks (e.g., answering questions about political and sports figures) as more challenging and women's self-evaluations of their performance were lower than the actual. In contrast, even though men perceived feminine-typed tasks (e.g., answering questions about stars and fashion) as more difficult, their self-evaluations remained accurate, potentially due to higher self-confidence. Analyzing over 1.6 million amateur runners' records, Krawczyk and Wilamowski (2019) find that men tend to have a more optimistic estimate of finishing time than females especially in more challenging competitions such as marathons, compared to shorter 10 km runs. In a controlled online experiment, Exley and Kessler (2022) demonstrate that most men overestimate their performance on math and science tests, whereas the majority of women tend to underestimate, even when they know the evaluations will be disclosed to potential employers.

Different from the aforementioned literature, our study seeks to investigate the role of gender differences in self-evaluation as potential contributors to the gender earnings gap

in online labor platforms. This exploration is crucial because, given the unique dynamics at play in digital environments, it is not clear whether self-evaluation biases persist or to what extent they influence workers' task selection behavior in online labor platforms. First, on the demand side, the potential for employer discrimination or gender-biased beliefs may be dampened in platforms where personal identifies can be easily concealed or manipulated (through anonymous IDs or fictional profile pictures), creating a more neutral environment (Chan & Wang, 2018; Litman et al., 2020; X. Chen et al., 2023). Then, anticipating reduced discrimination from employers, workers might show less regard for societal stereotypes or expectations in selecting tasks. Second, the nature of the online tasks is expected to mute the effects of self-evaluation biases, as many online tasks tend to be standardized in nature and can be completed without specialized training or advanced degrees (Litman et al., 2020). On the other hand, if gender differences in self-evaluation biases, albeit small, persist in online labor platforms, they might have an outsized impact on workers' task selection behavior. This is because job recommendation algorithms in online labor platforms are typically trained on data reflecting workers' past choices, including workers' job history³ and previous interactions with the job recommendations⁴. Given gender differences in self-evaluation biases, the resulting gender disparities in task preferences and selections can lead to historical biases (Mehrabi et al., 2021) that will be reflected in future recommendations (Mehrabi et al., 2021; Z. Chen, 2023). Hence, biased algorithmic recommendations might further affect workers' future task selection behavior, thus creating a "data-algorithm-user" feedback loop and potentially amplifying the impact of the biases. In addition to the empirical ambiguity, online labor platforms offer novel and practical opportunities for influencing workers' choices to help narrow the gender earnings gap, once the underlying behavioral mechanisms are well examined. These opportunities include gender quotas in recruitment, tailored task listings and recommendations, or offer-

³<https://community.upwork.com/t5/Freelancers/Best-Matched-Algorithm-for-quot-Recently-posted-quot-Jobs/m-p/1511756>

⁴<https://community.upwork.com/t5/Product-Release-Notes/quot-Interesting-Job-quot-Feature-and-Improving-Recommendations/bc-p/164411#M12251>

ing task-level feedback. In this paper, we examine how workers’ early exposure to more difficult tasks among micro-tasks of varying difficulty levels as well as performance feedback can be leveraged to mitigate the impact of self-evaluation biases on workers’ task selection behavior and earnings.

2.3 Method

2.3.1 Empirical Context

We conduct our experiments on Prolific⁵, a leading online labor platform. Prolific has a user base of over 130,000 active workers globally and is widely utilized by researchers for various purposes, including visual evaluation (Anwyl-Irvine et al., 2020), transcription (Gesiarz et al., 2020), surveys (Tandon et al., 2021), customer ratings collection (Schneider et al., 2021) and decision-making games (Read et al., 2017; S. Huang et al., 2020), among others.

Unlike other online labor platforms (e.g., Upwork and Freelancer), wherein demographic information such as gender, race, and age is visible or can be inferred from profile pictures during the recruitment process, Prolific anonymizes workers and assigns them unique 24-character alphanumeric IDs (Appendix section B.1 provides a comparison of the two interfaces). The recruitment process at Prolific does not display individual-level demographic information, in line with its commitment to inclusivity and fairness. Prescreening criteria are available for recruiters to match eligible workers with their target participant pool based on characteristics such as professions and demographics. Once the workers qualify, some of them cannot be prioritized over the others based on individual-level characteristics.

⁵<https://www.prolific.co/>

2.3.2 Experiment Design

The experiment is comprised of two phases, as depicted in Figure 1. At a high level, in Phase 1, workers are recruited to complete Task 1, which involves grading elementary school-level math test problems, filling out a short survey about their performance self-evaluation and demographics, and playing a simple investment game used to measure workers' risk preferences.

The math grading task has two difficulty levels, easy and hard, with low (\$1.5) and high (\$3) payment rates, respectively. In the following sections, we refer to the easy and low-payment task as “EL”, and the hard and high-payment task as “HH”. There are 15 questions in the EL task and 10 questions in the HH task, such that the estimated completion time of EL or HH task is consistent, both ranging from 10 to 15 minutes. Our task duration and payment structure is consistent with other typical tasks on Prolific. The median task duration on Prolific is around 10 minutes, with 50% of the tasks taking about 5 to 15 minutes, 25% taking less than 5 minutes, and 25% taking more than 15 minutes to complete⁶. Prolific enforces a minimum payment of \$8 per hour and recommends studies pay \$12 per hour. To encourage high levels of engagement and ensure good data quality, recruiters are advised to set payments in commensurate with the effort required for each task. As such, payments may vary depending on the difficulty level of tasks⁷.

Depending on the experimental condition that a worker was randomly assigned to, the worker either completed a task of his/her choice or an HH task in Phase 1. After completing Task 1, workers' responses to the performance self-evaluation question were compared to their actual performance in order to categorize them into positively-biased, negatively-biased, and unbiased groups. On the same day, all workers who have completed Task 1 were notified of another opportunity to work on a similar grading task (Task 2), which would be posted the next day around the same time. Workers were randomly assigned to receive a

⁶<https://community.prolific.co/t/length-of-studies/938>

⁷<https://researcher-help.prolific.co/hc/en-gb/articles/4407695146002>

notification with or without performance feedback on Task 1, with the random assignment stratified based on their performance self-evaluation biases. In Phase 2, all workers who decided to continue with Task 2 were instructed to select between EL or HH. Below, we describe each step in more detail. The experimental stimuli and protocols described below were approved by the institutional review board of the authors' university.

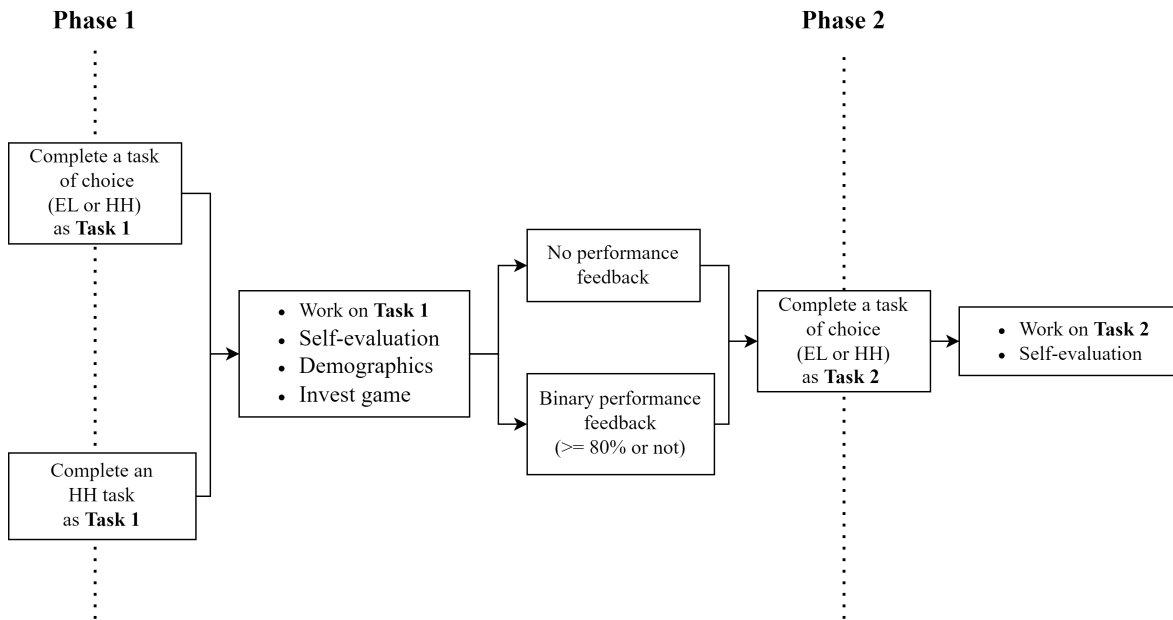


Figure 2.1: Field experiment design flow

Phase 1 – Math Grading

In our experiments, participants were asked to complete elementary-school math grading tasks in exchange for compensation. The simple math grading tasks have been preferred for the following reasons: First, these tasks do not require specialized knowledge, consistent with the homogenous nature of the tasks in online platforms. Indeed, evaluating the accuracy of answers to math questions is a common task completed by workers on online labor platforms (see Appendix section B.2 for more examples). Second, the literature has demonstrated that there are no gender differences in performance in easy math tasks (Niederle & Vesterlund, 2007). Finally, these tasks allow us to evaluate performance and, depending on the experimental condition, provide feedback in a quantitative manner almost immediately,

which is necessary for the two experimental phases to be completed back-to-back without any delays so that potential confounding factors are minimized.

We obtained the math test questions from the publicly available math worksheets on the K5 Learning website⁸, which provides learning tools for students from kindergarten to grade 5 and has been used by millions of parents and teachers since its launch in 2011. The EL and HH math grading task includes 15 questions at grade 3 and 10 questions at grade 5 level, respectively. The math test questions require students to apply elementary school-level math skills, such as operations with whole numbers and fractions and recognizing equivalent fractions and mixed numbers, to answer questions in practical contexts. We invited a grade 6 student to answer the math questions at two levels independently. The student's answer was handwritten along with a proof of her work (examples in Appendix section B.3). The student's initial accuracy rate was about 97% for grade 3 and 80% for grade 5. To avoid potentially careless grading (e.g., grade all answers as correct without careful evaluation) and to control for the accuracy rate of answers across the two levels, we asked the student to randomly modify some of the answers to be incorrect. Consequently, the two levels as presented in experiments have comparable accuracy rates (about 67% for grade 3 and 70% for grade 5).

Task 1 was advertised on Prolific on August 22, 2023. To investigate workers' task selection behavior in their real-world work context, workers were not made aware of their participation in a gender-related study and were instructed that "We are seeking workers to grade math tests for elementary school students. As a math test grader, you will be responsible for reviewing and grading math tests based on their accuracy." In addition, the workers would also see the piece rate of the task, estimated completion time, estimated hourly payment, and available slots upfront (Prolific recruitment ad is in Appendix section B.4). Once a worker accepted to work on Task 1, the worker was directed to a Qualtrics page where the workers completed the math grading task, a demographics survey, and an investment

⁸<https://www.k5learning.com/free-math-worksheets>

game. Workers were instructed to grade the math tests based on the students' answers and proof of work, and provide feedback on whether each answer was correct or wrong.

Before accepting the task, workers were informed upfront that we would conduct random accuracy checks on their submissions, consistent with a common practice adopted in online labor platforms for quality check with limited capacity (an example is provided in Appendix section B.5). We communicated that if the accuracy rate of the submission is 80% or higher (i.e., deemed eligible), the work will be approved for payment; otherwise, the worker will be asked to return the submission (i.e., deemed ineligible) and receive a merit payment of \$0.5 for completing the task. Before taking on the task, workers had up to two attempts to correctly answer a comprehension question that tested their understanding of the payment scheme. Workers who failed to pass the comprehension question within two attempts were routed to exit the task.

After all workers completed Task 1, we checked each submission for accuracy. This was necessary for identifying the performance self-evaluation biases of workers and for randomly assigning them to experimental conditions in Phase 2 based on a stratification of their self-evaluation biases. Among all the workers who completed the task in Phase 1, 79% of them met or exceeded the 80% accuracy threshold. Consistent with the random quality check scheme, we approved all eligible workers and a random subset (about 43%) of the ineligible workers for payment. Furthermore, in experimental conditions with performance feedback, we provided performance feedback to all workers (regardless of the accuracy rate) after completing Task 1 to maintain the same worker pool across the two phases as much as possible. This aims to avoid a potentially biased worker pool in Phase 2.

Phase 1- Survey

After the workers submitted the math grading work, we asked them to complete a short survey to help us better understand the worker pool. We emphasized that their responses would be anonymous and would not affect their earnings. The survey included three parts:

(1) Self-evaluation questions, including one question about their perceived difficulty level of the task and one question about their performance self-evaluation (i.e., “*How many questions do you think you answered correctly in the grading task?*”). (2) Standard demographic questions, asking about gender, age, ethnicity, education, usual hours per day spent on online labor platforms, annual household income, cumulative approval rate, and the primary reason for working on Prolific. (3) A bonus question, asking the workers to play a simple investment game, commonly adopted by researchers to gauge participants’ risk preferences (Gneezy & Potters, 1997; Gneezy et al., 2009). In the investment game, workers had to decide how many out of 10 tokens to invest in a lottery with a 50-50 chance of either succeeding or failing. The lottery returns three times the bet if succeeding, and returns nothing if failing. Workers were informed that the money earned (up to 60 cents) would be paid as a bonus.

Phase 2 - Math Grading and Survey

On the same day as Task 1, we notified all the workers who have completed Task 1 that a second batch of similar math grading tasks would be released the following day around the same time. In the same notification, workers could receive performance feedback on Task 1 or not, depending on the experimental condition they were randomly assigned to. We assigned workers randomly to one of the four conditions outlined in Table 2.1, and in order to ensure a balanced worker composition in terms of worker self-evaluation across different conditions, we stratified the random assignment based on workers’ self-evaluation biases. We measure each worker’s self-evaluation bias as the difference between the worker’s performance self-evaluation (i.e., estimated number of questions graded correctly) and the actual number of questions graded correctly (Dargnies et al., 2019). If the difference was positive (negative), the worker was positively (negatively) biased; If it was zero, the worker was categorized as unbiased.

Workers in the control group, who did not receive performance feedback, were notified

Table 2.1: Control and treatment conditions

		Intervention 1: HH Task exposure	
		No	Yes
Intervention 2:	No	Control group	Treatment 2
Performance feedback	Yes	Treatment 1	Treatment 3

that their submissions were under review and that they would have the opportunity to undertake a similar grading task the following day. Additionally, they were asked to respond to the notification to express their interest in participating in the subsequent task and to specify their preferred type of task. Conversely, participants in the treatment group, who received performance feedback, were notified about whether their accuracy met the 80% threshold or not. The remainder of the notification content was identical for both groups (Notification messages are in Appendix section B.6).

In Phase 2, workers could choose either an EL or HH task to work on, which contained a different set of questions randomly drawn from the same question pool as in Task 1. After completing the math grading task, the workers were asked to fill out a short survey with two questions about the task difficulty level and performance self-evaluation, respectively.

All workers, regardless of whether they participated in Phase 1 only, or Phase 1 and 2, received their payments for each task at the same time after Phase 2 is completed. In our experiments, since Phase 1 and 2 are conducted on different days, some workers who worked on Phase 1 may not have returned for Phase 2. By conducting Phase 2 in a different session before any payment information on Phase 1 is revealed, we aim to eliminate any income or hedging effects and ensure the worker pool between the two Phases remains consistent (which we later verified in Section 4.3).

2.3.3 Sample and Balance Check

In Phase 1, we targeted 500 workers (250 female and 250 male) residing in the United States, who were at least 18 years old and fluent in English, identified through Prolific’s prescreening option. To ensure data quality, we specified that the workers must have com-

pleted a minimum number of 100 previous submissions and achieved a cumulative approval rate of at least 95%. We eventually recruited 489 workers (241 male and 248 female) for Task 1 because 11 of the 500 slots were blocked by workers who failed comprehension check but did not return their tasks by the end of the day. At the end of Phase 1, the 489 workers were notified about Task 2. Among the workers notified, 407 (83%) of them chose to work on Task 2. Out of the 407 workers who have completed both tasks, 35 workers were removed from further analysis due to extremely short or long task completion duration⁹. To verify that workers start Phase 2 after reading the control and performance feedback messages, we requested workers to reply to the message and indicate their preferred task type for Phase 2. Among the remaining 372 workers, 292 (78%) workers replied before starting Phase 2. Therefore, our final sample consists of 292 workers who have completed both Task 1 and 2 in a reasonable amount of time and read the control and feedback messages.

Table 2 presents the characteristics of the workers, which were obtained from their responses to the demographic questions in the survey, except for gender. For gender information, we obtained it from the external data on worker demographics, which is verified and provided by Prolific and is accessible only after the task is completed. We confirmed that 98% of the self-reported genders were consistent with the records held by Prolific. We do not find any significant differences in observable characteristics of workers between gender groups, suggesting that the listed demographics are unlikely to contribute to any gender differences in workers' task selection behavior.

⁹The minimum completion times observed in a pilot experiment with 80 workers were 7 minutes for Phase 1 (which included the grading task, a demographic survey, self-evaluation questions and an investment game) and 5 minutes for Phase 2 (which included the grading task and self-evaluation questions). Given that our main experiment includes an additional comprehension question, these pilot times should represent the minimum expected duration for our experiment. Therefore, we applied the 7- and 5-minute minimum duration thresholds in the analysis of our experiments. The maximum allowed time automatically set by Prolific is 44 minutes, which is an estimation based on the estimated completion time of the study. Therefore, we applied the maximum duration threshold of 46 minutes (2 minutes were added for the demographics survey and investment game) for Phase 1, and 44 minutes for Phase 2.

Table 2.2: Worker characteristics by gender groups (balance check)

	Male	Female	$\Delta(Male-Female)$	<i>p</i> -value
Number of workers	137	155		
	<i>Age</i>			
Age_18_24	6%	8%	-3%	(0.13)
Age_25_34	31%	22%	9%	
Age_35_44	34%	28%	5%	
Age_45_54	18%	21%	-2%	
Age_55_or_above	12%	21%	-9%	
	<i>Race</i>			
White	70%	72%	-2%	(0.87)
Non_white	30%	28%	2%	
	<i>Education</i>			
Some_high_school	1%	1%	0%	(0.70)
High_school	9%	8%	1%	
Some_college	26%	30%	-4%	
College_grad	50%	45%	5%	
Postgrad	13%	17%	-4%	
	<i>Hours on online labor platforms</i>			
Hour_0_3	74%	73%	1%	(0.71)
Hour_4_8	24%	23%	1%	
Hour_8+	2%	4%	-2%	
	<i>Household income</i>			
Income_25k_less	15%	17%	-3%	(0.22)
Income_26_49k	23%	29%	-6%	
Income_50_99k	43%	32%	11%	
Income_100_149k	15%	13%	2%	
Income_150k_more	5%	9%	-4%	
Cumulative approval rate	98.43%	98.65%	-0.22%	(0.35)
	<i>Reason on Prolific</i>			
Primary income	9%	10%	-1%	(0.27)
Money is the main reason	53%	62%	-9%	
Productive use of my time	33%	23%	10%	
For fun	4%	5%	-1%	

Note: The gender differences in “Task 1 duration” and “Approval rate” are tested using the independent sample *t*-test. *P*-value of the null that the difference of means between two gender group equals 0 is reported in brackets. The gender differences in the composition of age, race, education, hours, income, approval rate, reason are tested using the Chi-squared test. *P*-value of the null that the composition is independent of the gender groups is reported in brackets.

2.4 Results

2.4.1 Empirical Approaches

In the following sections, we conduct both parametric and non-parametric tests to ensure the robustness of our results. Specifically, we employed t -test and Wilcoxon test to compare means and medians across groups, respectively. Furthermore, to account for potential confounding effects of workers' demographics, we conducted regression analyses of piece rate pay, hourly pay, and selection of HH in Phase 2 on gender, controlling workers demographics including cumulative approval rate, race (white or otherwise), age (above 55 or not), annual household income (above \$100,000 or not), education (above college or not), hours on online labor platforms (above 3 or not), and reasons on Prolific (a primary source of income or not)¹⁰.

2.4.2 Is There a Gender Earnings Gap in Phase 1?

Before analyzing the gender earnings gap, we first examine whether there is any evidence of gender difference in performance under the simple math grading task in our experiment. Figure 2 displays the gender differences in workers' performance in Task 1 across all experimental conditions. As illustrated in Figure 2, while men's average grading accuracy in Task 1 is 2% points lower than women's for the EL type, the difference is not statistically significant (p -value is 0.30). Therefore, within each group of workers taking on EL or HH tasks in Phase 1, men and women perform equally well¹¹.

We measure a worker's performance self-evaluation bias as the discrepancy between the worker's self-evaluation (i.e., the estimated number of questions graded correctly) and actual performance (i.e., the actual number of questions graded correctly). Figure 3 illustrates the gender differences in performance self-evaluation biases for the EL or HH types.

¹⁰We dichotomize the control variables to allow for greater degree of freedom in the regression analyses.

¹¹The gender difference in performance among the workers who were assigned to the HH task in Phase 1 is insignificant (difference is 0.0034, p -value is 0.90).

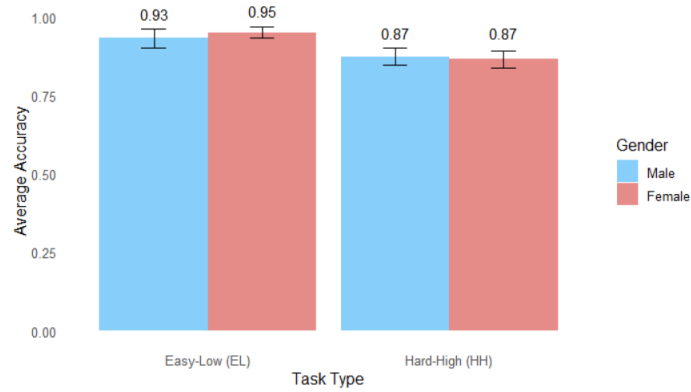


Figure 2.2: Average accuracy in Task 1 of workers who self-selected in Phase 1

Our findings reveal that women exhibit more negative biases than men across both task types. In line with our hypothesis, while men have unbiased self-evaluation for both types of tasks, women display a particularly pronounced negative bias for the HH task, with the bias being significantly different from zero ($p < 0.01$).

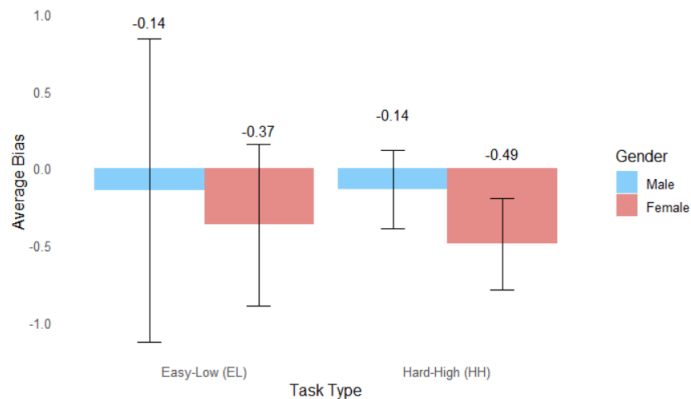


Figure 2.3: Average performance self-evaluation bias by gender

Next, we consider the motivating question for this paper: Do women earn less than men when they have the autonomy to choose which tasks to work on, a critical feature of online labor platforms? To answer this question, we compare the gender earnings gap between the subgroup of workers who worked on a task of their own choice in Phase 1 and those who were assigned to work on an HH task in Phase 1. Figure 4 presents the average pay (\$) and average hourly pay (\$/hour) in Task 1 by gender for each subgroup. From the left panel, we observe that within the subgroup of workers who self-selected into tasks in Phase 1,

men on average earn 0.43 (22%) more than women in terms of piece rate, which refers to the payment from completing the task, and \$0.99 per hour more in hourly pay. The gender earnings gap is significant for the piece rate (p -value < 0.01) and marginally significant for hourly pay (p -value < 0.10) according to a regression analysis, which controls for worker demographics¹². However, in the right panel, where workers work on the HH type tasks regardless of the gender, the gender earnings gap diminishes and is insignificant in our regression analyses.

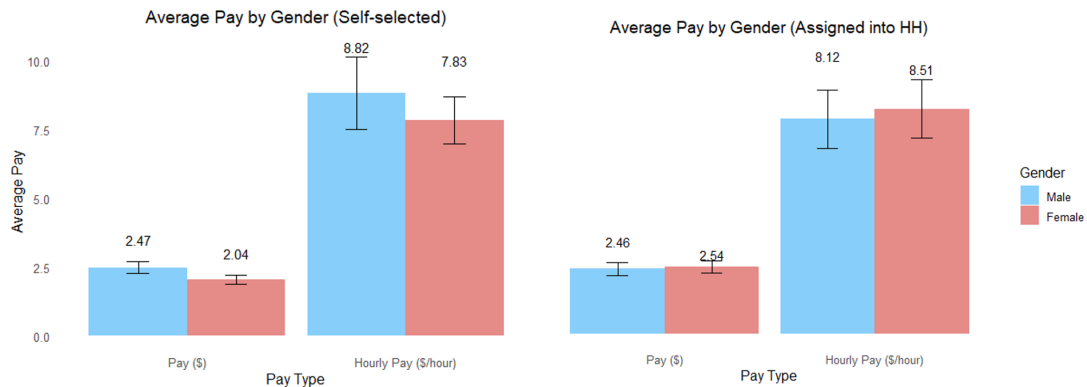


Figure 2.4: Average pay (\$) and hourly pay (\$/hour) of workers who self-selected (left) and assigned to HH (right) in Phase 1

Considering the comparable performance between genders overall, we suspect that the gender earnings gap in the case where workers self-select into tasks can be attributable to women being less likely to select an HH task. Figure 5, which displays the percentage of male and female workers opting for an HH task in Phase 1, indeed shows that while approximately 76% of the men choose to work on HH, only about 42% of the women do so, resulting in a significant 34% points difference, which is significant at 0.01 level based on our regression analysis.

2.4.3 Mitigation Strategies: Performance Feedback and Early Exposure to An HH Task

In the following section, we evaluate the effectiveness of early exposure to an HH task and performance feedback as potential interventions to mitigate the gender earnings gap

¹²Similar conclusions can be reached for comparison with Wilcoxon test.

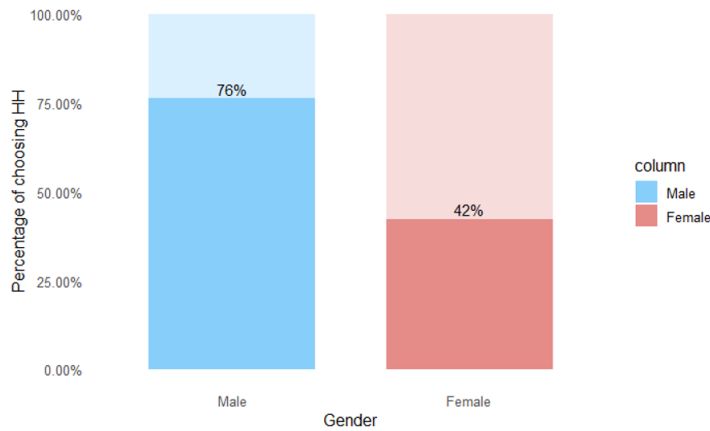


Figure 2.5: Gender differences in the percentage of choosing HH for Task 1

on online labor platforms¹³. Table 3 presents the workers’ piece rate pay, hourly pay, percentage of workers that choose the HH task, as well as the associated gender gaps in Phase 2 within each treatment condition.

- **Gender earnings gap persists in Phase 2 in the absence of any interventions:**

Our findings reveal that in the control group, where participants freely choose their tasks in Phase 2 without receiving performance feedback after Phase 1, the gender pay gap and the disparity in choosing the HH task in phase 2 persists. Specifically, as shown in Table 3, men earn about \$0.58 (29%) more in average pay than women, which is significant at 0.01 level with both non-parametric test and regression analysis. This marks a directional increase from \$0.43 (22%) in average pay in Phase 1. Additionally, although both genders show a higher propensity to opt for the HH task in Phase 2 than in Phase 1, there is a 41% point gap between genders in the percentage of workers choosing HH (which is significant at 0.01 level). Notably, the gap is

¹³To ensure that our worker pool between the two Phases remains consistent, we have checked for each gender whether groups of participants who returned or did not return demonstrated any significant differences in their performance (as measured in Phase 1) as well as their tendency to choose HH in Phase 1. We do not observe any such differences except for Treatment 1, wherein participants who chose to return had significantly higher accuracy rates than those who did not return, in each gender category. However, this is unlikely to bias our results as we do not find any significant differences between men and women who chose to return in terms of their accuracy performance in Phase 1 (difference is 1%, p-value is 0.50).

greater than the 34% point difference observed in Phase 1. These results indicate that in the absence of performance feedback, women, more so than men, may be inclined to follow a similar pattern of selecting easier and lower-paying tasks over time.

Table 2.3: Piece Rate Pay and Hourly Pay in Phase 2

	Male	Female	$\Delta(Male-Female)$	Coefficient of <i>gender</i>
Control: Self-select in Phase 1; No performance feedback (N=69)				
Average Pay (\$)	2.604 (0.821)	2.022 (0.898)	0.582 *** (0.214)	**
Average Hourly Pay (\$/hr)	10.036 (5.013)	9.817 (5.330)	0.219 (1.296)	NS
Median Pay (\$)	3	1.5	1.5 ***	
Median Hourly Pay (\$/hr)	9.033	9.045	-0.011	
% of Choosing HH in Phase 2	88%	47%	41%***	***
Treatment 1: Self-select in Phase 1; With performance feedback (N = 75)				
Average Pay (\$)	2.343 (0.968)	2.063 (0.962)	0.280 (0.223)	NS
Average Hourly Pay (\$/hr)	12.124 (8.413)	9.674 (5.805)	2.451 * (1.693)	NS
Median Pay (\$)	3	1.5	1.5 *	
Median Hourly Pay (\$/hr)	10.898	8.539	2.359	
% of Choosing HH in Phase 2	71%	53%	18%**	**
Treatment 2: Exposure to HH in Phase 1; No performance feedback (N=63)				
Average Pay (\$)	2.217 (1.088)	2.303 (0.856)	-0.086 (0.248)	NS
Average Hourly Pay (\$/hr)	10.882 (7.525)	11.001 (5.455)	-0.118 (1.670)	NS
Median Pay (\$)	3	3	0	
Median Hourly Pay (\$/hr)	10.103	9.845	0.258	
% of Choosing HH in Phase 2	80%	64%	16%*	NS
Treatment 3: Exposure to HH in Phase 1; With performance feedback (N=85)				
Average Pay (\$)	2.417 (0.958)	2.324 (0.922)	0.092 (0.205)	NS
Average Hourly Pay (\$/hr)	10.318 (6.767)	12.804 (6.676)	-2.486 * (1.469)	NS
Median Pay (\$)	3	3	0	
Median Hourly Pay (\$/hr)	8.551	13.300	-4.749 *	
% of Choosing HH in Phase 2	79%	70%	9%	NS

Note: "NS" indicates not significant, * $p < 0.10$, ** $p < 0.05$; *** $p < 0.01$. Significance in column (3) is based on the one-sided test with the alternative hypothesis of $\Delta(Male-Female) > 0$ in Control, Treatment 1, 2, and the two-sided test in Treatment 3. s.d. in the parentheses in column "Male" and "Female"; s.e. in the parentheses in column " $\Delta(Male-Female)$ ".

- **Task exposure and performance feedback combined are effective in increasing**

women’s hourly pay and likelihood of selecting HH tasks, but do not significantly affect men:

Table 4 compares the differences in pay, hourly pay, and HH task selection across treatment groups for men and women, respectively. Table 4 shows that while each mitigation strategy alone improves women’s pay, hourly pay, and likelihood of selecting HH task, the effects are not significant, as indicated by the positive and insignificant coefficients of Treatment 1 and Treatment 2 in the female subgroup analysis (right panel). The effects of Treatment 3 on women’s hourly pay and likelihood of selecting HH tasks are positive and significant (at 0.05 level), suggesting that a combination of both strategies is effective in improving both outcomes. On the other hand, in the male subgroup analysis, the coefficients of the treatment dummies are all insignificant, suggesting that men’s earnings and likelihood of choosing HH in Phase 2 are not affected by the mitigation strategies. In Figure 6, the ANOVA analysis shows that women’s hourly earnings significantly increase in Treatment 3 compared to other conditions (significant at the 10% level), but men do not experience a decrease in their hourly earnings. To sum up, the two interventions together appear to benefit women without hurting men.

- **Task exposure or performance feedback alone helps to narrow the gender earnings gap, but may not fully eliminate differences in task selection behavior:**

In Treatment 1, when workers self-select into tasks in Phase 1 and receive performance feedback afterwards, we continue to find directional evidence of a gender gap in earnings favoring men, which is a sizable difference of \$0.28 in average pay ($p = 0.11$) and \$2.45/hour in average hourly pay (significant at 0.10 level). Moreover, the effect of performance feedback on eliminating differences in task selection behavior is limited, as indicated by women still being 18% points less likely than men to choose the HH task in Phase 2 (significant at 0.05 level). Receiving feedback on

Table 2.4: Comparison among treatment groups

	Male			Female		
	Pay (\$)	Hourly Pay (\$/hour)	Choose HH	Pay (\$)	Hourly Pay (\$/hour)	Choose HH
Intercept	0.914 (4.043)	20.612 (28.476)	1.265 (10.543)	-3.782 (3.726)	7.209 (23.271)	-23.435** (9.787)
Treatment 1	-0.185 (0.263)	2.716 (1.852)	-1.157 (0.75)	0.053 (0.195)	0.283 (1.216)	0.123 (0.460)
Treatment 2	-0.252 (0.273)	0.668 (1.925)	-0.442 (0.800)	0.315 (0.209)	1.786 (1.305)	0.730 (0.502)
Treatment 3	-0.211 (0.245)	0.621 (1.726)	-0.805 (0.734)	0.270 (0.199)	2.757** (1.243)	1.029** (0.497)
Demographics	Controlled in all models					
<i>N</i>	137	137	137	155	155	155
<i>F</i>	0.879	2.090**		2.066**	3.219***	
Adj. <i>R</i> -squared	-0.009	0.074		0.064	0.126	
Pseudo <i>R</i> -squared			0.062			0.091

Note: * $p < 0.10$, ** $p < 0.05$; *** $p < 0.01$. s.e. in the parentheses.

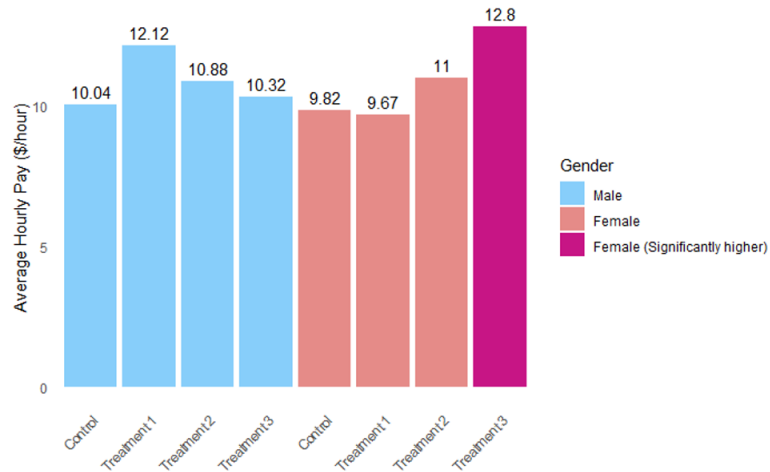


Figure 2.6: Average hourly pay by gender and treatment

Note: The ANOVA test shows that average hourly pay varies significantly across treatment conditions for women (p -value < 0.10), but not for men. Specifically, women's hourly pay in Treatment 3 is significantly higher (p -value < 0.10) than other conditions.

Phase 1 performance does not substantially impact task selections for women, because more than half of women choose to work on EL in Phase 1 and the negative self-evaluation biases among these women working on EL is insignificant to begin with, as evident from Figure 3.

In Treatment 2, when workers work on HH in Phase 1 but do not receive performance feedback, the gender gap in pay (hourly pay) diminishes to \$0.08 (\$0.12 per hour) and becomes insignificant, with women earning slightly higher than men. Moreover, women appear to be encouraged to break free from a pattern of choosing the EL task. Specifically, 64% of women choose the HH task in Phase 2, as opposed to 47% of women making the same choice within the control condition. Despite this improvement, a significant gender gap in task selection (16% points significant at 0.10 level, as noted in the last row for Treatment 2 in Table 3) still persists. The mitigated differences between genders in earnings and task selection can be attributed to the workers' increased familiarity with HH from Phase 1, which can be particularly influential on women. The literature on gender and ambiguity aversion¹⁴ shows that women are more ambiguity averse than men (Coffman et al., 2023), especially when the ambiguity is related to opportunities for a gain (Borghans et al., 2009; Schubert et al., 2000). Working on the HH task in Phase 1 is likely to reduce the ambiguity about the task (e.g., difficulty level and question types), thus disproportionately increasing women's likelihood of choosing the HH task.

In Treatment 3, when workers work on HH in Phase 1 and receive performance feedback, the gender gap in average pay remains insignificant, while the difference in task selection further diminishes and becomes insignificant. Specifically, men and women are similarly likely to choose the HH type for Task 2, with the gender difference in percentage of selection decreasing to 9% points. Furthermore, we find

¹⁴Ambiguity aversion refers to the preference of placing higher values on bets with known probabilities (risk) than bets with unknown probabilities (uncertainty) (Borghans et al., 2009).

evidence for the hourly pay of women dominating that of the men (significant at the 10% level in a two-sided test), with women earning \$2.49 (24%) more than men. This is due to an increase in hourly pay for women in this treatment compared to others, as opposed to a reduction in hourly pay for men (as previously discussed). Given the insignificant differences in average or median pay between genders, the differences in hourly pay likely stem from the differences in task completion times. This is indeed supported by our data. In Treatment 3, women complete the task in Phase 2 significantly faster than men (by 5.73 minutes, significant at 0.01 level), while we do not find differences in task completion times in Phase 1.

2.4.4 Mechanisms: The Role of Performance Self-evaluation Biases

Models

In the previous section, we observe that task exposure and performance feedback together help to reduce the gender gap in task selection and earnings, while also helping to increase women's pay per hour without hurting men. The primary purpose of this section is to investigate behavioral mechanisms behind these observations. Specifically, we hypothesize that in the absence of any intervention, gender differences in performance self-evaluation (which are particularly apparent for harder tasks) may influence workers' task selection behavior, thereby contributing to the gender earnings gap. Performance feedback is then expected to moderate the strength of the mediated relationship between bias and task selection. This is because serving as an additional performance indicator, performance feedback can alleviate workers' reliance on performance self-evaluation when choosing subsequent tasks, which is likely to also influence their earnings.

To test the proposed mechanism, we conduct a mediation analysis using the commonly-adopted causal steps strategy (Preacher & Hayes, 2008; Sterling et al., 2020). That is, we test whether the following three criteria are met: (1) Gender (independent variable) significantly explains the variations in the self-evaluation bias (mediator); (2) Gender significantly

explains the variations in the propensity to select HH in Phase 2 (dependent variable); and (3) The effect of gender on the propensity to select HH is noticeably reduced when the effect of bias is accounted for. Therefore, denoting the worker with i (where $i = 1, \dots, 292$) and the phase index with t (where $t = 1$ or 2), we test the first relationship hypothesized above with the following model:

$$Bias_{i,t=1} = \alpha_1 \times Female_i + \alpha_2 \times Task_{i,t=1} + \mathbf{a} \times Controls_i + \omega_{i,t=1}, \quad (2.1)$$

where $Bias_{i,t=1}$ measures worker i 's performance self-evaluation biases at the end of Phase 1, calculated as the discrepancy between the worker's estimated and actual number of correctly graded questions. $Female_i$ is a dummy variable that equals 1 if worker i is female and 0 otherwise. $Task_{i,t=1}$ is a categorical variable with two levels, indicating the task type (EL or HH) that worker i worked on in Phase 1. The variable $Controls_i$ includes a set of demographic and online work characteristics of worker i (as described in subsection 2.4.1), and $\omega_{i,t=1}$ is the error term assumed to be Normally distributed. The coefficient a_1 measures the gender gap in performance self-evaluation biases.

We test the hypothesized relationship (2) above by:

$$\ln \left[\frac{\text{Prob}(Pick_hh_{i,t=2} = 1)}{1 - \text{Prob}(Pick_hh_{i,t=2} = 1)} \right] = \beta_1 \times Female_i + \beta_2 \times Task_{i,t=1} + \beta_3 \times Risk_pref_i + \mathbf{b} \times Controls_i, \quad (2.2)$$

where $Pick_hh_{i,t=2}$ is a binary variable indicating whether worker i selected the HH task in Phase 2, with a value of 1 indicating the choice of an HH task. The variable $Risk_pref_i$ represents the degree of risk-seeking by worker i , measured by the number of tokens (ranging from 0 to 10) they invested in the investment game. A higher number of tokens indicates

a greater preference for risks. Finally, we test relationship (3) above with the model

$$\ln \left[\frac{\text{Prob}(\text{Pick_hh}_{i,t=2} = 1)}{1 - \text{Prob}(\text{Pick_hh}_{i,t=2} = 1)} \right] = \gamma_1 \times \text{Female}_i + \gamma_2 \times \text{Bias}_{i,t=1} \\ + \gamma_3 \times \text{Task}_{i,t=1} + \gamma_4 \times \text{Risk_pref}_i + \mathbf{c} \times \text{Controls}_i. \quad (2.3)$$

Equation 2.3 differs from Equation 2.2 by the inclusion of an additional variable, $\text{Bias}_{i,t=1}$. The coefficient γ_2 measures the effect of workers' performance self-evaluation biases on their task selection. With Equation 2.3, we aim to assess if the impact of *Female* on task selection (captured by γ_1) is substantially reduced when the effect of *Bias* is accounted for in the analysis.

Next, to test if the effect of self-evaluation bias on task selection is mitigated when the workers receive performance feedback, we conduct two subgroup analyses of Equation (3) with sub-samples of workers who did not receive performance feedback and who received feedback after Phase 1, respectively. We expect the effect of bias to be attenuated when performance feedback is provided, but to remain significant in the absence of the feedback.

Results

Table 2.5 presents the results of our hierarchical regression analysis. In Model (1), we observe a negative and marginally significant coefficient for *Female*, indicating that women tend to have a lower self-evaluation of performance compared to men. Additionally, women are less likely to choose HH in Phase 2 even after accounting for factors such as risk preferences and the task experience in Phase 1, as evident from the significantly negative coefficient of the *Female* indicator. In model (3), we include the self-evaluation bias to test whether controlling the effect of bias substantially decreases the effect of gender on task selection. The significantly positive coefficient of the *Bias* (0.298, significant at the 1% level) indicates that workers with more positive values of the self-evaluation bias are

Table 2.5: Mediation Analysis

Dependent variable	(1) <i>Bias_{i,t=1}</i>	(2) <i>Pick_hh_{i,t=2}</i>	(3) <i>Pick_hh_{i,t=2}</i>	(4-1) <i>Pick_hh_{i,t=2}</i> (no feedback)	(4-2) <i>Pick_hh_{i,t=2}</i> (with feedback)	(5-1) <i>Pay_{i,t=2}</i> (\$)	(5-2) <i>Hr_pay_{i,t=2}</i> (\$/hour)
<i>Intercept</i>	-2.960 (4.404)	-14.316** (6.743)	-14.845** (6.907)	-2.350 (11.471)	-23.876** (9.926)	-0.389 (1.237)	15.147 (15.853)
<i>Female_i</i>	-0.350* (0.185)	-0.683** (0.303)	-0.606* (0.309)	-1.134** (0.526)	-0.365 (0.427)	-0.011 (0.052)	0.725 (0.665)
<i>Bias_{i,t=1}</i>			0.298*** (0.106)	0.499*** (0.19)	0.219 (0.143)	-0.019 (0.017)	-0.201 (0.217)
<i>Pick_hh_{i,t=2}</i>						1.230*** (0.056)	4.229*** (0.719)
Control variables							
<i>Task_{i,t=1}</i>	-0.039 (0.226)	2.292*** (0.351)	2.432*** (0.367)	2.638*** (0.581)	2.419*** (0.517)		
<i>Risk_pref_i</i>		-0.005 (0.046)	-0.017 (0.047)	-0.053 (0.070)	0.013 (0.068)		
<i>Accuracy_{i,t=2}</i>						4.296*** (0.203)	16.920*** (2.606)
Demographics							
<i>N</i>	292	292	292	132	160	292	292
<i>F</i>	2.954***					106.3***	12.4***
Adj. <i>R</i> -squared	0.086					0.799	0.301
Pseudo <i>R</i> -squared		0.198	0.221	0.287	0.235		

Note: s.e. in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

In Model (5-1) and (5-2), the dependent variable $Pay_{i,t=2}$ is the piece rate payment (\$), and $Hr_pay_{i,t=2}$ is the hourly payment (\$/hour) earned by worker i for completing the task in Phase 2. Hourly payment is calculated as the piece rate payment divided by the total hours that the worker spent on the Phase-2 task. $Accuracy_{i,t=2}$ is the accuracy rate (%) of the worker's grading submission

more likely to select HH for subsequent tasks. Moreover, comparing Model (2) and (3), we find the effect of *Female* on task selection decreases in both significance (from 5% to 10% level) and magnitude (from -0.683 to -0.606) after controlling for self-evaluation biases. Despite the reductions, the effect of *Female* remains marginally significant at 10% level. Altogether, these results¹⁵ support the partial mediation role of self-evaluation bias in explaining the gender differences in task selection preferences.

To examine the moderating role of performance feedback, we analyze how workers' self-evaluation biases influence their task choices in Phase 2 in the absence and presence of performance feedback after Phase 1. Model (4-1) reveals that, in the absence of performance feedback, the impact of self-evaluation bias on task selection remains significantly positive (0.499, significant at the 1% level). Conversely, in Model (4-2), the insignificant coefficient of *Bias* implies that the effect of the bias is muted when workers receive

¹⁵We follow the three steps as in Model (1)-(3) to test the mediation role of risk aversion, but it is not supported. Detailed results are included in Appendix section B.7 Table B.1

performance feedback after Phase 1. These results collectively support the effectiveness of performance feedback in mitigating the influence of self-evaluation biases on workers' task choices. Furthermore, we find that this mitigation effect is achieved primarily through the positive performance feedback received by negatively-biased workers, while the effect of negative performance feedback is insignificant (results in Appendix section B.8 Table B.2). This is consistent with the patterns in Table 2.4, which has shown that performance feedback together with exposure to HH task in Phase 1 in Treatment 3 is particularly effective at encouraging women to choose HH in Phase 2 because women, who tend to be more negatively biased and even more so after working on HH in Phase 1, are more likely to receive positive feedback in this treatment.

Finally, to assess whether gender differences in task selection contribute to the gender earnings gap, we perform regression analyses with workers' Phase 2 Pay (\$) and Hourly Pay (\$/hour) as the dependent variables in Models (5-1) and (5-2), respectively, and including *Female*, *Bias*, and *Pick_hh* as explanatory variables in both models. The results from both regressions consistently demonstrate significant positive coefficients for *Pick_hh* (significant at the 1% level). Therefore, choosing the HH task in Phase 2 is associated with higher earnings, which significantly accounts for variations in workers' earnings.

In summary, as illustrated in Figure 2.7, our analysis indicates that performance feedback helps to bridge the gender gap in task selection. This is achieved by mitigating the impact of self-evaluation bias on task selections. Furthermore, the mitigation effect of performance feedback is more prominent when workers are previously exposed to the HH task, where the gender difference in self-evaluation bias is particularly pronounced.

2.5 Concluding Remarks

The gender earnings gap is a persistent issue not only in traditional labor markets, but also across various online platforms like Upwork (Foong et al., 2018), Freelancer (Liang et al., 2018), Mturk (Litman et al., 2020), and Uber (Cook et al., 2021). This gap is often

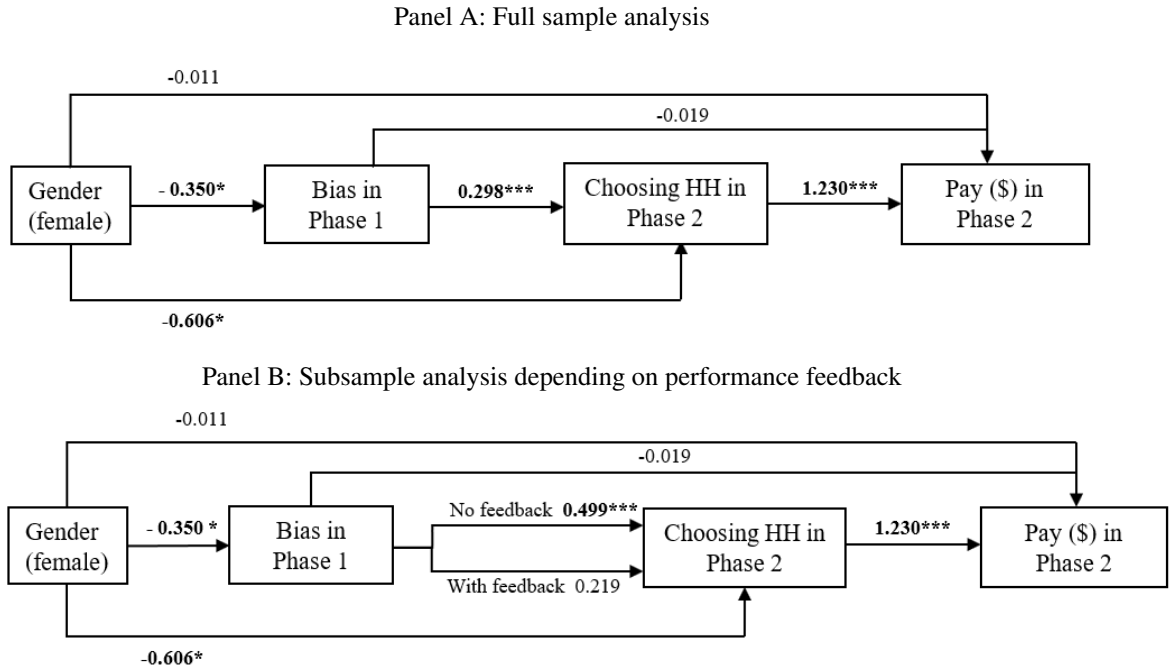


Figure 2.7: Full path mediation analysis

attributed to demand-side factors such as customers’ potential biases against specific service provider groups (Chan & Wang, 2018; Kricheli-Katz & Regev, 2016; Coffman et al., 2021), but paradoxically, also persists in gender-blind online labor platforms Litman et al. (2020). This limits the potential of these labor platforms for reducing disparities within the worker population. In this paper, we focus on gender differences in workers’ job selection as a contributor to the gender earnings gap, and performance self-evaluation as a driver of these differences. We then evaluate actionable interventions that can be practically utilized by online platforms to mitigate the gender earnings gap and enhance earnings prospects for all workers.

We conduct a two-phase experiment on Prolific, an online platform that presents work opportunities in the form of micro-tasks and is gender-blind in recruitment. In addition to its practical relevance as a micro-task platform, an advantage of Prolific from an experimental design perspective is that it allows us to isolate the impact of gender differences in workers’ self-evaluation on task selection, while minimizing the effect of other demand-

side confounding factors, such as actual and/or anticipated discrimination from employers/customers. In our experiments, we present workers with basic math grading tasks of varying difficulty levels (low corresponding to grade 3 and high corresponding to grade 5, respectively). We find that women are less likely to choose the more difficult task compared to men of comparable ability, despite a higher pay for the same expected amount of task completion time. This contributes to a gender earnings gap, even after accounting for other sources of heterogeneity among workers such as demographics, online work experiences, and actual performance in the task. We find these observations to be driven by women exhibiting a more negative self-evaluation of their performance, particularly for the more difficult task, while other explanations such as risk aversion can be ruled out.

Online platforms present new and practical opportunities for tackling the gender earnings gap. We evaluate the potential of two interventions, task-level performance feedback and exposure to more difficult tasks through task assignment, in reducing gender disparities in task selection and in earnings. These findings are pivotal in guiding platform designs and policies toward fostering a more inclusive and equitable online marketplace.

Performance feedback is a common tool used by labor platforms to provide quality signals to employers and improve the matching efficiency, as well as for their internal evaluation. Concerns with rating inflation and workers' gaming behavior have led some platforms to delay sharing the employer's performance feedback with workers or not revealing the feedback to them at all (Rahman, 2021; Upwork, 2024b). However, our findings point to a potential caveat of these changes. We find that providing unambiguous performance feedback to workers reduces the gender gap in task selection and earnings. For workers, performance feedback acts as an additional performance indicator, alleviating the salience of workers' self-evaluation in task selection and thus narrowing the gender earnings gap. This is particularly evident for women who worked on the more difficult task in the first phase, as they tend to exhibit a more negative self-evaluation. The impact of performance feedback alone as an intervention, however, is limited, because many women in our exper-

iments chose to work on the easier task in the first phase, for which they have a relatively unbiased performance self-evaluation. For them, performance feedback alone may not be sufficiently motivating to pursue the more difficult task later on.

As a second intervention, we considered an early exposure of all workers to difficult tasks through task assignments. Task assignment is a widely-adopted strategy by platforms to match customers with service providers. In practice, task assignment can be automatic or operationalized in less stringent forms through presenting tasks in a particular sequence to service providers (e.g., DoorDash and Uber) or sending notifications for jobs satisfying prespecified criteria (e.g., Upwork) (Shi, 2023) to influence workers' choices. In our experiments, in the absence of task assignments, women tend to choose the easier task in both phases. However, when all workers are exposed to the more difficult but higher-paying task in the earlier phase, it encourages women to break free from the pattern of picking easier tasks and instead to pursue difficult and higher-paying tasks later on. This observation corroborates the perils of algorithms that match workers with jobs based on their likelihood to respond or apply, as they may exacerbate gender gaps due to differences in the historical patterns of job-seeking behavior between genders (Wall & Schellmann, 2021). Working on the more difficult task in the earlier phase can reduce workers' perceived uncertainty and increase familiarity, and is particularly influential for women, who exhibit a negatively-biased self-evaluation regarding challenging tasks.

We also observe that combining early exposure to difficult tasks and providing performance feedback is the most effective strategy in mitigating the gender disparity in task selection and improving workers' overall welfare. This dual approach not only helps lessen the impact of self-evaluation biases, particularly among women, but it also significantly boosts women's hourly earnings without adversely affecting men's earnings.

Our study has certain limitations that can be addressed in future research. First, our experiment was conducted on Prolific using a math grading task. While we expect the behavioral mechanism explored in this paper to be relevant in other online micro-task plat-

forms, particularly for those high-paying and challenging tasks such as programming and data analysis¹⁶, we encourage further investigations into other task types and platforms to assess the generalizability of our results. For instance, it will be interesting to examine whether women shy away from choosing lucrative tasks traditionally dominated by females, such as editors and accountants. Another interesting extension is to compare the effects of performance self-evaluation against other potentially influential factors such as discrimination or its anticipation on other online labor platforms where gender is revealed or can be inferred in the hiring process. Second, we employed a binary performance feedback as an initial exploration of potential mitigation strategies on platforms. However, binary feedback might not be suitable for tasks where assessment is more subjective, such as writing and logo design. Future studies could examine the effects of other feedback formats (e.g., detailed performance reviews), varying levels of feedback granularity, or alternative types of interventions. By identifying performance self-evaluation bias as an underlying mechanism and evaluating the effectiveness of performance feedback and task exposure, our study lays the groundwork for further exploration of interventions and their potential to close the gender disparities in online labor platforms.

¹⁶<https://www.upwork.com/resources/highest-paying-freelance-jobs>

CHAPTER 3
BEYOND HEALTH RISKS: THE IMPACT OF COVID-19 ON LABOR
VIOLATIONS IN GLOBAL SUPPLY CHAINS

3.1 Introduction

Since the outbreak of the COVID-19 pandemic in March 2020, organizations such as the International Labor Organization (ILO) have been actively evaluating and addressing the threat posed by the COVID-19 pandemic to global labor practices (Berg et al., 2021; ILO, 2020). For instance, the ILO's International Program on the Elimination of Child Labour and Forced Labour (IPEC+) has allocated additional funding to support the research that assesses the impact of the pandemic on child labor and forced labor. It also developed business continuity plans to mitigate the income shocks to vulnerable populations. Such initiatives call for more comprehensive investigations of the impact of COVID-19 that extends beyond immediate health risks.

The effect of COVID-19 on labor violation incidents in global supply chains remains an empirical question to be explored. On the one hand, the economic downturn resulting from business shutdowns, production slowdowns, and reduced consumer demand has led to lower demand for labor. The reduction in total working hours worldwide is persistent and estimated to be 8.8% in 2020, 4.8% in the first quarter of 2021, and 4.4% in the second quarter of 2021 (Berg et al., 2021). The decrease in labor demand potentially leads to reduced incentives for labor exploitation. Additionally, the restrictions on movement imposed by governments worldwide may have disrupted the existing trafficking systems, making it more challenging for traffickers to recruit new workers.

On the other hand, COVID-19 has created conditions which may trigger unethical supply chain practices to emerge¹ (L. Chen et al., 2022). Disruptions in production due to

¹https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---ipec/documents/publication/wcms_854733.

unexpected changes in orders have led to instability in employment. According to a survey of 376 Bangladesh garment factories in 2020 (Anner, 2022), 53% of the suppliers experienced cancellations for at least 25% of the in-process orders. Worse still, 72% of the fashion brands and retailers which canceled or suspended orders refused to pay for the cost of raw materials that were already purchased for canceled orders. Consequently, 80% of the Bangladesh garment factories surveyed had to cut employment or shut down. The resulting economic shocks such as income losses² along with school closures have increased the vulnerability of children to child labor. The ILO's 2020 child labor and forced labor report (ILO, 2020) estimates a rise in child poverty, with an additional 42-66 million children living in extreme poverty compared to 2019, making them more susceptible to exploitation. Furthermore, stricter border controls, while introduced to stem the spread of the virus, have inadvertently increased the risk of human trafficking, as individuals seeking work may resort to using irregular and unsafe migration routes³. The capacity for compliance and enforcement of labor standards may be compromised. This is because resources and attention can be diverted towards addressing the pandemic, and travel restrictions can make onsite auditing impractical, weakening the mechanisms designed to protect workers from violations⁴.

The challenges induced by COVID-19, such as supply chain disruptions, can have heterogeneous effects on firms with different supplier networks. For example, firms with more fragmented and diversified supplier bases as well as shorter-term relationships with suppliers such as apparel retailers can be more susceptible to labor violations in their supply chain. This is presumably because buyers' significant bargaining power over suppliers may prompt suppliers to cut corners and compromise on labor standards to meet buyers' demands for lower prices and faster production. Even though such unfair practices may be

pdf

²<https://www.theguardian.com/global-development/2020/dec/03/garment-workers-going-hungry-as-fall-out-from-cancelled-orders-takes-toll-report>

³https://www.unodc.org/documents/human-trafficking/SOM_and_COVID-19_Publication_final_EN_final.pdf

⁴<https://www.just-style.com/comment/social-compliance-auditing-during-and-after-covid-19/?cf-view>

exposed and catch attention in normal times, they can be justified or go unnoticed during challenging times such as the pandemic by claiming the market conditions unpredictable (L. Chen et al., 2022). Therefore, it is important to take the perspective of supply chain fairness and examine the heterogeneous impact of COVID-19 across firms with different supplier networks.

Motivated by the empirical ambiguity and to identify the firms that are more susceptible to labor violations, our study aims to address the following research questions: (1) What is the impact of the pandemic on the prevalence of worst forms of labor violations in global supply chains and whether the impact (if any) persists over time? (2) Does the impact vary across firms with different characteristics of supplier networks, including the number of suppliers, the geographic dispersion of suppliers, and the duration of relationship with the suppliers?

In our study, we leverage the country-specific COVID-19 outbreak as exogenous shocks and perform Year-to-Year Difference-in-Difference (DID) analyses (Sim et al., 2022; Han et al., 2022) to examine how the shocks impact the number of labor violation incidents in the US and Chinese buyers' global supplier network. We obtain data on four types of the most severe labor violation incidents, including child labor, forced labor, migrant workers, and human trafficking from RepRisk database. We get supply chain relationship data from FactSet and firm fundamentals from the Compustat database. Our analysis covers 2,385 globally public-listed firms and their suppliers in two different industries (apparel and electronics) and across 8 years (from 2015 to 2022). We find that the pandemic increases labor violations in global supply chains, and the impact persists over a 2- to 3-year post period. Moreover, heterogeneity analyses show that the impact is exacerbated for buying firms with more suppliers, and more diversified geographic exposure of suppliers, but is mitigated by longer-term relationships with suppliers.

Our study makes the following contributions. First, while there is anecdotal evidence⁵,

⁵<https://www.theguardian.com/global-development/2020/dec/03/garment-workers-going-hungry-as-fall-out-from-cancelled-orders-takes-toll-report>

qualitative discussions⁶, and surveys (Anner, 2022; Organization, 2017) showing the competing effects of COVID-19 on labor violations, large-scale and comprehensive empirical research remains scarce. Our study addresses this gap by examining the causal relationship between the COVID-19 outbreak and the prevalence of labor incidents in global supply chains, as well as the enduring nature of the impact. We find that the pandemic has intensified labor violations in global supply chains and the effect lasts for 2 to at least 3 years, which calls for future research and policy interventions to mitigate the negative impacts.

Second, we provide nuanced insights by identifying the firms and industries that are more susceptible to labor violations during the pandemic. Our study provides practical guidance for managers, policymakers, and industry regulators. Specifically, recognizing which industries are at higher risk for labor violations can help managers to proactively assess the company's supply chains and implement strategies to mitigate the risk of labor violations. Additionally, this information can inform targeted regulation and enforcement efforts, ensuring that resources are allocated efficiently and effectively. Ultimately, our research contributes to the ongoing efforts to promote ethical labor practices and protect workers' rights during and beyond the COVID-19 pandemic.

3.2 Literature Review

3.2.1 Antecedents of Labor Violations

Our study is related to the research that seeks to identify the factors contributing to labor practices within firms. Researchers have explored the effect of various operational and management practices. For example, based on the data from Nike factories in developing countries, researchers find that lean manufacturing adoption increases compliance with labor standards, achieved through improved labor relations (Distelhorst et al., 2017). de Zegher et al. (2021) show that low-skilled labor shortages are associated with an increase

⁶https://www.unodc.org/documents/human-trafficking/SOM_and_COVID-19_Publication_final_EN_final.pdf

in abuse reported by workers subsequently. By comparing the labor practices in a social enterprise versus a commercial company, Vanpoucke and Klassen (2023) find that management practices such as the underlying institutional logic affects a firm's willingness to act beyond minimal efforts to reduce the risks of forced labor in supply chains.

Another stream of research focuses on the effect of compliance regulations. For example, Kwon and Wu (2021) examine the role of the California Transparency in Supply Chains Act, which requires large manufacturers and retailers to disclose their efforts in eradicating slavery and human trafficking from their supply chains, in shaping corporate behavior and labor practices. They highlight the importance of legal frameworks in promoting transparency and ethical labor practices within businesses. Furthermore, Amengual and Distelhorst (2020) compares the effectiveness of a cooperative approach and a penalty approach, which threatens to discontinuing business relationships with noncompliant suppliers. They find that the penalty approach outperforms the cooperative approach in incentivizing compliance behavior of suppliers.

Recent research has emerged to investigate the effect of major supply chain disruptions, such as those caused by the COVID-19 pandemic, on labor practices. With limited availability of comprehensive data, existing research often adopts an interview (Uddin et al., 2023) or survey approach (Anner, 2022) to gather preliminary findings on the potential influence of the pandemic on occurrences of labor violations within factory settings. These studies show that buyers' order cancellations, price discount requests, and payment delays often lead to business shutdowns in the factories, which further worsen their working conditions and can further expose them to human rights abuses (Anner, 2022; Uddin et al., 2023).

Despite these valuable insights into the impact of COVID-19 on labor violations, a large-scale and comprehensive causal inference of the impact of COVID-19 on labor violations is lacking. By examining this relationship and its heterogeneity across countries, industries, and firms, our study fills the gap and allows managers and policymakers to im-

plement targeted strategies and regulations to mitigate these risks and promote ethical labor practices.

3.2.2 The Impact of COVID-19 on Operations

The COVID-19 pandemic has significantly influenced operations at both firm and individual levels across various sectors. A growing body of literature has emerged to study the challenges faced by businesses during this period. Delasay et al. (2022) investigate the impact of COVID-19 on retail operations, highlighting the importance of leveraging digital platforms, such as online shopping and contactless payments, to adapt to the changing consumer behavior and maintain business resilience during the crisis. Similarly, Han et al. (2022) emphasize the challenges faced by brick-and-mortar stores due to lockdowns, social distancing, and shifting customer preferences. Their findings underscore the need for retailers to enhance logistics capacity, such as home delivery services, to stay competitive in the rapidly evolving market landscape.

Additionally, at the individual level, researchers have investigated the impact on online gig workers' labor supply decisions (Cao et al., 2022), research productivity (Cui et al., 2022), and music consumption (Sim et al., 2022). For example, Cao et al. (2022) investigate how online gig workers changed their behavior after the COVID-19 outbreak, revealing that unemployment and nonpharmaceutical interventions such as shelter-in-place orders and closure of daycares and restaurants drove the increase in gig economy labor supply. Cui et al. (2022) examine the impact on research productivity and find that the pandemic disrupted traditional research activities and collaboration patterns, leading to an overall decline in research output. Moreover, female researchers' productivity dropped by a greater magnitude relative to that of male researchers.

Our study complements this growing body of literature by focusing on the impact of COVID-19 on labor violation incidents and examining its enduring nature, which may not have been well understood.

3.3 Method

3.3.1 Data

Labor Violation Incidents

We obtain the data on labor violation incidents from the RepRisk database⁷, a comprehensive and reliable source widely used by researchers to identify information on environmental, social, and governance (ESG) risks associated with companies and infrastructure projects (Kwon & Wu, 2021; Li & Wu, 2020; Cen et al., 2022). RepRisk was established in 2006 and has grown to cover more than 210,000 public and private companies and 55,000 infrastructure projects of all sizes, in every sector and market. The database is updated daily and aggregates data from over 100,000 public sources, spanning 23 languages, ensuring that it captures the most relevant and up-to-date ESG risk information. This broad coverage enables us to conduct a comprehensive analysis of labor incidents across various industries in different countries. Furthermore, the database's daily updates enable us to track how firms' labor practices evolve as the pandemic unfolds.

In this study, we focus on four of the most severe labor violation categories: child labor, forced labor, migrant workers, and human trafficking. We first obtain the daily incident records associated with each firm (if any), including variables of incident date, related countries, related companies, and characteristics of the incident (e.g., severity, reach, novelty, and etc.). We then aggregate the daily incident records into the number of labor violation incidents at the quarter-firm level for further analysis.

Supply Chain Relationships

We obtain the data on supply chain relationships from FactSet database⁸. FactSet is by far the most comprehensive dataset of supply chain relationships. The database covers 33,280

⁷<https://www.reprisk.com/>

⁸<https://www.factset.com/marketplace/catalog/product/factset-supply-chain-relationships>

public firms across 125 countries. It gathers information from a variety of resources including annual and quarterly filings (Securities and Exchange Commission forms 10-K, 8-K, and 10-Q), conference call transcripts, company websites, and company press releases. FactSet is comprehensive in capturing the supply chain relationships as it includes not only the relationship disclosed by a focal firm itself, but also the relationship disclosed by its supply chain partners.

For each firm, we rely on supply chain relationships captured by FactSet to construct the supplier characteristics. Specifically, a focal firm's number of suppliers (*n_supplier*) is calculated as the total count of its suppliers in a given quarter, geographic dispersion of suppliers (*n_supplier_iso*) is calculated as the count of distinct headquarter countries for the firm's suppliers in a given quarter, and duration of relationship with the suppliers (*duration_median*) is calculated as the median duration (in years) of the firm's supplier relationships in a given quarter.

Firm Fundamentals

We obtain the data on firm fundamentals from the Compustat North America and Global database, a widely recognized source of financial, statistical, and market information on global companies. Based on Compustat, we construct the following control variables that reflect different dimensions of a firm's operations. To measure a firm's profitability, we construct Return-on-assets (ROA), calculated as operating income before depreciation divided by total assets. To measure a firm's efficiency, we construct operational slack, calculated as fixed assets divided by total revenue, as well as financial slack, calculated as total stockholders equity divided by the sum of long-term debts and debt in current liabilities.

3.3.2 Model

In this study, we analyze the impact of COVID-19 outbreak on labor violations by employing the Year-to-Year DID approach. Since almost all major countries are eventually

affected by the pandemic, we follow the literature (Sim et al., 2022; Han et al., 2022) and define comparable periods as treatment and control groups. Specifically, to account for the temporal trend of labor violation incidents, we compare the average number of violations (aggregated across buyer firms) that occurred in each quarter with those in a comparable quarter in the previous period. The comparable periods for treatment and control groups are defined relative to the shock timing specific to the country where the firm’s headquarter is located. For example, in the U.S., the outbreak was in 2020 Quarter 1 (Q1), as declared by the World Health Organization on March 11, 2020. As illustrated in Figure 1, for the control group, we define it as the period from 2015 Q1 to 2018 Q4, representing a pre-pandemic baseline. The treatment group covers the period from 2019 Q1 to 2022 Q4, capturing the period of the pandemic. The shock timing of 2020 Q1 in the treatment period corresponds to 2016 Q1 in the control period. By subtracting the difference between the pre- and post-shock period in the control group from the difference in the treatment group, we examine whether and to what extent the difference widened after the outbreak the pandemic in 2020 Q1.

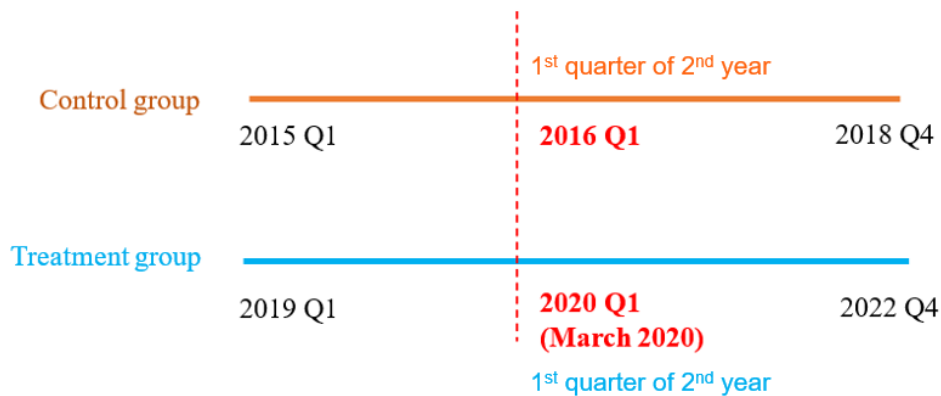


Figure 3.1: Control and Treatment Groups (the US)

The Year-to-Year DID model is described as follows:

$$\begin{aligned}
 Violations_{i,t} = & \beta_0 \times Year2019_2022_t + \beta_1 \times Year2019_2022_t \times Post_t \\
 & + \gamma_t + m_i + Controls_{i,t} + \epsilon_{i,t} \quad (3.1)
 \end{aligned}$$

where i indicates firm, and t indicates the quarters relative to the first quarter of 2020, which is the shock timing of the pandemic. $Violations_{i,t}$ is the total number of labor violation incidents associated with firm i in Year-Quarter t . $Year2019_2022_t$ is a dummy variable that equals one if the violations occurred between 2019 and 2022 (i.e., treatment group), and equals zero otherwise. $Post_t$ is a dummy variable that equals one if the violations occurred after the first quarter of the second year in the control or treatment group. γ_t represents time-fixed effects that capture unobserved time-varying factors common to all companies. m_i denotes company-fixed effects, capturing unobserved time-invariant factors specific to each company. $Controls_{i,t}$ includes the firm-year-quarter control variables, which include ROA, Operational slack, and financial slack. $\epsilon_{i,t}$ is the error term. The coefficient of interest, β_1 , captures the average treatment effect of COVID-19 on the number of labor violations.

3.4 Results

Our analysis covers 2,385 globally public-listed firms with headquarters in the US or China and their suppliers in two industries (apparel⁹ and electronics¹⁰) across 8 years from 2015 to 2022.

In this section, we present the analysis results based on the firms with headquarters in the US for brevity. Similar analyses were conducted based on Chinese firms and are

⁹The apparel industries are identified through two-digit SIC codes. We include SIC Code 23 - Apparel and other Finished Products Made from Fabrics and Similar Materials; SIC Code 56 - Apparel and Accessory Stores.

¹⁰The electronics industries are identified through two-digit SIC codes. We include SIC Code 36 - Electronic and other Electrical Equipment and Components, except Computer Equipment; SIC Code 35 - Industrial and Commercial Machinery and Computer Equipment.

included in section C.2 and C.3 of the Appendix. For an overview, we summed the number of labor incidents at the year-quarter level and computed descriptive statistics for the US firms. The minimum number of incidents in a given quarter is 5 in 2015 Q1, while the maximum is 68 in 2020 Q3. The mean number of incidents in a quarter is 23, and the standard deviation is 16. These statistics suggest that there is considerable variability in the frequency of labor incidents over time.

3.4.1 Parallel Trend Test

Before we conduct formal analyses, we plot the trend of the number of labor violation incidents averaged across all associated public-listed firms with headquarters in the US. Figure 3.2 shows that the number of labor violation incidents seems to be parallel prior to the shock.

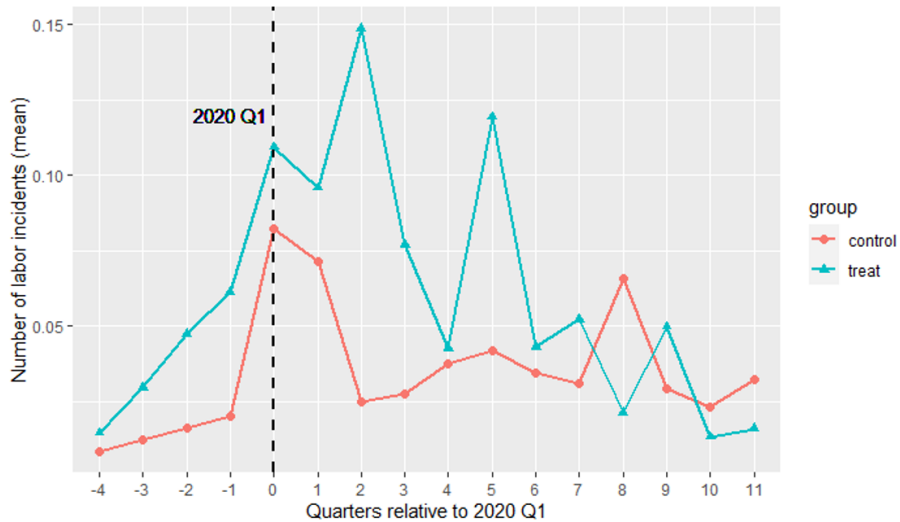


Figure 3.2: Parallel trend (companies with HQs in the US)

Next, we modify Equation 3.1 to formally test whether the parallel trend is violated.

$$\begin{aligned}
 Violations_{i,t} = & \beta_0 \times Year2019_2022_t + \beta_1 \times Year2019_2022_t \times Quarter_t \\
 & + \gamma_t + m_i + Controls_{i,t} + \epsilon_{i,t} \quad (3.2)
 \end{aligned}$$

where $Quarter_t$ is a dummy variable that equals one if the observation is in a specific year-quarter. If there is no difference in pre-trends across the treatment and control groups, the coefficient of β_1 should not be significant before 2020 Q1 (i.e., for $t < 0$), which is supported by Table 3.1.

3.4.2 Effect of COVID-19 on Labor Violations

Table 3.2 presents the results of the DID analysis, which demonstrate the positive impact of the pandemic on the number of labor violations over a three-year post-treatment period. The treatment effect remains significantly positive for up to 2 years after the treatment, as indicated by the coefficients for $Treat \times Post$ (ranging from 0.024 to 0.052 with significance levels at 10% and 1%, respectively). Notably, the treatment effect of increasing labor violations is diminishing over time, with the magnitude and significance of coefficients peaking in the 1-year post period (0.052 with $p < 0.01$). The treatment effect of increasing labor violations reduces in the 2-year post period (0.029 with $p < 0.05$) and becomes insignificant in the 3-year period after the COVID-19 outbreak.

3.4.3 Heterogeneity across Characteristics of Suppliers

As the treatment effect is the most prominent in the 1-year post period, we focus on the heterogeneity in the characteristics of suppliers over the 1-year post period in this subsection. The heterogeneity effects observed over the 2- and 3-year post periods are consistent and included in Appendix section C.1.

To further explore how the impact of COVID-19 on labor violation incidents depends on firms' supplier characteristics, we conduct moderation analyses using the number of suppliers, geographic dispersion of suppliers, and relationship duration as moderating variables, respectively. In Table 3.3 Column (1) and (2), the coefficients of the three-way interaction terms ($Treat \times Post \times n_supplier$ and $Treat \times Post \times n_supplier_iso$) are both significantly positive (0.009, significant at 1% level and 0.053, significant at 1%), sug-

Table 3.1: Parallel trend test (US buying firms)

	(1)	(2)
<i>Treat</i>	0.032 (0.020)	0.037* (0.022)
<i>Treat</i> × <i>Quarter</i> ₋₄	-0.025 (0.027)	-0.031 (0.030)
<i>Treat</i> × <i>Quarter</i> ₋₃	-0.008 (0.027)	-0.013 (0.030)
<i>Treat</i> × <i>Quarter</i> ₋₂	0.007 (0.027)	0.003 (0.030)
<i>Treat</i> × <i>Quarter</i> ₋₁	0.014 (0.027)	0.012 (0.030)
<i>Treat</i> × <i>Quarter</i> ₁	-0.003 (0.027)	0.000 (0.031)
<i>Treat</i> × <i>Quarter</i> ₂	0.097*** (0.028)	0.112*** (0.031)
<i>Treat</i> × <i>Quarter</i> ₃	0.024 (0.028)	0.020 (0.030)
<i>Treat</i> × <i>Quarter</i> ₄	-0.019 (0.028)	-0.024 (0.031)
<i>Treat</i> × <i>Quarter</i> ₅	0.055** (0.028)	0.060* (0.031)
<i>Treat</i> × <i>Quarter</i> ₆	-0.016 (0.028)	-0.021 (0.031)
<i>Treat</i> × <i>Quarter</i> ₇	-0.005 (0.028)	-0.008 (0.031)
<i>Treat</i> × <i>Quarter</i> ₈	-0.069** (0.028)	-0.078** (0.031)
<i>Treat</i> × <i>Quarter</i> ₉	-0.007 (0.028)	-0.010 (0.031)
<i>Treat</i> × <i>Quarter</i> ₁₀	-0.037 (0.028)	-0.042 (0.031)
<i>Treat</i> × <i>Quarter</i> ₁₁	-0.030 (0.029)	-0.034 (0.032)
<i>Controls</i> _{<i>i,t</i>}		Included
Year-Quarter FE	Yes	Yes
Firm FE	Yes	Yes
N	15,610	14,227
R-squared	0.013	0.014

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

Table 3.2: DID analysis results (US buying firms)

	1-year post period		2-year post period		3-year post period	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Treat</i>	0.031*** (0.011)	0.035*** (0.013)	0.031*** (0.010)	0.033*** (0.011)	0.028*** (0.009)	0.030*** (0.010)
<i>Treat</i> × <i>Post</i>	0.044** (0.017)	0.053*** (0.020)	0.024* (0.013)	0.028** (0.014)	0.004 (0.011)	0.006 (0.012)
<i>Controls</i> _{<i>i,t</i>}		Included		Included		Included
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
N	8,096	7,155	11,946	10,608	15,610	13,956
R-squared	0.012	0.013	0.011	0.012	0.009	0.010

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

gesting that the increase in labor violations is more pronounced for the US firms with more suppliers and more diversified geographic exposure of suppliers. Conversely, in Column (3), the coefficient of $Treat \times Post \times duration_median$ is negative and marginally significant ($p < 0.1$). This shows that the increase in labor violations is mitigated for the US firms with longer-term relationship with suppliers.

3.5 Concluding Remarks

Many global organizations such as the ILO have called for an in-depth understanding of the impacts of COVID-19 on labor practices globally beyond the immediate health risks. Responding to these calls, our study examines the impact of the pandemic on the prevalence of worst forms of labor violations in global supply chains. We conduct Year-to-year DID analyses and find that the pandemic increases the labor violation incidents associated with child labor, forced labor, migrant workers, and human trafficking in global supply chains. The effect persists for two to three years after the COVID-19 outbreak. Furthermore, firms with more suppliers and more diversified geographic exposure of suppliers are more susceptible to such violations, while firms with longer relationships with suppliers mitigate the impact.

Table 3.3: Heterogeneity Analysis over 1-year post period (US buying firms)

	(1)	(2)	(3)
	n_supplier	n_supplier_iso	duration_median
<i>Treat</i>	-0.001 (0.021)	0.054 (0.027)	-0.050 (0.036)
<i>n_supplier</i>	0.014*** (0.001)	0.016*** (0.001)	0.017*** (0.001)
<i>n_supplier_iso</i>	-0.035*** (0.005)	-0.033*** (0.005)	0.008*** (0.005)
<i>duration_median</i>	0.000*** (0.006)	0.000 (0.007)	0.001 (0.008)
<i>Treat × Post</i>	-0.090*** (0.029)	-0.192*** (0.037)	0.166*** (0.050)
<i>Post × n_supplier</i>	-0.000 (0.001)		
<i>Treat × n_supplier</i>	-0.001 (0.000)		
<i>Treat × Post × n_supplier</i>	0.009*** (0.001)		
<i>Post × n_supplier_iso</i>		-0.004 (0.004)	
<i>Treat × n_supplier_iso</i>		-0.016*** (0.004)	
<i>Treat × Post × n_supplier_iso</i>		0.053*** (0.005)	
<i>Post × duration_median</i>			0.001 (0.009)
<i>Treat × duration_median</i>			0.005 (0.010)
<i>Treat × Post × duration_median</i>			-0.024* (0.014)
<i>Controls_{i,t}</i>	Included	Included	Included
Year-Quarter FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
N	4619	4619	4619
<i>R</i> ²	0.254	0.205	0.156

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

Our study provides nuanced insights into the moderating role of firm-level dynamics such as supplier characteristics. We show that supply chain complexity increases firms' risk profile regarding labor violations in the context of the COVID-19 pandemic. Specifically, firms engaging with an extensive network of suppliers, particularly those spread across vast geographic regions, can face heightened challenges in auditing. Given the expanded scope of supplier networks, there are naturally limited resources and capacity to conduct thorough audits for each supplier to ensure compliance. Moreover, it can be difficult to build trust and effectively communicate given the cultural and language barriers across national borders. These barriers may also impede the enforcement of compliance standards and resolving any issues that may arise. Lastly, a diversified supplier pool increases the likelihood of sourcing from regions where the regulations are less stringent, potentially increasing the risks of not adhering to labor standards. On the other hand, the increased risks of labor violations can be mitigated with long-standing supplier relationships. Firms forming long-term relationships with their suppliers are less likely to have labor violations in their global supply chains. This mitigated effect is possibly a result of open communications, mutual trust, and collective efforts to uphold labor standards. These long-term supplier relationships function not merely as transactional engagements but as strategic alliances that contribute to the overall resilience of the supply chain.

3.6 Future Steps

First, we plan to further investigate why there is an improvement (i.e., reduction in labor violations) associated with US firms, but not Chinese firms (as shown in Appendix section C.2), especially in the third year after the pandemic. Do the US buying firms shift to more compliant suppliers, or they are actively improving labor conditions in their supply chain? We can shed light on this question by analyzing whether there are changes in the buying firms' supplier base, and if so, how the new suppliers differ from the original suppliers in terms of their labor practices.

Second, we plan to consider other moderators to better understand the heterogeneous effects of the pandemic and provide more nuanced managerial implications. For example, we may consider other supply chain network characteristics such as power asymmetry. We may also compare industries with different demand shocks such as the personal protective equipment (PPE) industry where experienced a demand surge and the apparel industry which experienced cancellation of orders. These analyses will help us understand how the pandemic has affected different types of business sectors and inform policy recommendations targeted at specific sectors.

Third, we plan to conduct robustness analysis to address potential concerns about the validity of our results. One concern is that after COVID-19, media may be more sensitive to labor incidents and report more of them, leading to an increase in the number of violations observed. To address this concern, we may measure and control the media exposure of industries by collecting the trend index in the Google Trends¹¹. This analysis will help us distinguish between the effects of COVID-19 and media attention on labor violation incidents. We may also be able to indirectly alleviate the concern by testing whether firms in industries that should expect greater media attention had more violations, such as those producing products with surging demand.

¹¹<https://trends.google.com/trends/>

Appendices

APPENDIX A
TRIPS FOR TIPS? IMPLICATIONS OF TIPS ON DRIVERS' SEARCH
BEHAVIOR

A.1 Data pre-processing steps

We further removed trips records which are likely to be generated due to system errors or drivers' misreports. We removed trips with:

- driver-reported passengers more than 5 or fewer than 1;
- distance shorter than 1 mile;
- identical pickup and drop-off longitude and latitude;
- negative fare amount;
- total fare amount not equal to the sum of all itemized fares;
- time duration shorter than 2 minutes or longer than 180 minutes;
- drop-off time earlier than pickup time;
- average speed lower than 2 MPH or higher than 70 MPH;
- tip percentage exceeds 200%.

After the above cleaning process, about 91.62% of the trip observations remained in our data set.

A.2 Machine learning prediction about cash tips

A common challenge in the research of tips is that we only observe customers' tips paid with credit cards (which we call *credit card tips*), but not those paid with cash (which we call *cash tips*) (Alexander et al., 2021; Haggag & Paci, 2014). In our specific research context, every trip taken by a driver is likely to contribute to the learning process and should be taken into account. Therefore, unlike previous research, we choose not to remove the trips paid with cash (which account for about 70% of the total trips in our dataset), but instead use the machine learning approach to predict the cash tip amount (\$).

As explained in subsection 1.6.1, our analysis focuses on a randomly sampled 1,044 (about 5%) experienced drivers and their respective (1,065,591) trips from April 1st to July 31st for tractability, and our goal is to predict the cash tip amount received by this sample of drivers. Since we only observe the tips from trips paid with the credit card, we use these trips as the training-validation set. Specifically, we use the remaining 95% of the drivers' (5,550,705) trips paid with credit card from April 1st to July 31st as the training-validation set. We shuffle these trips and split them evenly to implement the tenfold cross-validation. We first use the data in nine buckets to train the model, and then validate the model's out-of-sample predictive performance in the remaining bucket. We iterate the aforementioned process for 10 times, each time reserving a different tenth for validation.

We consider 39 features including customer-specific characteristics (e.g., pickup and drop-off GPS coordinates and time, and the number of passengers), trip-related characteristics (e.g., base fare, surcharge, MTA tax, toll amount, distance, duration, speed, and direction), and market conditions (e.g., weather conditions, and holidays). All features are listed in Table A.1.

To select the model with the best predictive performance, we employ the following metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Root Mean Square Logarithmic Error (RMSLE), Mean Absolute Percent-

Table A.1: Definitions of features

Type	Feature	Measure
Trip-related attributes	shift_pickup	The shift index associated with the pickup
	shift_dropoff	The shift index associated with the dropoff
	is_holiday_pickup	A dummy variable indicating whether the pickup occurred during holidays
	is_holiday_dropoff	A dummy variable indicating whether the dropoff occurred during holidays
	is_weekend_pickup	A dummy variable indicating whether the pickup occurred during weekends
	is_weekend_dropoff	A dummy variable indicating whether the dropoff occurred during weekends
	dayofweek_pickup	A categorical variable indicating the day of week when the pickup occurred
	dayofweek_dropoff	A categorical variable indicating the day of week when the dropoff occurred
	direction_NS	A categorical variable indicating whether the trip is from north to south or south to north
	direction_EW	A categorical variable indicating whether the trip is from east to west or west to east
	with_tolls	A dummy variable indicating whether tolls are attached
	with_add_fee	A dummy variable indicating whether extra charges are attached
	trip_distance	The distance of the trip (in miles)
	minute_pickup	The minute of the pickup time
	hour_pickup	The hour of the pickup time
	day_pickup	The day of the pickup time
	month_pickup	The month of the pickup time
	minute_dropoff	The minute of the dropoff time
	hour_dropoff	The hour of the dropoff time
	day_dropoff	The day of the dropoff time
	month_dropoff	The month of the dropoff time
	start_lon	The longitude of the pickup location
	start_lat	The latitude of the pickup location
	end_lon	The longitude of the dropoff location
	end_lat	The latitude of the dropoff location
	fare_amt	The dollar amount of base fare
	surcharge	The dollar amount of surcharge
	mta_tax	The dollar amount of tax
	tolls_amt	The dollar amount of tolls
	tot_amt_before_tips	The total amount excluding tips
	trip_duration	The duration of the trip
	Speed	The average speed of the trip
vendor_name	A categorical variable indicating the name of the credit card machine vendor	
p_uwr_code	A categorical variable indicating the name of the pickup location-time set	
d_uwr_code	A categorical variable indicating the name of the dropoff location-time set	
customer-specific attributes	pickup_income	The median household income in the pickup census tract
	passenger_count	The number of passengers in the car
Market conditions	fog_haze	An ordinal variable of fog and haze intensity
	PRCP	Daily precipitation volume

age Error (MAPE), and R-squared (R^2). Among the 18 models evaluated, we select Light Gradient Boosting Machine (LightGBM) as our predictive model since it has the best average out-of-sample predictive performance in terms of the aforementioned metrics. The out-of-sample predictive performance evaluated across these models are summarized in Table A.2. We further tune the performance of the selected LightGBM model by calibrating the hyper-parameters based on its predictive performance over the 10 out-of-sample validation buckets.

Table A.2: Out-of-sample predictive performance

Model	MAE	MSE	RMSE	RMSLE	MAPE	R^2
Extreme Gradient Boosting	0.535	0.666	0.816	0.277	0.448	0.272
Gradient Boosting Regressor	0.540	0.668	0.817	0.277	0.451	0.269
Bayesian Ridge	0.544	0.675	0.822	0.279	0.451	0.261
Linear Regression	0.544	0.676	0.822	0.279	0.451	0.261
Orthogonal Matching Pursuit	0.545	0.676	0.822	0.279	0.451	0.260
Ridge Regression	0.544	0.677	0.823	0.279	0.451	0.260
Random Forest Regressor	0.543	0.679	0.824	0.280	0.455	0.257
Extra Trees Regressor	0.550	0.693	0.833	0.283	0.459	0.242
Huber Regressor	0.572	0.734	0.857	0.286	0.463	0.197
Lasso Regression	0.579	0.745	0.863	0.289	0.475	0.185
Elastic Net	0.579	0.745	0.863	0.289	0.475	0.185
K Neighbors Regressor	0.621	0.845	0.919	0.307	0.495	0.076
Lasso Least Angle Regression	0.664	0.914	0.956	0.316	0.537	-0.000
Dummy Regressor	0.664	0.914	0.956	0.316	0.537	-0.000
Passive Aggressive Regressor	0.661	0.925	0.951	0.328	0.479	-0.011
Decision Tree Regressor	0.765	1.445	1.202	0.401	0.584	-0.581
AdaBoost Regressor	4.529	27.294	5.117	1.014	3.407	-28.956
Light Gradient Boosting Machine	0.536	0.664	0.815	0.277	0.448	0.273

Note: Boldface indicates the best-performing model. The unit of the errors is \$. The performance reported here are averaged across 10 validations.

Finally, we apply the calibrated LightGBM model to predict the cash tip amount (from 772,995 trips) received by our sample of experienced drivers between April and July. Since we do not observe the actual cash tips, we use the credit card tips (from 292,596 trips) received by the same 5% of the drivers between April and July as the testing set to gauge the final predictive performance on cash tips. Specifically, we hold out the 5% of the drivers' credit card transactions and apply the calibrated LightGBM model to predict the credit card tips. The final prediction performance is: MAE = 0.534, MSE = 0.663, RMSE

= 0.814, RMSLE = 0.276, MAPE = 0.443, and $R^2 = 0.274$ (Unit: \$).

A.3 Variations in attributes observed in the data

Since the focus of our study is on learning about tip percentages, we illustrate the spatial and temporal variations in tip percentages in the data. Similar variations are found in other five attribute realizations.

In the figure below, we plot the average and standard deviation of tip percentages associated with the pickups from different neighborhoods in Manhattan and during three time periods (early morning 3-5 am, daytime 8-10 am, and night 9pm -12 am). Average tip percentages range from 15% to 22%, and the standard deviation ranges from 7% to 15%. Specifically, we find that tip percentages tend to be higher (reddish dots) in the Midtown Manhattan area where tourist spots and business districts are located than in other areas. Tip percentages also vary across different times of a day. Tip percentages tend to be higher in early mornings between 3 and 5 a.m., with greater variations within the neighborhood (reddish and bigger dots). Tip percentages tend to be lower during weekday mornings between 8 and 10 am, with smaller variations within the neighborhood (greenish and smaller dots).

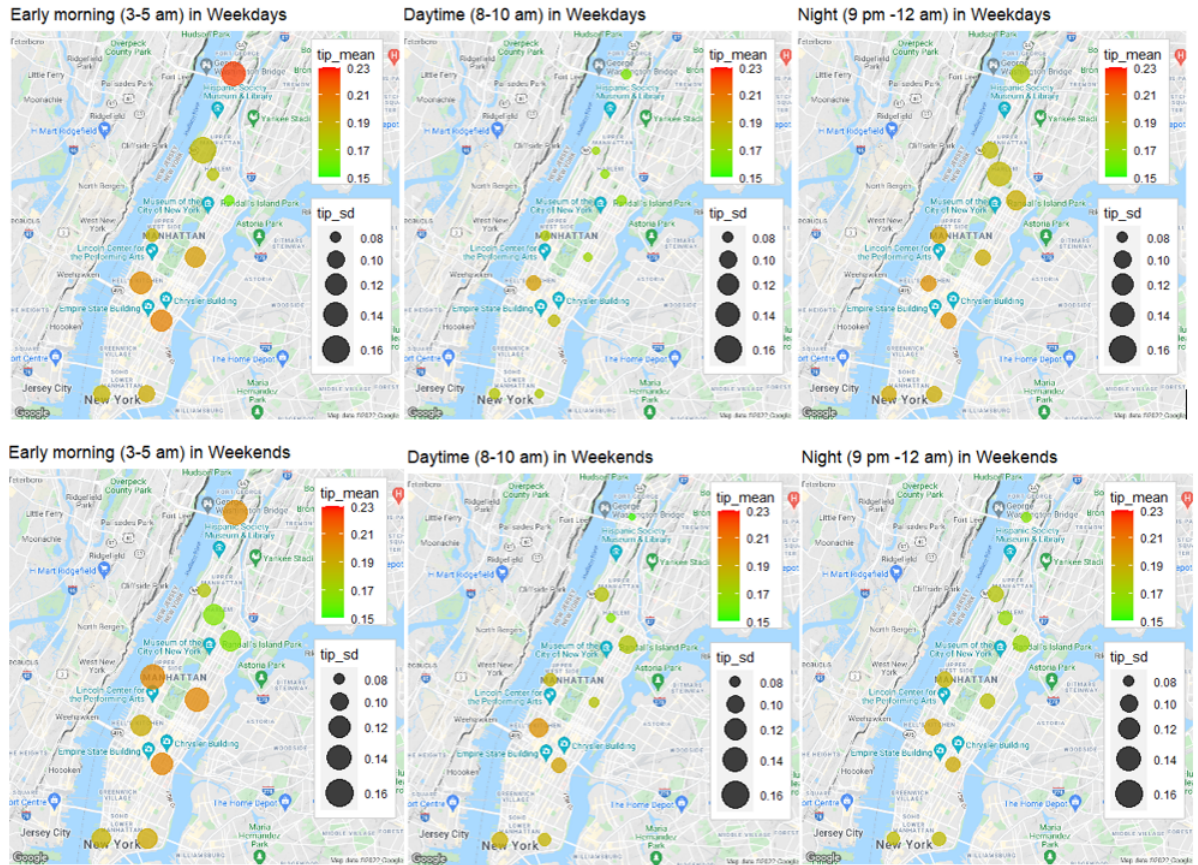


Figure A.1: Temporal and spatial variations in tip percentages

A.4 Empirical evidence of learning

As robustness checks, we first adopt an alternative specification to measure drivers' cumulative experiences. We set the time window $m = 5$ when calculating the term $\bar{X}_{il} = \frac{1}{m} \times \sum_{\tau=t-m+1}^t X_{il\tau}$, which is the moving average over 5 periods prior to the relocation at t that measures driver i 's prior experience. Additionally, we focus on new drivers' relocation behavior in their first shift, when the relocation decisions are mainly driven by intrinsic preferences rather than learning. Specifically, we identify 1,044 full-time new drivers who did not drive between January and March 2009, but started their first shift after March. As such, we are able to observe these drivers' first shift occurred after March, 2009. We conduct the multinomial logit regression analysis following the procedure described in sec-

tion 1.4. The robustness analysis results are shown in Table A.3. Overall, we find consistent evidence suggesting drivers' history-dependent relocation behavior.

Table A.3: Multinomial logit regression results - moving averages over 5 periods

	Dependent variable: relocation decisions	
	(1) $m = 5$	(2) New drivers
Moving average of tip percent	2.773*** (0.070)	6.401*** (0.177)
Moving average of fare	0.010*** (0.001)	0.024*** (0.002)
Moving average of speed	-0.019*** (0.001)	-0.011*** (0.001)
Moving average of the number of pickups	-0.529*** (0.003)	-0.625*** (0.004)
Moving average of the number of dropoffs	0.275*** (0.002)	0.257*** (0.003)
Moving average of wait time	-0.000*** (0.000)	-0.004*** (0.000)
Location-specific intercepts	Included	Included
Weekday indicator	Included	Included
Rush hour indicator	Included	Included
Driver cluster indicator	Included	Included
-LL	690,003	822,579
AIC	1,380,599	1,645,610
BIC	1,383,839	1,648,116

Note: Independent variables are the moving averages calculated over 5 periods prior to the relocation. Standard errors are in the parentheses. **** indicates significance at 1%.

A.5 Box-Cox transformation of attributes

Consistent with (Erdem et al., 2008), we adopt a Box-Cox transformation on all the six attributes to bring their distribution closer to normality. Since the Box-Cox likelihood is not well behaved when the dependent variable is zero, we use $1 + A_{ijt}$ rather than A_{ijt} for the transformation, which is given by:

$$[(1 + A_{ijt})^\lambda - 1]/\lambda = A_{ijt}^{(\lambda)},$$

where A_{ijt} is the attribute realization in the original scale, λ is the Box-Cox parameter, whose optimal value is the one that results in the best approximation of a Normal distribution curve (Box & Cox, 1964), and $A_{ijt}^{(\lambda)}$ is the transformed attribute realizations. The distribution of the transformed attribute realizations is as below.

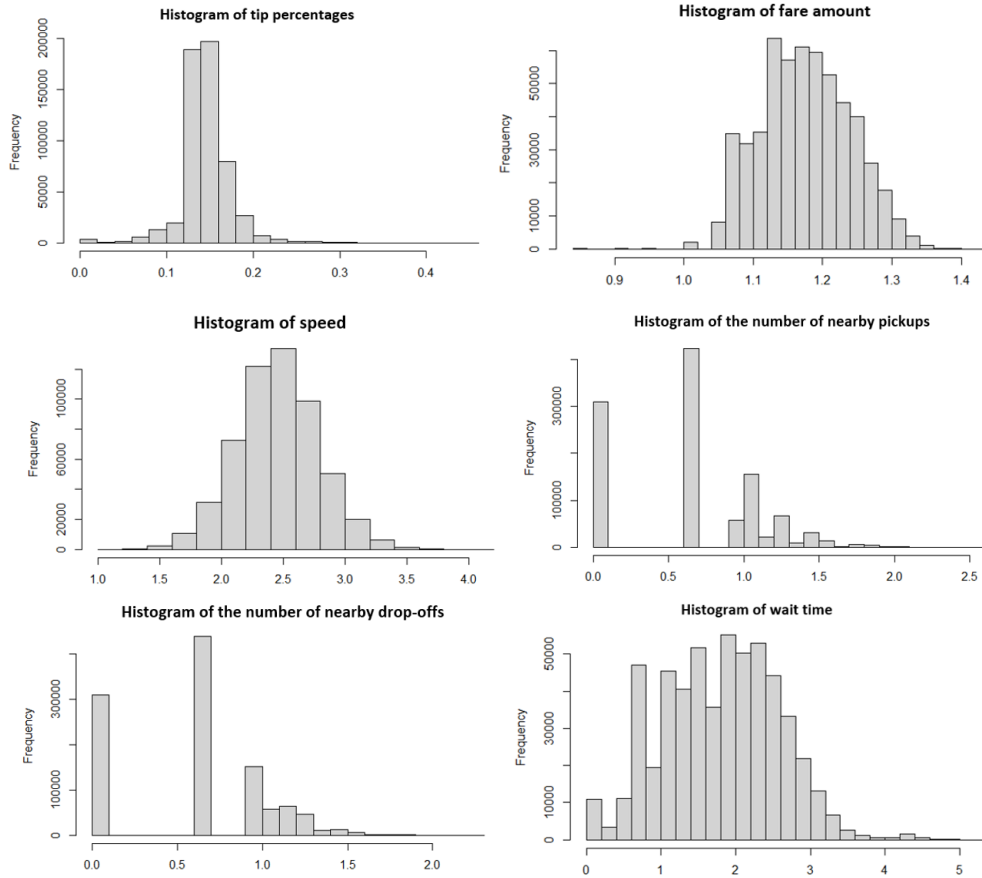


Figure A.2: The distribution of Box-Cox transformed attribute realizations

A.6 The calculation of the effect size of tip vs. fare

In Location 8 (PUMA name: Murray Hill, Gramercy & Stuyvesant Town), the average tip percentage observed in the data is 19.3%, the closest to the average tip percentages among all locations (19.1%). The average fare amount earned in Location 8 is \$8.526. We calculate the change in the relative odds given one dollar increase in tip amount on relocation probability following the steps below:

- Calculate the average tip amount observed in the data is \$1.626;
- The increased tip amount is $\$1.626 + 1 = \2.626 ;
- The corresponding increase in tip percentage is from $\frac{1.626}{8.526} = 19.1\%$ to $\frac{2.626}{8.526} = 30.8\%$

;

- The corresponding increase in tip percentage after applying the Box-Cox transformation using the optimal parameter found is from 14.8% to 20.8%;
- Calculate the odds ratio given one dollar increase in tip as $\exp((0.208 - 0.148) \times 5.645) = 1.403$, which implies that one dollar increase in the tip amount in Location 8 increases the odds of relocating to Location 8 by 40%, holding all else equal.

Similarly, we calculate the change in the relative odds given one dollar increase in fare amount following the steps as below:

- The increased fare amount is $\$8.526 + 1 = \9.526 ;
- The corresponding increase in fare amount after applying the Box-Cox transformation using the optimal parameter found is from \$1.127 to \$1.147;
- Calculate the odds ratio given one dollar increase in fare as $\exp((1.147 - 1.127) \times 0.186) = 1.004$, which implies that one dollar increase in fare amount in Location 8 increases the odds of relocating to Location 8 by 0.4%, holding all else equal.

A.7 Estimated coefficients of spatial control variables and drivers' prior belief parameters

Table A.4 presents the coefficients of spatial control variables, which show interesting patterns. In Table A.4, higher coefficients indicate that drivers have higher intrinsic preferences for certain locations. Consistent with our expectation, drivers prefer to relocate to more popular areas with higher earning opportunities such as the business districts in lower Manhattan, but are less likely to relocate to the residential areas in upper Manhattan which can have lower earning opportunities.

Table A.4: Estimates of the location-specific dummy variable coefficients

PUMA name	Dummy variable coefficients (γ)
Washington Heights, Inwood & Marble Hill	1 (fixed)
Central Harlem	1.286*** (0.051)
East Harlem	2.408*** (0.040)
Hamilton Heights, Manhattanville & West Harlem	2.670*** (0.039)
Chinatown & Lower East Side	3.313*** (0.036)
Upper West Side & West Side	3.732*** (0.035)
Battery Park City, Greenwich Village & Soho	3.780*** (0.035)
Murray Hill, Gramercy & Stuyvesant Town	3.782*** (0.037)
Upper East Side	3.821*** (0.036)
Chelsea, Clinton & Midtown Business District	3.828*** (0.036)

Note: The standard errors are in the parenthesis. “***” indicates significance at 1%.

Table A.5 compares the distribution of drivers' imputed prior beliefs about the averages of attributes with their actual averages. As explained in subsection 1.6.1, we use the individual-location-time-specific posteriors imputed at the end of the pre-estimation period (April-May) as the prior means and variances at the beginning of the estimation period (June-July), over which the utility coefficients are estimated. As such, we account for

drivers' heterogeneous prior beliefs about the attribute levels in different location-time sets depending on their past experiences.

Table A.5: Comparison of the imputed prior belief parameters and the actual value of the attribute mean

	Prior mean (μ_{B_0})	Prior variance ($\sigma_{B_0}^2$)	Actual mean (μ_{A_j})
Tip percentage	0.107 (0.029)	0.038 (0.043)	0.142 (0.030)
Fare	1.032 (0.060)	0.093 (0.107)	1.140 (0.066)
Speed	2.990 (0.432)	0.004 (0.002)	2.697 (0.313)
Number of pickups	0.074 (0.093)	0.001 (0.000)	0.550 (0.428)
Number of dropoffs	0.444 (0.144)	0.468 (0.522)	0.479 (0.392)
Wait time	0.335 (0.319)	0.010 (0.002)	1.908 (0.654)

Note: The prior means and variances reported here are averaged among drivers and across location-time sets for brevity. The standard deviations across location-times are in the parenthesis.

In Table A.5, for each attribute, the distribution parameters of drivers' prior beliefs are averaged among drivers and then across location-time sets. We find that, on average, drivers underestimate attributes associated with earning opportunities such as tip percentages, fare amount, and number of nearby pickups. At the same time, they also underestimate the search costs, indicated by higher prior beliefs about the speed and low prior beliefs about the wait time.

A.8 Implementation details about the model comparison

We assess the model fit using both in-sample and out-of-sample measures. First, we divide the June-July data (used for main analysis) into the training set (Day 1-43 which account for 70% of the data) and the test set (Day 44-61 which account for 30% of the data).

We then estimate both the Bayesian learning model and the rational expectations model using the training set. In the Bayesian learning model, drivers' beliefs are updated following a Bayesian manner, while in the rational expectations model, we assume that the true distribution of all attributes is common knowledge to all drivers and thus set all drivers' beliefs of attributes as the true distribution parameters. Based on the training sample, we calculated the in-sample negative log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) for each model.

Next, we used the parameter estimates obtained from the training set to calculate the drivers' probability of relocation on the test set for both models. Given the relocation probabilities, we computed the out-of-sample negative log-likelihood, AIC, and BIC for both models using the testing sample.

A.9 More details about the simulation procedures

A.9.1 Drivers' prior beliefs

Drivers' prior beliefs about tips reflect their existing knowledge of the tip distribution across different locations and time periods. In line with our estimation, we assume that drivers have heterogeneous prior beliefs about the tip distribution across different location-time sets, depending on their trajectory in the past. We draw drivers' prior beliefs about tips from their posterior tip beliefs imputed at the end of the pre-estimation period (explained in subsection 1.6.1), which differ both among drivers and across location-time sets.

A.9.2 The generation process of attribute realizations

Since the average speed of a trip is related to the congestion level in both pickup and drop-off location-time set, we model speed attribute realization as a function of the number of simulated total pickups and drop-offs in the pickup and drop-off location-time sets, and control for location and time dummies.

$$\begin{aligned} Speed_{ijt} = & c_0 + c_1 \times total_pick_at_pick_loc_{ijt} + c_2 \times total_drop_at_pick_loc_{ijt} \\ & + c_3 \times total_pick_at_drop_loc_{ijt} + c_4 \times total_drop_at_drop_loc_{ijt} \\ & + c_5 \times pick_loc_{ijt} + c_6 \times drop_loc_{ijt} + c_7 \times pick_hour_{ijt} \\ & + c_8 \times drop_hour_{ijt} + \epsilon_{ijt}, \end{aligned}$$

where c_0 is the intercept, $total_pick_at_pick_loc_{ijt}$ is the total number of pickups observed by driver i in the pickup location-time set j at time t , $total_drop_at_pick_loc_{ijt}$ is the total number of drop-offs observed by driver i in the pickup location-time set j at time t , and $total_pick_at_drop_loc_{ijt}$ and $total_drop_at_drop_loc_{ijt}$ are defined similarly. Location-time set dummy variables ($pick_loc_{ijt}$, $drop_loc_{ijt}$) and hour dummy variables ($pick_hour_{ijt}$, $drop_hour_{ijt}$) are also included in the model.

As explained in subsection 1.6.3, at the end of each trip, drivers' wait time after the current drop-off will not be realized until the next pickup occurs. Therefore, when drivers make the relocation decision at the end of the current trip, their beliefs about the wait time are updated with the most recent observation from the previous drop-off location-time set. The lagged wait time attribute realization is modeled as a function of the supply and demand mismatches (i.e., the demand-to-supply ratio ¹) in the drop-off location-time set of the previous trip, location and time dummies of the previous trip, and the current pickup location.

$$\begin{aligned}
 Wait_time_{ij,t-1} = & d_0 + d_1 \times \frac{total_pick_at_drop_{ij,t-1}}{(total_drop_at_drop_{ij,t-1} + total_pick_at_drop_{ij,t-1})} + \\
 & d_2 \times drop_loc_{ij,t-1} + d_3 \times drop_hour_{ij,t-1} + \\
 & d_4 \times pick_loc_{ijt} + \epsilon_{ij,t-1}
 \end{aligned}$$

For the number of nearby pickup attributes (i.e., the number of pickups within a 150-meter radius of each pickup and drop-off point defined in section 1.3), we assume that it is proportional to the total number of pickups in a location-time set. To infer the number of nearby pickups, we (1) impute the average ratio of the number of nearby pickups to the number of total pickups in each location (i.e., PUMA location grids defined in section 1.3) from the data, and then (2) impute the simulated number of nearby pickups by multiplying the number of total simulated pickups in a location-time set with the ratio found in step (1) for that location. The drop-off attribute is calculated in a similar manner.

A.9.3 Attribute variances

Per our model setup and consistent with the previous literature (Sriram et al., 2015), attribute realizations follow a Normal distribution $A_{ijt} \sim N(\mu_{A_j}, \sigma_{A_j}^2)$ and the attribute variance $\sigma_{A_j}^2$ is known to all drivers. Recall that in the estimation, we estimate $\sigma_{A_j}^2$ from the

¹We use the number of pickups as a proxy for demand, and the number of pickups and drop-offs as a proxy for supply.

data (i.e., the actual variance of attribute realizations) since the attribute realizations received by drivers are observable to us. However, in our simulation, we cannot impute $\sigma_{A_j}^2$ at the beginning of the simulation horizon anymore because the attribute realizations are system-dependent (i.e., determined by drivers' simulated relocations) and hence not observable upfront. To address this challenge, we initialize the simulation with the observed attribute variances, run the simulation, impute the updated $\sigma_{A_j}^2$ based on the observations of the simulated attribute, and then re-run the simulation based on the updated $\sigma_{A_j}^2$. The iterative process is repeated until the assumed $\sigma_{A_j}^2$ converges to the actual $\sigma_{A_j}^2$ in the simulation.

A.9.4 Relocation decisions

Now that we have drivers' prior beliefs, attribute realizations, and attribute variances, we can impute drivers' posterior beliefs about all attributes at the end of their respective first trip in Bayesian fashion, as specified in subsection 1.5.2. Based on the posterior beliefs, we impute drivers' utilities from relocation for all reachable location-time sets (defined earlier in subsection 1.6.3) and from choosing the outside option. We then simulate drivers' relocation decisions by assigning the option with the highest utility, a sum of the imputed nominal utility and the idiosyncratic shock randomly drawn from the Type I Extreme Value distribution. If a driver chooses to relocate to one of the reachable location-time sets, the next pickup location-time set j is determined by his choice and the pickup time t is calculated by adding the simulated wait time to the previous drop-off time. Once the driver finishes relocation, the customer drop-off location and time (j, t) will be drawn from the frequency distribution of the pickup and drop-off location-time pairs observed in the data. If a driver chooses the outside option, the current shift ends and the next pickup and drop-off location and time (j, t) are drawn from the empirical frequency distribution of pickup and drop-off location-time pairs of the first trip in a shift, similar to drivers' first trips. Finally, we repeat the iterative process from subsection A.9.2 to subsection A.9.4 over the

entire simulation horizon for each driver.

APPENDIX B

REDUCING GENDER DISPARITIES IN ONLINE LABOR PLATFORMS: THE ROLE OF PERFORMANCE FEEDBACK AND TASK EXPOSURE

B.1 Comparison of Upwork and Prolific hiring process

Hiring process on Upwork: Employers search for workers and browse their personal profiles before making hiring decisions. Profiles often contain the worker's first name and a headshot, which may allow customers to make inferences about the worker's gender and race.

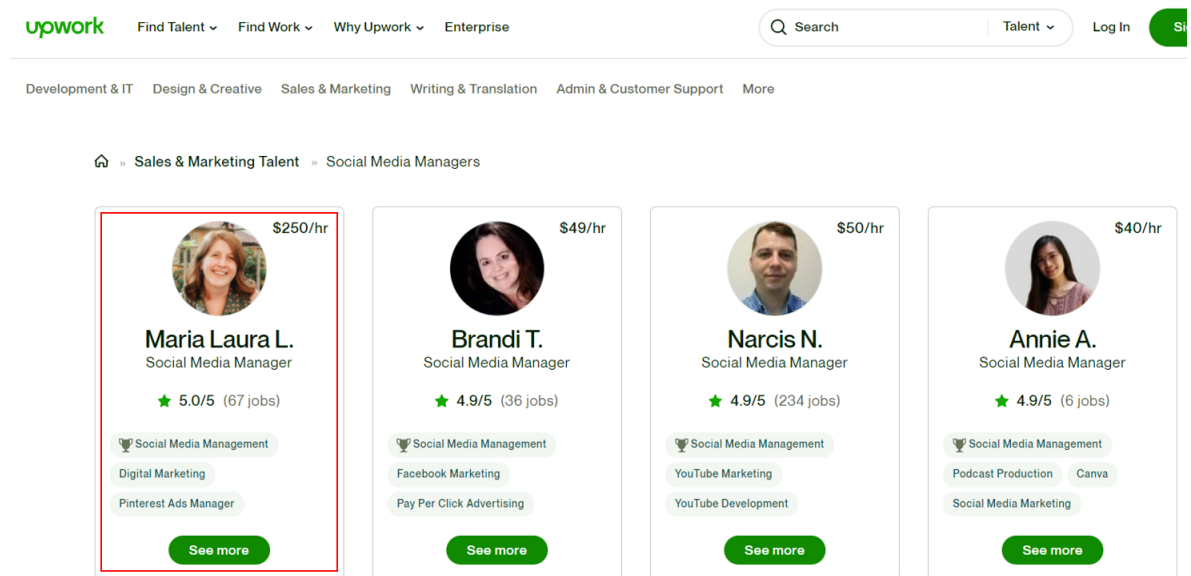


Figure B.1: User interface on Upwork

Hiring process on Prolific: Employers post a task and specify pre-screening conditions (if any) to better target a suitable worker pool. The hiring decision is not based on individual demographics.

Prolific WORKSPACES MESSAGES 62 APPS & INTEGRATIONS HELP CENTRE

My Workspace

- Projects
- Finance
- Team
- Settings

22 Feb 2023, 09:51 Published

\$21.93/hr Average reward per hour

00:14:53 Median time

221 / 241 Submissions in progress

35,255 of 121,365 Eligible participants

Find by ID...

Download demographic data Bulk actions

Approved (221) Rejected (10) Returned (32) Timed-out (2)

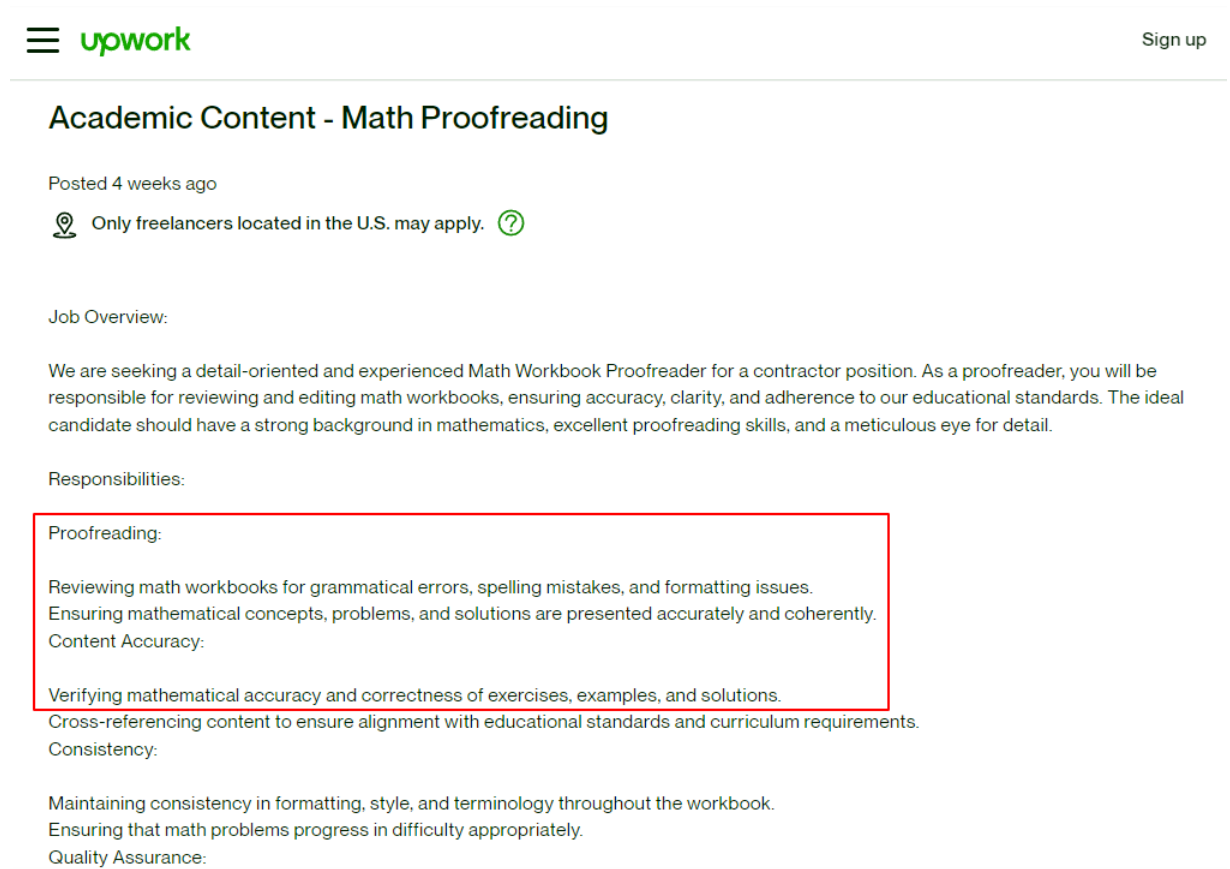
<input type="checkbox"/>	PARTICIPANT PROLIFIC ID	STARTED	TIME TAKEN	COMPLETION CODE	BONUS	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	Sa8e095f1408d000176d5d1	22 Feb 2023, 10:05	00:09:17	QUALIFIED_EASY	\$0.04	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	5d4709f1291942001c2bf782	22 Feb 2023, 10:07	00:12:07	QUALIFIED_HARD	\$3.00	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	60c2e8f25a37360c41db6ac9	22 Feb 2023, 10:12	00:19:20	QUALIFIED_HARD	\$3.00	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	614111dd566906232003da2f	22 Feb 2023, 10:13	00:07:44	QUALIFIED_EASY	\$0.10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	6283b2454867524b04f115bb	22 Feb 2023, 10:17	00:22:21	QUALIFIED_HARD	\$3.00	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	606de1b3a2dc0742a154d180	22 Feb 2023, 10:19	00:07:44	QUALIFIED_EASY	\$0.10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	56f542877e3935000ce43685	22 Feb 2023, 10:20	00:09:49	QUALIFIED_EASY	\$0.60	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	5fc2e0cfd80f9025c6b4810	22 Feb 2023, 10:20	00:12:47	QUALIFIED_HARD	\$3.10	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	6290e41cb4d2e1a893a8a032	22 Feb 2023, 10:20	00:14:48	QUALIFIED_HARD	\$3.06	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	5e3606315bfa435ce4d6246c	22 Feb 2023, 10:21	00:15:02	QUALIFIED_HARD	\$3.00	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	5dc88a6643755e5f227e0488	22 Feb 2023, 10:23	00:14:01	QUALIFIED_HARD	\$3.00	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Approve selected (0) Message selected (0) Bonus payment selected (0)

Figure B.2: User interface on Prolific

B.2 Math tasks on online platforms

Math-related tasks in online platforms include math online teaching, math tutoring, math proofreading, and developing math worksheets, etc. Evaluating the accuracy of answers to math questions is one of the commonly required responsibilities. An example is as below.



The image shows a screenshot of an Upwork job listing. At the top left is the Upwork logo, and at the top right is a 'Sign up' link. The job title is 'Academic Content - Math Proofreading'. Below the title, it says 'Posted 4 weeks ago' and 'Only freelancers located in the U.S. may apply.' with a question mark icon. The 'Job Overview' section states: 'We are seeking a detail-oriented and experienced Math Workbook Proofreader for a contractor position. As a proofreader, you will be responsible for reviewing and editing math workbooks, ensuring accuracy, clarity, and adherence to our educational standards. The ideal candidate should have a strong background in mathematics, excellent proofreading skills, and a meticulous eye for detail.' The 'Responsibilities' section is divided into several categories: 'Proofreading:' (highlighted with a red box) includes 'Reviewing math workbooks for grammatical errors, spelling mistakes, and formatting issues.' and 'Ensuring mathematical concepts, problems, and solutions are presented accurately and coherently.'; 'Content Accuracy:' includes 'Verifying mathematical accuracy and correctness of exercises, examples, and solutions.'; 'Cross-referencing content to ensure alignment with educational standards and curriculum requirements.'; 'Consistency:' includes 'Maintaining consistency in formatting, style, and terminology throughout the workbook.' and 'Ensuring that math problems progress in difficulty appropriately.'; and 'Quality Assurance:'.

Figure B.3: Math proofreading task on Upwork

B.3 Math grading task in the experiment

Question 5 out of 15

3. Ken had 14 chocolate bars and 6 candies. He ate 2 chocolate bars and shared another 4 chocolate bars with his cousin. How many chocolate bars does he have left?

$$(14 - 2) + 4 = x$$

$$\begin{array}{r} 14 \\ -2 \\ \hline 12 \end{array} \quad \begin{array}{r} 12 \\ +4 \\ \hline 16 \end{array}$$

$$x = 16$$



The answer is correct. (1 pt)

The answer is wrong. (0 pt)

Figure B.4: Math grading example question (EL)

Question 3 out of 10

Hint: 16 ounces = 2 cups = 1 pint
 2 pints = 1 quart
 4 quarts = 1 gallon
 1 lb. = 16 oz.

1. Before a party, Ashley bought a tray of 24 water bottles; each bottle has a capacity of 12 oz. There are 30 guests. If each guest drinks 1 cup of water, is there enough water for all the guests?

$$24 \times 12 = x$$

$x \Rightarrow x \text{ cups}$
 $x \text{ cups} \div 30$

$$\begin{array}{r} 24 \\ \times 12 \\ \hline 48 \\ 240 \\ \hline 288 \end{array} \quad \begin{array}{r} 036 \\ 8 \overline{)288} \\ \underline{240} \\ 48 \\ \underline{48} \\ 0 \end{array} \quad 30 \overline{)36} \downarrow$$

Yes

The answer is correct. (1 pt)

The answer is wrong. (0 pt)

Figure B.5: Math grading example question (HH)

B.4 Prolific recruitment ad

Math Test Graders Wanted (elementary school level; with bonus you can get \$8 - \$10+/hr!)

 \$1.50 - \$9.00/hr  10 mins  500 places

We are seeking workers to grade math tests for elementary school students. As a math test grader, you will be responsible for reviewing and grading math tests based on their accuracy. You should have a strong understanding of elementary math concepts and be able to provide feedback on the student answers.

This is a paid opportunity and you will be able to work from the comfort of your own home. If you are interested in participating, please click the link below to start.

Thank you for considering this opportunity. We look forward to working with you.

Devices you can use to take this study:

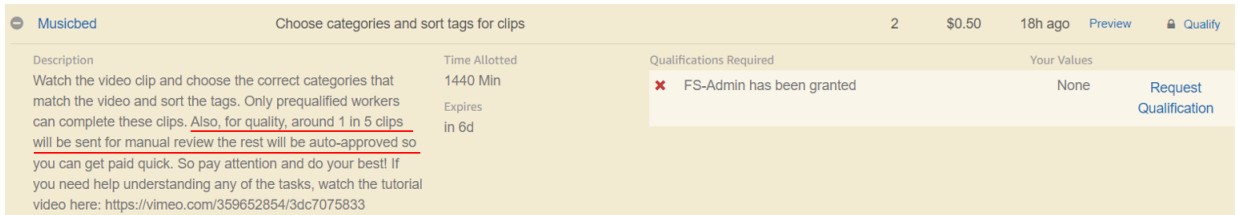
 Desktop  Tablet

[Open study link in a new window](#)

Figure B.6: Recruitment ad on Prolific

B.5 Random accuracy checks in online labor platforms

Due to limited capacity, employers sometimes check a random subset of submissions to motivate high-quality work. Below is an example of random quality check execution in Amazon Mechanical Turk, one of the major online labor platforms.



The screenshot shows a task page for 'Musicbed' with the instruction 'Choose categories and sort tags for clips'. The task has 2 units, a payment of \$0.50, and was posted 18 hours ago. A 'Quality' check is active. The task description includes a warning: 'Also, for quality, around 1 in 5 clips will be sent for manual review the rest will be auto-approved so you can get paid quick. So pay attention and do your best! If you need help understanding any of the tasks, watch the tutorial video here: https://vimeo.com/359652854/3dc7075833'. The 'Qualifications Required' section shows a red 'x' icon and the text 'FS-Admin has been granted'. The 'Your Values' section shows 'None' and a 'Request Qualification' button.

Musicbed		Choose categories and sort tags for clips		2	\$0.50	18h ago	Preview	Quality
Description	Time Allotted	Qualifications Required		Your Values				
Watch the video clip and choose the correct categories that match the video and sort the tags. Only prequalified workers can complete these clips. <u>Also, for quality, around 1 in 5 clips will be sent for manual review the rest will be auto-approved so you can get paid quick. So pay attention and do your best! If you need help understanding any of the tasks, watch the tutorial video here: https://vimeo.com/359652854/3dc7075833</u>	1440 Min	✘ FS-Admin has been granted		None				
	Expires in 6d			Request Qualification				

Figure B.7: Random quality check on Mturk

B.6 Performance feedback and control messages

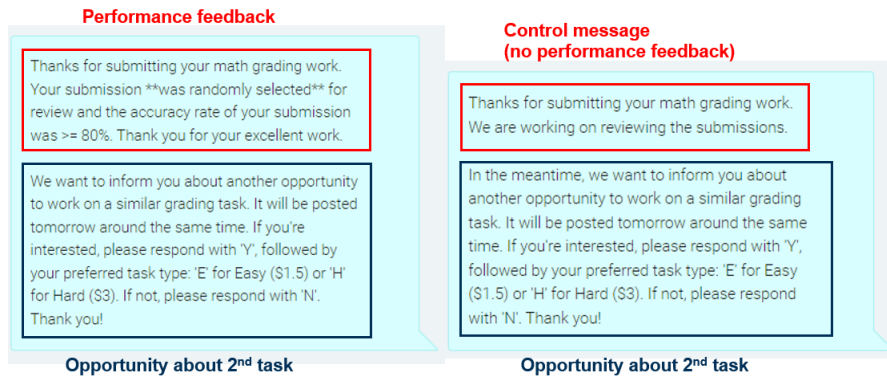


Figure B.8: Performance feedback and control messages in Prolific

B.7 The mediation role of risk aversion

In Table B.1 Model (1), the coefficient of $Female_i$ is negative, consistent with our expectation that women are less risk-seeking and invest fewer tokens in a risky investment game. However, the insignificant coefficient does not satisfy the first criterion of the mediation relationship proposed in (Preacher & Hayes, 2008; Sterling et al., 2020) (three criteria explained in subsection 2.4.4). Comparing Model (2) and (3), after $Risk_pref_i$ is added in Model (3) as an additional explanatory variable, the significance of $Female_i$ remains stable across the two models, which does not satisfy criterion (3). Altogether, the analysis results show that the mediation role of risk aversion is not supported.

Table B.1: Mediation effect of risk aversion

Dependent variable	(1) $Risk_pref_i$	(2) $Pick_hh_{i,t=2}$	(3) $Pick_hh_{i,t=2}$
<i>Intercept</i>	-1.204 (9.424)	-14.769** (6.914)	-14.845** (6.907)
$Female_i$	-0.394 (0.383)	-0.604* (0.309)	-0.606* (0.309)
$Risk_pref_i$			-0.017 (0.047)
Control variables			
$Task_{i,t=1}$		2.416*** (0.364)	2.432*** (0.367)
$Bias_{i,t=1}$		0.294*** (0.105)	0.298*** (0.106)
Demographics			
N	292	292	292
F	2.131**		
Adj. R -squared	0.030		
Pseudo R -squared		0.221	0.221

Note: s.e. in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

B.8 Moderation effect of positive and negative performance feedback

Table 7 presents the moderation analysis of performance feedback within the subgroup of negative and positive feedback, respectively. The dependent variable is $Pick_hh_{i,t=2}$, a binary variable indicates the selection of HH in Phase 2. While the coefficient of $Feedback_i \times Bias_{i,t=1}$ is insignificant within the negative feedback group, it is marginally significant within the positive feedback group. This result suggests that the effect of performance feedback is mainly through alleviating the effect of negative self-evaluation biases in workers' subsequent task selections.

Table B.2: The moderation effect of positive and negative feedback

Variable	Negative Feedback Subgroup	Positive Feedback Subgroup
<i>Intercept</i>	-11.795 (22.213)	-21.226 (14.766)
$Feedback_i \times Bias_{i,t=1}$	-1.719 (1.244)	-1.114* (0.633)
<i>Female_i</i>	-0.27 (0.777)	-0.746 (0.656)
<i>Feedback_i</i>	2.191 (1.956)	-2.399* (1.315)
$Bias_{i,t=1}$	0.237 (0.99)	1.321** (0.534)
$Task_{i,t=1}$	4.575*** (1.447)	3.02*** (0.994)
<i>Risk_pref_i</i>	-0.021 (0.123)	0.015 (0.115)
$Accuracy_{i,t=1}$	4.673 (2.999)	9.259*** (2.722)
Demographics	Controlled in all models	

Note: s.e. in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. $Feedback_i$ is a dummy variable which equals 1 if worker i received the performance feedback after phase 1, and 0 otherwise

APPENDIX C
BEYOND HEALTH RISKS: THE IMPACT OF COVID-19 ON LABOR
VIOLATIONS IN GLOBAL SUPPLY CHAINS

C.1 Heterogeneity analysis based on alternative post periods (US buying firms)

Table C.1 and Table C.2 present the heterogeneity analysis based on the 2-year and 3-year post period, respectively. The results based on alternative post periods show a consistent pattern with the ones based on 1-year post period as in Table 3.3.

Table C.1: Heterogeneity Analysis over 2-year post period (US buying firms)

	<i>n_supplier</i>	<i>n_supplier_iso</i>	<i>duration_median</i>
<i>Treat</i>	-0.002 (0.019)	0.030 (0.023)	-0.006 (0.017)
<i>n_supplier</i>	0.010 *** (0.001)	0.012 *** (0.001)	0.009 *** (0.001)
<i>n_supplier_iso</i>	-0.017 *** (0.003)	-0.024 *** (0.004)	-0.014 *** (0.003)
<i>duration_median</i>	0.001 (0.005)	-0.000 (0.005)	0.000 (0.004)
<i>Treat</i> × <i>Post</i>	-0.037 * (0.023)	-0.142 *** (0.028)	-0.042 ** (0.021)
<i>Post</i> × <i>n_supplier</i>	-0.001 (0.000)		
<i>Treat</i> × <i>n_supplier</i>	-0.001 (0.000)		
<i>Treat</i> × <i>Post</i> × <i>n_supplier</i>	0.006 *** (0.001)		
<i>Post</i> × <i>n_supplier_iso</i>		-0.009 *** (0.003)	
<i>Treat</i> × <i>n_supplier_iso</i>		-0.009 *** (0.003)	
<i>Treat</i> × <i>Post</i> × <i>n_supplier_iso</i>		0.044 *** (0.004)	
<i>Post</i> × <i>duration_median</i>			0.004 (0.007)
<i>Treat</i> × <i>duration_median</i>			0.006 (0.009)
<i>Treat</i> × <i>Post</i> × <i>duration_median</i>			-0.020 * (0.011)
<i>Controls</i>	Included	Included	Included
Year-Quarter FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
N	6,905	6,905	6,905
R^2	0.168	0.146	0.109

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses

Table C.2: Heterogeneity Analysis over 3-year post period (US buying firms)

	<i>n_supplier</i>	<i>n_supplier_iso</i>	<i>duration_median</i>
<i>Treat</i>	-0.006 (0.017)	0.023 (0.022)	-0.028 (0.029)
<i>n_supplier</i>	0.009 *** (0.001)	0.011 *** (0.001)	0.011 *** (0.001)
<i>n_supplier_iso</i>	-0.014 *** (0.003)	-0.020 *** (0.003)	-0.019 *** (0.003)
<i>duration_median</i>	0.000 (0.004)	-0.001 (0.004)	0.001 (0.006)
<i>Treat</i> × <i>Post</i>	-0.042 ** (0.021)	-0.145 *** (0.026)	0.136 *** (0.035)
<i>Post</i> × <i>n_supplier</i>	-0.001 ** (0.000)		
<i>Treat</i> × <i>n_supplier</i>	-0.000 (0.000)		
<i>Treat</i> × <i>Post</i> × <i>n_supplier</i>	0.006 *** (0.000)		
<i>Post</i> × <i>n_supplier_iso</i>		-0.010 *** (0.003)	
<i>Treat</i> × <i>n_supplier_iso</i>		-0.007 ** (0.003)	
<i>Treat</i> × <i>Post</i> × <i>n_supplier_iso</i>		0.043 *** (0.004)	
<i>Post</i> × <i>duration_median</i>			0.004 (0.006)
<i>Treat</i> × <i>duration_median</i>			0.005 (0.008)
<i>Treat</i> × <i>Post</i> × <i>duration_median</i>			-0.019 * (0.010)
<i>Controls</i>	Included	Included	Included
Year-Quarter FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
N	8,082	8,082	8,082
<i>R</i> ²	0.163	0.139	0.099

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses

C.2 DID main analysis (Chinese buying firms)

For the Chinese buying firms, the shock timing is modified to be 2019 Q4. We first perform parallel trend test to check whether this assumption is violated or not. As indicated in Table C.3, the coefficients over pre-shock periods are all insignificant, suggesting that the parallel trend assumption is unlikely to be violated.

Table C.3: Parallel trend test (Chinese buying firms)

	(1)	(2)
<i>Treat</i>	0.000 (0.004)	-0.000 (0.005)
<i>Treat</i> × <i>Quarter</i> ₋₃	0.007 (0.005)	0.015 (0.012)
<i>Treat</i> × <i>Quarter</i> ₋₂	-0.001 (0.005)	-0.004 (0.012)
<i>Treat</i> × <i>Quarter</i> ₋₁	-0.001 (0.005)	0.000 (0.007)
<i>Treat</i> × <i>Quarter</i> ₁	0.000 (0.005)	0.000 (0.007)
<i>Treat</i> × <i>Quarter</i> ₂	0.004 (0.005)	0.006 (0.006)
<i>Treat</i> × <i>Quarter</i> ₃	0.007 (0.005)	0.008 (0.007)
<i>Treat</i> × <i>Quarter</i> ₄	0.007 (0.005)	0.009 (0.006)
<i>Treat</i> × <i>Quarter</i> ₅	0.006 (0.005)	0.020 * (0.012)
<i>Treat</i> × <i>Quarter</i> ₆	0.032 *** (0.005)	0.036 *** (0.006)
<i>Treat</i> × <i>Quarter</i> ₇	0.014 *** (0.005)	0.017 *** (0.006)
<i>Treat</i> × <i>Quarter</i> ₈	0.002 (0.005)	0.004 (0.006)
<i>Treat</i> × <i>Quarter</i> ₉	0.005 (0.005)	0.001 (0.014)
<i>Treat</i> × <i>Quarter</i> ₁₀	0.002 (0.005)	0.003 (0.006)
<i>Treat</i> × <i>Quarter</i> ₁₁	0.012 ** (0.005)	0.014 ** (0.006)
<i>Treat</i> × <i>Quarter</i> ₁₂	0.008 (0.005)	0.010 (0.006)
<i>Controls</i>		Included
Year-Quarter FE	Yes	Yes
Firm FE	Yes	Yes
N	39,933	27,630
R ²	0.005	0.006

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

The DID analysis results based on Chinese buying firms are presented in Table C.4. Similar to US buying firms, COVID-19 has a positive impact on the number of labor violation incidents in their suppliers (coefficients ranging from 0.005 to 0.015, all significant at 1% level). Furthermore, the impact on Chinese firms persists for at least 3 years after the shock, contrasting with the observed trend in the US firms where such an impact attenuates by the third year.

Table C.4: DID analysis results (Chinese buying firms)

	1-year post period		2-year post period		3-year post period	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Treat</i>	0.001 (0.001)	0.005 (0.003)	0.001 (0.002)	0.005 (0.004)	0.001 (0.002)	0.005 (0.004)
<i>Treat</i> × <i>Post</i>	0.005 *** (0.002)	0.011 *** (0.004)	0.009 *** (0.002)	0.015 *** (0.005)	0.008 *** (0.002)	0.013 *** (0.004)
<i>Controls</i>		Included		Included		Included
N	18,949	8,383	29,328	14,336	39,933	21,124
R^2	0.002	0.007	0.004	0.008	0.004	0.007

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

C.3 Heterogeneity analysis (Chinese buying firms)

The heterogeneity analyses based on Chinese buying firms show that the increase in labor violations is greater for firms with more suppliers (0.011, significant at 1% level) and more diversified geographic exposure of suppliers (0.049, significant at 1%), consistent with the patterns observed for the US firms in Table 3.3. However, different from the US firms, the impact is not attenuated for the Chinese firms engaging in longer supplier relationships.

Table C.5: Heterogeneity analysis over 1-year post period (Chinese buying firms)

	<i>n_supplier</i>	<i>n_supplier_iso</i>	<i>duration_median</i>
<i>Treat</i>	0.048 (0.048)	0.123** (0.054)	0.120 (0.079)
<i>n_supplier</i>	0.026*** (0.008)	0.026*** (0.004)	0.031*** (0.004)
<i>n_supplier_iso</i>	-0.078*** (0.016)	-0.067*** (0.018)	-0.113*** (0.016)
<i>duration_median</i>	-0.014 (0.011)	-0.015 (0.011)	0.002 (0.014)
<i>Treat</i> × <i>Post</i>	-0.096** (0.047)	-0.166*** (0.055)	0.022 (0.070)
<i>Post</i> × <i>n_supplier</i>	-0.005 (0.004)		
<i>Treat</i> × <i>n_supplier</i>	-0.007 (0.006)		
<i>Treat</i> × <i>Post</i> × <i>n_supplier</i>	0.011*** (0.004)		
<i>Post</i> × <i>n_supplier_iso</i>		-0.014 (0.011)	
<i>Treat</i> × <i>n_supplier_iso</i>		-0.045*** (0.016)	
<i>Treat</i> × <i>Post</i> × <i>n_supplier_iso</i>		0.049*** (0.012)	
<i>Post</i> × <i>duration_median</i>			-0.000 (0.017)
<i>Treat</i> × <i>duration_median</i>			-0.051* (0.027)
<i>Treat</i> × <i>Post</i> × <i>duration_median</i>			-0.020 (0.023)
<i>Controls</i>	Included	Included	Included
Year-Quarter FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
N	502	502	502
<i>R</i> ²	0.266	0.249	0.188

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$ (two-tailed tests). Standard errors in parentheses.

REFERENCES

- Cramer, J., & Krueger, A. B. (2016). Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5), 177–82.
- Azar, O. H. (2003). The implications of tipping for economics and management. *International Journal of Social Economics*.
- Rapier, G. (2019). Doordash is under fire for its controversial tipping policy. we asked uber, lyft, instacart and other gig-economy startups how much of your tips go directly to their workers. *Business Insider*, July, available at: <https://www.businessinsider.com/which-delivery-app-lets-workers-keep-their-tips-2019-7>.
- Uber. (2020a, November). Our approach to tipping (Medium, Ed.) [Online; posted 28-April-2016].
- Abramson, M. (2023). \$388 in sushi. just a \$20 tip: The brutal math of uber eats and doordash. *The New York Times*.
- Xu, T. (2019, August). Working to strike a better balance [Online; Retrieved on 3-July-2022].
- Postmates. (2021, June). How does tipping work [Online; Retrieved on 21-June-2021].
- Chandar, B., Gneezy, U., List, J. A., & Muir, I. (2019). *The drivers of social preferences: Evidence from a nationwide tipping field experiment* (tech. rep.). National Bureau of Economic Research.
- Zhang, Y., Li, B., & Krishnan, R. (2020). Learning individual behavior using sensor data: The case of global positioning system traces and taxi drivers. *Information Systems Research*.
- Azar, O. H. (2020). The economics of tipping. *Journal of Economic Perspectives*, 34(2), 215–36.
- Kabra, A., Belavina, E., & Girotra, K. (2020). Bike-share systems: Accessibility and availability. *Management Science*, 66(9), 3803–3824.
- Erkan, H., Stamatopoulos, I., Agarwal, A., & Muthuraman, K. (2022). Why so many scooters? a policy analysis. *A policy analysis* (May 5, 2022).
- Feldman, P., Li, J., & Tsai, H.-T. (2022). Welfare implications of congestion pricing: Evidence from sf park. *Manufacturing & Service Operations Management*, 24(2), 1091–1109.

- Allon, G., Chen, D., & Moon, K. (2023). Measuring strategic behavior by gig economy workers: Multihoming and repositioning. *Available at SSRN*.
- Cachon, G. P., Daniels, K. M., & Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management, 19*(3), 368–384.
- Chen, M. K. (2016). Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. *Proceedings of the 2016 ACM Conference on Economics and Computation, 455–455*.
- Hall, J. V., Horton, J. J., & Knoepfle, D. T. (2019). Pricing efficiently in designed markets: The case of ride-sharing. *Available john-joseph-horton. com*.
- Bimpikis, K., Candogan, O., & Saban, D. (2019). Spatial pricing in ride-sharing networks. *Operations Research, 67*(3), 744–769.
- Guda, H., & Subramanian, U. (2019). Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science, 65*(5), 1995–2014.
- Ozkan, E., & Ward, A. R. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems, 10*(1), 29–70.
- Cui, R., Li, J., & Zhang, D. J. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science, 66*(3), 1071–1094.
- Karacaoglu, N., Moreno, A., & Ozkan, C. (2018). Strategically giving service: The effect of real-time information on service efficiency. *Available at SSRN 3260035*.
- Kulkarni, S., & Kalkanci, B. (2021). *Spatial information sharing on on-demand service platforms: A behavioral examination* [Working paper].
- Azar, O. H. (2009). Incentives and service quality in the restaurant industry: The tipping–service puzzle. *Applied Economics, 41*(15), 1917–1927.
- Chen, Y.-J., Dai, T., Korpeoglu, C. G., Korpeoglu, E., Sahin, O., Tang, C. S., & Xiao, S. (2020). Om forum-innovative online platforms: Research opportunities. *Manufacturing & Service Operations Management, 22*(3), 430–445.
- Buchholz, N. (2018). *Spatial equilibrium, search frictions and dynamic efficiency in the taxi industry* (tech. rep.). Working Paper.

- Jiang, Z.-Z., Kong, G., & Zhang, Y. (2020). Making the most of your regret: Workers' relocation decisions in on-demand platforms. *Manufacturing & Service Operations Management*.
- Küçükgül, C., Özer, Ö., & Wang, S. (2022). Engineering social learning: Information design of time-locked sales campaigns for online platforms. *Management Science*, 68(7), 4899–4918.
- Blair, M., Alizamir, S., & Wang, S. (2022). *The impact of climate change: An empirical analysis of smart thermostat data* [Working paper].
- Haggag, K., McManus, B., & Paci, G. (2017). Learning by driving: Productivity improvements by new york city taxi drivers. *American Economic Journal: Applied Economics*, 9(1), 70–95.
- Alexander, D., Boone, C., & Lynn, M. (2021). The effects of tip recommendations on customer tipping, satisfaction, repatronage, and spending. *Management Science*, 67(1), 146–165.
- Lynn, M., Zinkhan, G. M., & Harris, J. (1993). Consumer tipping: A cross-country study. *Journal of Consumer Research*, 20(3), 478–488.
- Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing science*, 15(1), 1–20.
- Narayanan, S., & Manchanda, P. (2009). Heterogeneous learning and the targeting of marketing communication for new products. *Marketing science*, 28(3), 424–441.
- Shin, S., Misra, S., & Horsky, D. (2012). Disentangling preferences and learning in brand choice models. *Marketing Science*, 31(1), 115–137.
- Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science*, 32(1), 153–169.
- Huang, Y., Vir Singh, P., & Srinivasan, K. (2014). Crowdsourcing new product ideas under consumer learning. *Management science*, 60(9), 2138–2159.
- Wu, C., Che, H., Chan, T. Y., & Lu, X. (2015). The economic value of online reviews. *Marketing Science*, 34(5), 739–754.
- Sriram, S., Chintagunta, P. K., & Manchanda, P. (2015). Service quality variability and termination behavior. *Management Science*, 61(11), 2739–2759.

- Emadi, S. M., & Swaminathan, J. M. (2018). Customer learning in call centers from previous waiting experiences. *Operations Research*, 66(5), 1433–1456.
- Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2), 407–441.
- Farber, H. S. (2015). Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. *The Quarterly Journal of Economics*, 130(4), 1975–2026.
- Hathaway, B. A., Emadi, S. M., & Deshpande, V. (2022). Personalized priority policies in call centers using past customer interaction information. *Management Science*, 68(4), 2806–2823.
- Haggag, K., & Paci, G. (2014). Default tips. *American Economic Journal: Applied Economics*, 6(3), 1–19.
- Brewster, Z. W., & Mallinson, C. (2009). Racial differences in restaurant tipping: A labour process perspective. *The Service Industries Journal*, 29(8), 1053–1075.
- Hahn, J., & Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3), 863–881.
- Bonhomme, S., Lamadon, T., & Manresa, E. (2017). Discretizing unobserved heterogeneity. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-16).
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411–423.
- Teixeira, T., Wedel, M., & Pieters, R. (2012). Emotion-induced engagement in internet video advertisements. *Journal of marketing research*, 49(2), 144–159.
- Corritore, M., Goldberg, A., & Srivastava, S. B. (2020). Duality in diversity: How intrapersonal and interpersonal cultural heterogeneity relate to firm performance. *Administrative Science Quarterly*, 65(2), 359–394.
- Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3), 203–238.
- Erdem, T., Keane, M. P., & Sun, B. (2008). A dynamic model of brand choice when price and advertising signal product quality. *Marketing Science*, 27(6), 1111–1125.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.
- Donovan, C. (2016). Instacart cancels plans to scrap tips amid threats of strikes. *BuzzFeed News*, October, available at: <https://www.buzzfeednews.com/article/carolineodonovan/instacart-cancels-plans-to-scrap-tips-amid-threats-of-strike>.
- Rosenbloom, S. (2016). To tip or not to tip your uber driver. *The New York Times*.
- Matarese, J. (2021, January). No tip, no trip: Here's why your food order may be late [Online; posted 28-January-2021].
- Samuels, A., & Burnley, M. (2019, August). Low wages, sexual harassment and unreliable tips. this is life in america's booming service industry [Online; posted 22-August-2019].
- Sainato, M. (2019, March). I made \$3.75 an hour': Lyft and uber drivers push to unionize for better pay [Online; posted 22-March-2019].
- Azar, O. H. (2010). Tipping motivations and behavior in the us and israel. *Journal of Applied Social Psychology*, 40(2), 421–457.
- Uber. (2020b, November). \$600 million worth of connections and counting (Uber, Ed.) [Online; posted 21-June-2018].
- Hoisl, K., & Mariani, M. (2017). It's a man's job: Income and the gender gap in industrial research. *Management Science*, 63(3), 766–790.
- Chan, J., & Wang, J. (2018). Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, 64(7), 2973–2994.
- Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of “blind” auditions on female musicians. *American economic review*, 90(4), 715–741.
- Das, S., & Kotikula, A. (2019). *Gender-based employment segregation: Understanding causes and policy interventions*. World Bank.
- Goldin, C. (2014). A grand gender convergence: Its last chapter. *American economic review*, 104(4), 1091–1119.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3), 1067–1101.

- Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1), 122–155.
- Leibbrandt, A., & List, J. A. (2015). Do women avoid salary negotiations? evidence from a large-scale natural field experiment. *Management Science*, 61(9), 2016–2024.
- Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3), 789–865.
- Foong, E., Vincent, N., Hecht, B., & Gerber, E. M. (2018). Women (still) ask for less: Gender differences in hourly rate in an online labor marketplace. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–21.
- Litman, L., Robinson, J., Rosen, Z., Rosenzweig, C., Waxman, J., & Bates, L. M. (2020). The persistence of pay inequality: The gender pay gap in an anonymous online labor market. *PloS one*, 15(2), e0229383.
- Daniels, K., & Grinstein-Weiss, M. (2019). The impact of the gig-economy on financial hardship among low-income families. *Available at SSRN 3293988*.
- Kasliwal, R. (2020). Gender and the gig economy: A qualitative study of gig platforms for women workers. *ORF Issue Brief*, 359, 1–14.
- Kricheli-Katz, T., & Regev, T. (2016). How many cents on the dollar? women and men in product markets. *Science advances*, 2(2), e1500599.
- Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1914–1933.
- Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? executive search and gender inequality in hiring for top management jobs. *Management Science*, 62(12), 3636–3655.
- Liang, C., Hong, Y., Gu, B., & Peng, J. (2018). Gender wage gap in online gig economy and gender differences in job preferences. *Available at SSRN 3266249*.
- Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14.
- Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8, 321–340.

- Wall, S., & Schellmann, H. (2021). LinkedIn, ai bias, ziprecruiter, monster, artificial intelligence [Accessed: 2024-03-11].
- Coffman, K. B., Exley, C. L., & Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6), 3551–3569.
- Coffman, K. B., Collis, M. R., & Kulkarni, L. (2023). Whether to apply. *Management Science*.
- Cook, C., Diamond, R., Hall, J. V., List, J. A., & Oyer, P. (2021). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. *The Review of Economic Studies*, 88(5), 2210–2238.
- Rahman, H. A. (2021). The invisible cage: Workers' reactivity to opaque algorithmic evaluations. *Administrative Science Quarterly*, 66(4), 945–988.
- Kokkodis, M., & Ipeiritis, P. G. (2016). Reputation transferability in online labor markets. *Management Science*, 62(6), 1687–1706.
- Paul, C. S. (2018). The architecture of digital labour platforms: Policy recommendations on platform design for worker well-being. *ILO future of work research paper series*, 3.
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics*, 118(3), 1049–1074.
- Upwork. (2024a). Search for jobs [Accessed: 2024-03-03]. *Upwork*.
- Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American journal of Sociology*, 106(6), 1691–1730.
- Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological bulletin*, 84(1), 1.
- Exley, C. L., & Kessler, J. B. (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137(3), 1345–1381.
- Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of personality and social psychology*, 59(5), 960.
- Krawczyk, M., & Wilamowski, M. (2019). Task difficulty and overconfidence. evidence from distance running. *Journal of Economic Psychology*, 75, 102128.
- Chen, X., Zeng, X., & Zhang, C. (2023). Does concealing gender identity help women win the competition? an empirical investigation into online video games. *Marketing Science*, 42(3), 551–568.

- Mehrabian, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1–35.
- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1), 1–12.
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52, 388–407.
- Gesiarz, F., De Neve, J.-E., & Sharot, T. (2020). The motivational cost of inequality: Opportunity gaps reduce the willingness to work. *Plos one*, 15(9), e0237914.
- Tandon, A., Kaur, P., Bhatt, Y., Mäntymäki, M., & Dhir, A. (2021). Why do people purchase from food delivery apps? a consumer value perspective. *Journal of Retailing and Consumer Services*, 63, 102667.
- Schneider, C., Weinmann, M., Mohr, P. N., & vom Brocke, J. (2021). When the stars shine too bright: The influence of multidimensional ratings on online consumer ratings. *Management Science*, 67(6), 3871–3898.
- Read, D., Olivola, C. Y., & Hardisty, D. J. (2017). The value of nothing: Asymmetric attention to opportunity costs drives intertemporal decision making. *Management Science*, 63(12), 4277–4297.
- Huang, S., Hwang, B.-H., Lou, D., & Yin, C. (2020). Offsetting disagreement and security prices. *Management Science*, 66(8), 3444–3465.
- Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The quarterly journal of economics*, 112(2), 631–645.
- Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5), 1637–1664.
- Dargnies, M.-P., Hakimov, R., & Kübler, D. (2019). Self-confidence and unraveling in matching markets. *Management Science*, 65(12), 5603–5618.
- Borghans, L., Heckman, J. J., Golsteyn, B. H., & Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3), 649–658.

- Schubert, R., Gysler, M., Brown, M., & Brachinger, H.-W. (2000). *Gender specific attitudes towards risk and ambiguity: An experimental investigation* (tech. rep.). Economics Working Paper Series.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, *40*(3), 879–891.
- Sterling, A. D., Thompson, M. E., Wang, S., Kusimo, A., Gilmartin, S., & Sheppard, S. (2020). The confidence gap predicts the gender pay gap among stem graduates. *Proceedings of the National Academy of Sciences*, *117*(48), 30303–30308.
- Upwork. (2024b). Approaching clients about feedback [Accessed: 2024-03-03]. *Upwork*.
- Shi, P. (2023). Optimal matchmaking strategy in two-sided marketplaces. *Management Science*, *69*(3), 1323–1340.
- Berg, J., Hilal, A., El, S., Horne, R., et al. (2021). World employment and social outlook: Trends 2021. *International Labour Organization*.
- ILO. (2020). Covid-19 impact on child labour and forced labour: The response of the ipec+ flagship programme.
- Chen, L., Lee, H. L., & Tang, C. S. (2022). Supply chain fairness. *Production and Operations Management*, *31*(12), 4304–4318.
- Anner, M. (2022). Power relations in global supply chains and the unequal distribution of costs during crises: Abandoning garment suppliers and workers during the covid-19 pandemic. *International Labour Review*, *161*(1), 59–82.
- Sim, J., Cho, D., Hwang, Y., & Telang, R. (2022). Frontiers: Virus shook the streaming star: Estimating the covid-19 impact on music consumption. *Marketing Science*, *41*(1), 19–32.
- Han, B. R., Sun, T., Chu, L. Y., & Wu, L. (2022). Covid-19 and e-commerce operations: Evidence from alibaba. *Manufacturing & Service Operations Management*, *24*(3), 1388–1405.
- Organization, I. L. (2017). Global estimates of modern slavery: Forced labour and forced marriage. *International Labour Organization*.
- Distelhorst, G., Hainmueller, J., & Locke, R. M. (2017). Does lean improve labor standards? management and social performance in the nike supply chain. *Management Science*, *63*(3), 707–728.

- de Zegher, J. F., Liu, B., Rende Taylor, L., & Taylor, M. (2021). Low-skilled labor shortages contribute to forced labor. *Available at SSRN 3899489*.
- Vanpoucke, E., & Klassen, R. D. (2023). Reducing forced labour in supply chains: What could traditional companies learn from social enterprises? *International Journal of Operations & Production Management*, (ahead-of-print).
- Kwon, C., & Wu, D. A. (2021). Disclosure-driven social engagement in supply chains. *Available at SSRN 3900182*.
- Amengual, M., & Distelhorst, G. (2020). Cooperation and punishment in regulating labor standards: Evidence from the gap inc supply chain. *Available at SSRN 3466936*.
- Uddin, S., Ahmed, M. S., & Shahadat, K. (2023). Supply chain accountability, covid-19, and violations of workers' rights in the global clothing supply chain. *Supply Chain Management: An International Journal*, 28(5), 859–873.
- Delasay, M., Jain, A., & Kumar, S. (2022). Impacts of the covid-19 pandemic on grocery retail operations: An analytical model. *Production and Operations Management*, 31(5), 2237–2255.
- Cao, X., Zhang, D., & Huang, L. (2022). The impact of the covid-19 pandemic on the behavior of online gig workers. *Manufacturing & Service Operations Management*, 24(5), 2611–2628.
- Cui, R., Ding, H., & Zhu, F. (2022). Gender inequality in research productivity during the covid-19 pandemic. *Manufacturing & Service Operations Management*, 24(2), 707–726.
- Li, J., & Wu, D. (2020). Do corporate social responsibility engagements lead to real environmental, social, and governance impact? *Management Science*, 66(6), 2564–2588.
- Cen, L., Han, Y., Liu, C., & Wu, J. (2022). The rise of conscious consumers: The cash flow implications of corporate workplace equality. In *The rise of conscious consumers: The cash flow implications of corporate workplace equality: Cen, ling—uhan, yanru—uliu, chang—uwu, jing*. [SI]: SSRN.
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243.

VITA

Li Ding is a Ph.D. candidate at Scheller College of Business, Georgia Institute of Technology. Her research focuses on people-centric operations, the social impacts of gig economy, and socially responsible supply chains. She employs empirical methodologies including structural estimation, econometrics, experiments, and event studies.

Before her current pursuit at Georgia Tech, Li received her Ph.D. in Operations Management from Hong Kong Polytechnic University in 2020 and her B.A. from Chongqing University in 2014.