Summer 5-15-2023

Data-driven Platform and Digital Operations

Bing Bai
Washington University in St. Louis, bing.bai@wustl.edu

WASHINGTON UNIVERSITY IN ST. LOUIS

Olin Business School
Department of Supply Chain, Operations, and Technology

Dissertation Examination Committee:
Dennis J. Zhang, Co-Chair
Fuqiang Zhang, Co-Chair
Tat Y. Chan
Hengchen Dai
Jacob Feldman

Data-driven Platform and Digital Operations

by

Bing Bai

A dissertation presented to
Olin Business School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am filled with gratitude to have the support of an amazing group of people throughout this journey.

I would like to start by thanking my co-advisor, Dennis J. Zhang, whose encouragement and guidance were instrumental in helping me discover my research passions and steer me toward this career path. He is a role model for me as both a researcher and mentor. I want to thank my co-advisor Fuqiang Zhang for introducing me to this field, and for his continuous support throughout the doctoral program. I would also like to thank Tat Y. Chan and Hengchen Dai for providing insightful advice and spending significant effort on my research and academic career. I am also grateful to Lingxiu Dong, Jacob Feldman, and Fasheng Xu for their valuable advice, and the time and energy they have dedicated to supporting me. I would like to express my sincere appreciation to Zhenling Jiang, Xiaoyang Long, Jiankun Sun, Renyu Zhang, and Heng Zhang for their insightful advice in my job search process. I would also like to express my gratitude to the entire Olin community and my PhD friends for their support.

Finally, I want to especially thank my parents, Lixin Sun and Xuedong Bai, and my fiancé, Zhi Huang. Their belief in me and constant support have made it possible for me to get through this journey.

Bing Bai

Washington University in Saint Louis

May 2023

Dedicated to my family.

ABSTRACT OF THE DISSERTATION

Data-driven Platform and Digital Operations

by

Bing Bai

Doctor of Philosophy in Business Administration

Washington University in St. Louis, 2023

Professor Dennis J. Zhang, Co-Chair

Professor Fuqiang Zhang, Co-Chair

The objective of this dissertation is to study the emerging operations issues on data-driven platforms and digital operations. With the increasing availability of data and the development of information technologies, platforms process a large amount of data in order to efficiently make daily operational decisions. Understanding human behaviors and the human-algorithm connection is instrumental to the success of this process. In my research, I implement field experiments and use structural models to study in-warehouse worker behavior and out-of-warehouse customer behavior in the last mile of logistics.

In Chapter 1, "The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity: Evidence from Field Experiments", we study in-warehouse worker behavior. We study how algorithmic (vs. human-based) task assignment processes change task recipients' fairness perceptions and productivity. In a 15-day-long field experiment with Alibaba Group in a warehouse where workers pick products following orders (or "pick lists"), we randomly assigned half of the workers to receive pick lists from a machine that ostensibly relied on an algorithm to distribute pick lists, and the other half to receive pick lists from a human distributor. Despite that we used the same underlying rule to assign pick lists in both groups, workers perceive the algorithmic (vs. human-based) assignment process as fairer by 0.94-1.02

standard deviations. This yields productivity benefits: receiving tasks from an algorithm (vs. a human) increases workers' picking efficiency by 15.56%-17.86%. These findings persist beyond the first day when workers were involved in the experiment, suggesting that our results are not limited to the initial phrase when workers might find algorithmic assignment novel. We replicate the main results in another field experiment involving a nonoverlapping sample of warehouse workers. We also show via online experiments that people in the U.S. also view algorithmic task assignment as fairer than human-based task assignment. We demonstrate that algorithms can have broader impacts beyond offering greater efficiency and accuracy than humans: introducing algorithmic assignment processes may enhance fairness perceptions and productivity. This insight can be utilized by managers and algorithm designers to better design and implement algorithm-based decision making in operations.

In Chapter 2, "The Value of Logistic Flexibility in E-commerce", we study out-of-warehouse customer behavior in the last mile of logistics. We use the opening of hundreds of pick-up stations as a natural experiment to study the impact of these stations on consumers. We find that the introduction of pick-up stations has increased total sales by 3.9%. In contrast with past literature, we show that shipping time reduction is not the driving factor on the impact of pick-up stations. Yet, the logistic flexibility introduced by pick-up stations explains the sales impact. To explicitly examine how logistic flexibility affects consumers' decisions on purchases, we develop and estimate a structural model of consumer choice. In our model, consumers value two types of logistics flexibility—the flexibility to pick up their items at their preferred time, denoted as the value of time flexibility, and the flexibility to delay pickup decisions until after packages arrive at a local station, denoted as the value of choice flexibility. We show that the value of time flexibility accounts for 76.2% of the impact on sales, while the value of choice flexibility accounts for the remaining 23.8%. Using our estimated model, we develop a counterfactual strategy in building pick-up stations that could achieve the sales

lift with 56.4%-63.6% fewer stations. Last but not least, using our estimated time flexibility, we also develop a novel shipping strategy without pick-up stations that could improve sales by 8.4%. Our estimates suggest that our counterfactual logistic strategies could increase consumer welfare by 2.0%-10.0%.

Chapter 1

The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity: Evidence from Field Experiments

1.1 Introduction

With the increasing availability of data and the development of information technologies, companies are rapidly implementing algorithms to process a large amount of data in order to efficiently make daily operational decisions. For example, digital service platforms such as Uber and Airbnb instantly match customers with service providers, taking high-dimensional information into account (*e.g.*, customers' willingness to pay, service providers' availability) in their algorithms. Ad platforms such as Facebook and Google combine advertising algorithms

with rich data about consumers to identify specific audience groups for which to display certain ads.

Such growing interest in using algorithms in practice has inspired a large stream of research dedicated to improving algorithms' performance [1, 2]. However, in many domains of daily operations, algorithms rely on human involvement to complete tasks. For example, retailing platforms such as Alibaba use algorithms to determine which set of items should be packed into which box but need human workers in warehouses to pack the items according to algorithmic prescriptions [3]. Similarly, sales platforms such as Salesforce use algorithms to decide which product to be advertised to whom but need human salespeople to make sales pitches following algorithmic recommendations.

Thus, a fundamental question about algorithm development in operations management is how humans perceive and interact with algorithms. A growing body of work has begun to study this question from both operational and psychological perspectives [3, 4, 5, 6, 7, 8, 9, 10]. This literature has largely documented that people are reluctant to use algorithms and prefer instead to defer to judgments made by a human. Another growing concern regarding human and algorithm collaboration is that algorithms may produce or reproduce discriminatory outcomes and lead to new or more systematic biases than what humans have historically exhibited (see [11] for a review). Motivated by this concern, researchers have studied various ways of defining and enforcing fairness when designing algorithms [12].

Extending these two lines of work on how to improve human and algorithm collaboration, we study how workers *perceive* the fairness of algorithmic decisions and how such fairness perceptions affect their behavior when algorithms are used to make decisions related to workers' tasks.

In work settings, algorithms are increasingly replacing human decision makers to allocate resources and tasks (*e.g.*, delivery trips, customers, cases) across workers. We specifically examine how an algorithmic task assignment process, relative to a human-based task assignment process, changes task recipients' fairness perceptions and productivity. To causally answer this question, we conducted two field experiments in collaboration with Alibaba—the largest retailing platform in China—in its warehouse setting. In recent years, e-commerce warehouses have started digitizing equipment and applying algorithms to many key tasks, such as picking, routing, scheduling, and bin packing [3]. We focus on picking tasks for which workers follow a "pick list" to collect products from different stocking shelves.

During our experiments, workers in our partner warehouse were randomly assigned to one of two groups: workers in the algorithm group received picking tasks from a machine that ostensibly relied on an algorithm to distribute pick lists, whereas workers in the human group received picking tasks from a human distributor. We kept the objective nature of these two task assignment processes as well as the characteristics of the assigned pick lists as similar as we could, so the difference we observe between these two assignment processes can be attributed to workers' beliefs and perceptions about the differences between algorithmic and human-based assignment.

The first, main experiment involved 50 temporary workers for 15 days in August-September, 2019. We collected data about all $4,486$ pick lists completed by these workers during this experiment, along with 108 daily questionnaires from them. We present three key findings. First, we find that workers hold different views about the fairness of these two assignment processes: workers receiving tasks from an algorithm on average perceive their assignment process as fairer than workers receiving tasks from a human distributor, and the difference is 0.94-1.02 standard deviations (depending on our model specifications). Second, we document productivity differences between these two assignment processes: receiving tasks from an

algorithm significantly increases workers' picking efficiency by 15.56%-17.86%, compared to receiving tasks from a human distributor. Third, since unobserved variables such as worker ability, can be correlated with both how fair workers believe they are treated and their productivity, we estimate via an instrumental variable approach the effect of fairness perceptions on productivity. We show that a one-standard-deviation increase in fairness perceptions is associated with a boost of picking efficiency by 12.97%-16.98%. We conducted the second field experiment with a nonoverlapping sample of workers in December, 2019-January, 2020 and validated the robustness of our main results.

To further validate our findings from the field, we conducted an online experiment to study the effect of algorithmic (vs. human-based) task assignment on perceived fairness among a different population—201 people in the United States recruited from an online labor market (Amazon's Mechanical Turk). Study participants imagined working in a warehouse and receiving picking tasks from either a machine or a human distributor. Despite imagining receiving the same picking tasks, people on average perceived the assignment process run by a machine based on an algorithm as fairer than the process run by a human. We replicated this pattern in another online experiment with a slightly different design.

In sum, we examine people's psychological and behavioral responses to algorithmic decision-making processes across experiments and settings, and we present the first field experiments to our knowledge in an actual workplace to study this issue. By keeping the nature of pick lists the same between conditions, our design provides a clean test of how people perceive algorithmic (vs. human-based) decision-making processes and how people behave after receiving decisions made by these processes. Theoretically, this angle differentiates our study from the large body of research that examines sources of algorithm-engendered biases and compares algorithms with humans in the actual level of inequality they produce [12, 13, 14, 15, 16, 17]. Practically, through this unique design, our findings can help companies

understand workers' perceptions about algorithmic decision-making processes and optimize the framing of task assignment processes. Altogether, our research complements the existing literature about human-algorithm collaboration, highlights the importance of understanding workers' fairness perceptions when utilizing algorithms, and provides insights for designing better human-algorithm collaboration in daily operations.

1.2 Literature Review and Theoretical Contributions

Our work is mainly related to four research areas: human collaboration with algorithms, algorithmic bias, operations management research about automation, and behavioral operations.

1.2.1 Human Collaboration with Algorithms

Our work is closely connected to the growing stream of literature studying how people perceive and react to algorithms and automation. The primary focus of this literature has been on examining whether humans, as *users* of algorithms, are willing to rely on algorithmic prescriptions and utilize automated systems. With a few exceptions [4, 8], research in this area has largely documented *algorithm aversion*, whereby people are reluctant to utilize algorithms and automation (compared to their own judgment, human experts' advice, or their peers' aid), despite the fact that algorithms give identical output or, in some cases, even superior performance than humans [5, 6, 7, 9].

More recently, this literature has begun to examine how people as *recipients* of decisions (*e.g.*, employees receiving personnel decisions) respond to algorithmic decision processes. This line of research so far has found that people view algorithms as less capable of taking into account their unique, contextual, and personal characteristics [18]; as a result, people perceive

algorithmic decision-making as less procedurally fair and express less commitment to their organizations if algorithms (rather than humans) drive decision making [10].

We make several contributions to this literature. First, while recent research suggests that people disfavor algorithms when they want decision-making processes to consider their unique and personal characteristics [10, 18, 19], we recognize that people often have the equality motive—that is, they would like to receive equal treatment and opportunity relative to others [20, 21, 22]. We complement prior research by documenting that algorithmic decision-making processes are viewed more favorably than human-based decision-making processes in settings where people prioritize the equality motive over other motives that consider personal characteristics. Second, while prior research has focused on how people collaborate with algorithms on prediction tasks and consumer decision-making, we examine how employees perceive algorithms that determine their tasks at work. Our empirical context in the field studies—a labor-intensive working environment—is also a complement to the literature. Third, while the prior research reviewed above has largely used laboratory and online experiments, we conducted field experiments in a common operation setting (warehouse operations) to provide more external validity of our insights. Fourth, going beyond examining people's *perceptions* of algorithms, we further study employees' work behaviors and find a downstream consequence of algorithmic work assignment process on productivity.

1.2.2 Algorithmic Bias

This chapter is also related to the emerging literature about biases and discrimination engendered by algorithms. Scholars are concerned that algorithms may reproduce, codify, or even amplify disparities due to biases in objective functions, people building the algorithms, or historical data [11], and have provided evidence that algorithms perpetuate existing inequality across domains. For example, algorithms have been shown to derive semantics

that associate female names more with family than career-related words [14], produce gender discrimination in ad delivery [16], and yield racial bias in health risk assessments [17]. This concern has motivated researchers to study how to define and enforce fairness when designing algorithms [12]. Despite the concern around algorithmic bias, the scarce research that compares algorithms to human decision makers suggests that algorithmic judgment actually appears less biased than human judgement, even when algorithms are trained on historical data involving biased human decisions [15]. This provides some empirical support for the more optimistic view that the use of AI could have positive implications for social equality and fairness.

While prior work in this literature has focused on identifying when algorithms produce biased outcomes and has compared algorithms to humans in the actual level of fairness generated, we study people's *perceptions* of algorithms' ability to deliver fair treatments. We ask the fundamental question of how knowing that one's outcome is determined by an algorithm (vs. a human) affects people's perceived fairness about the decision process, which subsequently influences their behaviors. To examine this question, we keep the underlying decision-making logic and the assigned outcomes the same but investigate how people's perceptions change when they are led to believe that their outcomes are decided by an algorithm rather than a human decision maker.

1.2.3 Automation in Operations Management

Our work adds to the literature in operations management studying problems that arise in the presence of automation, particularly research that incorporates the role of humans in the design of automated systems [2, 3, 23, 24, 25]. For example, [23] and [3] study how managers' and workers' deviations from algorithmic prescriptions could yield insights for improving operational efficiency. While prior research in this area has focused on how to make

algorithms and automated systems more powerful, we study how people's perceptions about automation affect their efficiency. We show that in the presence of automation, psychological factors such as fairness perceptions impact worker productivity.

1.2.4 Behavioral Operations

Finally, our work builds on the behavioral operations literature. This literature has documented a number of behavioral and psychological drivers of productivity, such as team familiarity, time pressure, peer pressure, quality monitoring, and free-rider effect [26, 27, 28, 29, 30, 31], mostly based on archival data analysis. Through longitudinal field experiments, we document that perceived fairness about task assignment is another important driver of productivity. Also, the behavioral operations literature has shown that people in various operation settings fall prey to behavioral biases, such as framing [32, 33]. For example, [33] shows that the framing of patients' admission time may affect doctors' discharge decisions. Our finding—that people view a decision process as fairer and work more productively when the process is seemingly driven by an algorithm (vs. a human)—provides an example where the framing of work assignment affects operational efficiency.

1.3 Hypothesis Development

We present two hypotheses regarding how assigning tasks via algorithms versus humans affects task recipients' fairness perceptions and productivity in prevalent work settings where workers tend to prioritize the equality motive and desire equal treatment and opportunity relative to others.

When evaluating the fairness of a process that is used to make allocation decisions, people often consider whether the process is free from decision makers' personal biases, applies decision

rules consistently across people and across time, and uses appropriate factual information to make decisions [34, 35]. People may worry that a human decision maker would consciously or unconsciously make favorable or disadvantageous decisions about some individuals for unjustifiable reasons (*e.g.*, close relationships, physical attractiveness), but they may expect algorithms to be free of these personal biases and more capable of consistently applying rules and providing equal treatment across individuals. The differential beliefs about human versus algorithmic decision makers may arise because people often evaluate the fairness of an allocation based on how they attribute the outcome to the decision maker's intentions [36] and people tend to perceive algorithmic decision makers as less intentional than humans [37, 38]. Even in settings as our field experiments where, as explained later in Section 1.4.1, human distributors actually allocated tasks following a specific guideline (rather than using their own discretion), task recipients may still infer that assignment outcomes are more attributable to human distributors' preferences and biases than algorithms'. This psychological process may reflect people's beliefs about the intentionality of different assignment processes. Altogether, we propose the following hypothesis for settings where workers prioritize the equality motive:

Workers perceive a task assignment process as fairer if they believe the process is implemented by an algorithm than if they believe the process is implemented by a human.

Our next hypothesis pertains to how algorithmic (vs. human-based) assignment affects productivity. Research in psychology, organizational behavior, and behavioral economics consistently suggests that people desire fair treatments and behave differently at work in accordance to whether they think they are fairly treated in their organizations (see [35, 39, 40] for reviews of relevant research). In particular, meta-analyses of hundreds of studies suggest that procedural fairness perceptions have a moderately positive correlation with work performance on average ($r = 0.30$; [35]) and that the relationship is stronger among actual employees in work settings ($r = 0.47$) than among students in laboratory studies

[39]. Building on prior research, we predict that as an algorithmic task assignment process increases people's perceived fairness (Hypothesis 1), it should subsequently have a positive impact on their productivity.

Workers are more productive if they believe their task assignment process is implemented by an algorithm than if they believe the process is implemented by a human.

1.4 Experiment Design and Data

1.4.1 Field Setting and Experiment Design

Our field experiments were conducted in collaboration with Alibaba. In 2013, along with five package delivery companies, Alibaba co-founded Cainiao Network (hereafter, "Cainiao"), a logistic platform operator dedicated to digitizing the shipping industry and building a smart logistic network nationally and globally. Cainiao has the largest bonded warehouse network in China and manages more than 60% packages from Alibaba's Chinese retail marketplaces.

We study one core task that workers perform in warehouses: picking, which requires workers to collect certain products from different shelves following specific "pick lists". When an online purchase order is placed on Alibaba, Cainiao's warehouse management system first decides which warehouse should fulfill the order based on the Stock Keeping Units (SKUs) included in the order and their stocking information. After accumulating a number of purchase orders for a given warehouse, the system generates a set of pick lists for this warehouse, with each pick list usually covering multiple purchase orders. A pick list contains information about products that a worker should pick, including SKU name, the quantity the worker should pick for each SKU, and the stocking location of each SKU (see Appendix A.1 for an example pick list).

We conducted two experiments in one of Cainiao's warehouses where picking workers were paid hourly. These experiments had the same design but were run at different times involving nonoverlapping samples of workers. We focus on the first field experiment in this chapter, and report the second field experiment as a replication study in Appendix A.2. Our first, main experiment involved 50 temporary workers and spanned 15 days, starting from August 20, 2019 and ending on September 6, 2019. On August 25-26 and August 30, 2019, the warehouse had a much heavier workload than usual due to Alibaba's platform-wide "shopping holidays" so the experiment was temporally halted on these days.

Before and in between our field experiments, a staff member in the warehouse periodically printed out pick lists as physical hard copies and placed them on a table at the pick list distribution station so that workers could get pick lists themselves based on their own preferences. During our experiments, we manipulated how workers received pick lists and randomly assigned workers into either the human group or the algorithm group. Once assigned, workers stayed in the same group during the span of the experiment. We set up two tables side by side at the distribution station, one for the human group and the other for the algorithm group.

In the human group, hard-copy pick lists were printed and assigned by a human distributor. Specifically, a human distributor stood at the human-based assignment table and periodically printed out a stack of hard-copy pick lists from the pool of available pick lists. The selection of pick lists from the pool and the printing order were designed to be random. When a worker in the human group came to the human-based assignment table, the human distributor handed the worker the hard-copy pick list at the top of the stack. Upon receiving the pick list, the worker scanned the bar code on the pick list using a radio-frequency hand-held monitor, at which time Cainiao's system would record the starting time of this pick list. After scanning the bar code on the hard-copy pick list, the picking worker no longer needed the hard-copy

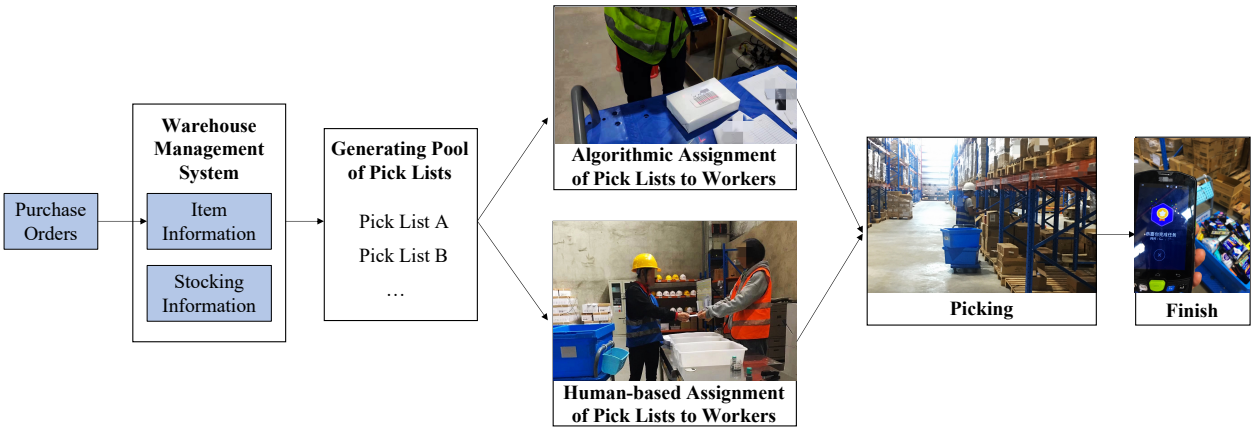


Figure 1.1: Flow Chart of Picking Process for Both Groups

pick list since she could access the pick list's information on her hand-held monitor. In the algorithm group, picking tasks were assigned to workers by an algorithm. Specifically, when a worker in the algorithm group came to the distribution station for her next pick list, she would scan a bar code on the algorithmic assignment table at the station. This would trigger the algorithm to randomly choose a pick list from the pool of available pick lists and then display the selected pick list on the worker's hand-held monitor. At that time Cainiao's system would record the starting time of this pick list.

From this point on, the process of completing a pick list was the same in the algorithm and human groups. Upon a pick list showing up on their monitor, workers in both groups would walk to the stocking location of the first SKU on the pick list. Once they found the first SKU and put the corresponding quantity into a cart, they scanned the bar code of the first SKU to record the time. When workers picked the last SKU in a pick list and scanned its bar code, the completion time of the pick list would be recorded. Figure 1.1 illustrates the pick list generation and picking process for both groups of workers.

Note that the pick list assignment process in both the human and algorithm conditions differed from how pick lists were assigned in this warehouse before and in between our

experiments. Thus, the preexperiment assignment process could not have served as an anchor that differently affected workers' perceptions of their assignment process during our experiments. Workers, human distributors, and the operation manager in this warehouse were unaware of the objectives of our experiments or our hypotheses, and they did not have information about the algorithm in use.

To cleanly examine workers' perceptions of algorithmic (vs. human-based) assignment processes and the subsequent implications for productivity, we took several measures to ensure that the accessibility of pick lists and the assignment outcomes are comparable between conditions so that the two conditions only differ in workers' *perceived* distributor (*i.e.*, algorithm vs. human).

First, in both conditions, pick lists assigned to workers at any given point were drawn from the same pool of pick lists using the same underlying rule. Specifically, in the human group, we instructed human distributors to give out pick lists in the same order as how pick lists were randomly selected from the pool and printed. This instruction prevented human distributors from handing out pick lists at their own discretion. In the algorithm group, the algorithm by design randomly selected a pick list from the pool of available pick lists each time. Therefore, in essence, workers in both groups received pick lists that were randomly drawn from a common pool of pick lists. Our following randomization check in Section 1.4.3 also confirms that the characteristics of pick lists are comparable between conditions.

Second, workers in both groups had to walk to the same location to obtain pick lists, which eliminated the effects of different walking distances on productivity. In other words, it cannot be the case that workers in one group walked less, were less fatigued, and thus were more productive than workers in the other group. Third, we sought to make the process of receiving pick lists equally simple for workers in both groups, so differences in productivity could not

be driven by how inconvenient workers found the assignment process. Indeed, as shown in 1.4.3, workers in both groups rated the process of receiving pick lists as similarly convenient.

We also tried to address the potential influence of operational transparency. Research about operational transparency suggests that when people have more knowledge about how things work during an operation process, their trust and engagement will be enhanced [41]. For example, showing customers who is serving them and how service is done can increase customers' service satisfaction [32, 41, 42]; and allowing employees to observe customers can improve service quality and efficiency [42]. Particularly related to our research, [43] find that people disfavor algorithms relative to their own decision-making processes because they perceive algorithmic decision processes as less transparent. To mitigate the influence of transparency in our field experiments, we decided not to tell workers how the human distributor or the algorithm assigned pick lists. In addition, as we have mentioned, the assignment process in both the human and algorithm conditions differed from how pick lists were assigned in the warehouse prior to our experiments; thus, workers in both conditions may similarly have uncertainty about how the assignment of pick lists was exactly determined.

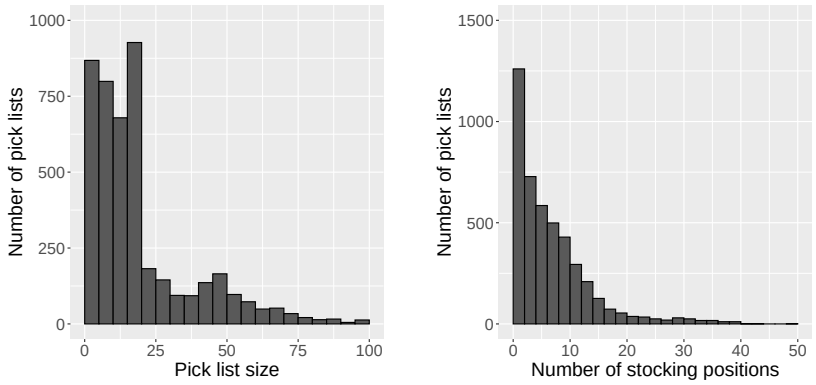
Another potential concern about our experimental design is the interference between workers; that is, the behavior of a particular worker may depend not only on her own pick list assignment process but also on the assignment process experienced by others in the warehouse (*e.g.*, because they may communicate with each other). We use a post-experiment statistical test [44] to show that the behavior of workers in our experiments was unlikely to have been affected by the assignment process of others working around them. We discuss this test later in Section 1.6.2. In Online Appendix C, we also explain why we consider our design (similar to [45, 46]) the cleanest among all feasible approaches given the constraints.

1.4.2 Data and Survey Design

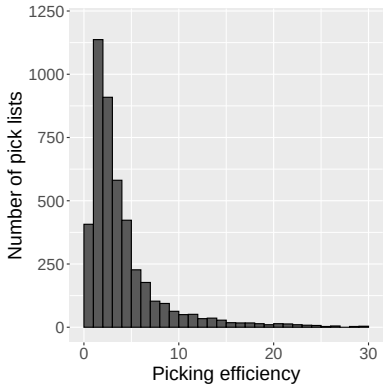
During our experiments, we collected two types of data: (1) operations data from the warehouse management system tracking the characteristics and processing time of each pick list, and (2) workers' responses to surveys that we administered every day. For each pick list, we tracked the pick list size (*i.e.*, the total quantity of items to be picked), the number of stocking positions (*i.e.*, the number of shelf positions in which SKUs in the pick list were stocked), the identifier of the worker who handled this pick list, and the times when the worker started versus completed the pick list.

On average, workers in the first experiment worked 2.16 days during the experimental period, yielding a total of 108 worker-day level observations. These workers completed 4,486 pick lists in total. Figure 1.2 provides the distributions of pick list characteristics across all pick list observations in our first experiment. Figure 1.2(a) shows the distribution of pick list size. Since the picking carts where workers temporarily stored products they collected had capacity limits, most (72.96%) pick lists contained no more than 20 items. Figure 1.2(b) shows the distribution of the number of stocking positions, which is highly skewed to the right. Since pick lists were designed to combine items in the same stocking position, the number of stocking positions was generally smaller than the number of items in a pick list. Figure 1.2(c) displays the distribution of picking efficiency. Picking efficiency equals the total quantity of items in a pick list divided by how long (in minutes) it took a worker to complete the pick list. It captures the average quantity of items a worker picked per minute while working on a pick list.

At the end of each day, we distributed surveys to all picking workers who showed up that day. Workers were told that their responses would be kept confidential and would be used exclusively for research purposes. Our daily survey collected workers' perceptions about their



(a) Pick List Size (b) Number of Stocking Positions



(c) Picking Efficiency

Figure 1.2: Distributions of Pick List Characteristics in the First Field Experiment

pick list assignment process as well as their demographics. When assessing people's attitudes towards algorithmic and human-based decision-making, prior research has often had people make head-to-head comparisons of these two methods [5, 6, 18]. Therefore, we developed two questions to assess workers' perceived fairness about their current assignment process (relative to the alternative assignment process; see Table 1.1).

First, we asked workers whether they thought it would be fairer to assign pick lists using the alternative process than using their current process. Specifically, workers in the algorithm group were asked, "Do you think it would be fairer if pick lists were assigned by a human distributor?" Workers in the human group were asked, "Do you think it would be fairer if pick lists were assigned by an algorithm?" Workers in both groups responded using a

Table 1.1: Measures of Fairness Perceptions

Question number	Group	Question wording
1	Algorithm	Do you think it would be fairer if pick lists were assigned by a human distributor? (1=Definitively would, 5=Definitively would not)
	Human	Do you think it would be fairer if pick lists were assigned by an algorithm? (1=Definitively would, 5=Definitively would not)
2	Algorithm	Which assignment process do you think would be more appropriate if you were paid by item instead of by time? (1=Definitively algorithmic assignment, 5=Definitively human-based assignment)
	Human	Which assignment process do you think would be more appropriate if you were paid by item instead of by time? (1=Definitively algorithmic assignment, 5=Definitively human-based assignment)

Note: The English translation does not match the Chinese version of our survey word by word, but it captures the meaning of our survey questions and scale response anchors well after considering the context.

five-point Likert scale from 1 ("Definitively would") to 5 ("Definitively would not"). In both groups, choosing a higher value (relative to a lower value) indicates that the worker viewed their current assignment process more favorably and less strongly believed the alternative assignment process would be fairer.

Second, we asked workers, "Which assignment process do you think would be more appropriate if you were paid by item instead of by time?" (with the five-point scale ranging from 1 = "Definitively algorithmic assignment" to 5 = "Definitively human-based assignment"). We framed the question this way because people are generally sensitive to fairness in task assignment when receiving performance-based incentives [47]. Thus, we expected workers to report what they deemed as a fairer assignment process when they were asked to pick their preferred assignment process under a piece-rate pay scheme. For workers in both groups, choosing a higher number in response to our second question indicates that the worker viewed human-based assignment more favorably. Since we wanted to compare between groups how fair workers believed their *current assignment process* to be, we reverse coded the responses of workers in the algorithm group so that a higher value instead would indicate that the worker viewed the *algorithmic* assignment process—their current assignment process—as fairer. Reverse coding scale items is a common practice in psychology and other fields that

use survey responses (*e.g.*, [48, 49]). Specifically, we used six to subtract the original answer of each worker in the algorithm group. For example, if a worker in the algorithm group gave an answer of one, the worker's reverse-coded answer would be five. For workers in the human group, we made no adjustment to their original answers. In the end, for workers in both groups, a higher (vs. lower) value indicates that the worker more strongly viewed their current assignment process as fairer than the alternative process.

Workers' responses to the aforementioned two questions (after we reverse coded the second one) are significantly and positively correlated ($r = 0.31$; $p = 0.001$). For each worker each day, we averaged her responses to these two questions to measure the extent to which she perceived her current assignment process as fairer than the alternative process (*Perceived Fairness*). In Appendix A.4, we show that our results are largely robust to separately using each fairness question. To facilitate the interpretation of how fairness perceptions affect productivity, we constructed *Standardized Perceived Fairness*, which equaled *Perceived Fairness* divided by its standard deviation in the whole sample. Moving forward, we report results using this standardized measure.

To evaluate the convenience of their assignment process, we asked workers, "How convenient do you feel it was to receive your pick lists today?" Workers responded using a five-point Likert scale (from 1 = "Very convenient" to 5 = "Very inconvenient"). We reverse coded their answers such that a higher value indicates greater convenience. We collected this measure to confirm that it did not seem easier to receive pick lists in one group than in the other.

To evaluate workers' emotional sensitivity to receiving difficult tasks, we asked workers how often they would feel upset if they received pick lists that were difficult to handle. Workers responded using a five-point Likert scale (from 1 = "Always" to 5 = "Never"). We reverse coded their answers to this question such that a higher value indicates that the worker was

more sensitive to task difficulty. We used this variable for an analysis of heterogeneous treatment effect.

Finally, we asked workers for their gender (female or male), education (middle school or under, high school, or college or above), residence (rural or urban), and age. Since two workers did not report residence (both of whom only worked one day during our first experiment), when we add demographic controls to regressions, two observations are dropped from regressions predicting perceived fairness and 71 observations are dropped from regressions predicting picking efficiency.

1.4.3 Randomization Check

To confirm that our randomization process was successful, we compare workers' demographics and the number of days they came to work in the warehouse between the algorithm and human groups. As shown in Panel A of Table 1.2, the proportion of females, education levels, the proportion of workers born in urban areas, age, and the number of work days during our experiment do not significantly differ between two groups. The Kolmogorov–Smirnov test further shows that the distributions of age and work days (the two continuous demographics variables) are comparable between the two groups of workers. These findings suggest that we have a comparable sample of workers between groups and thus our randomization process was successful.

In addition, workers' survey responses confirm that they found it similarly convenient to receive pick lists in the algorithm and human groups (Panel A in Table 1.2). Moreover, as mentioned earlier, an important feature of our experiment is that pick lists were distributed to workers in two groups using the same underlying process. Indeed, key pick list characteristics—pick list size and the number of stocking positions—are quite similar between groups (Panel

Table 1.2: Randomization Check

	Pick list assignment process		*Statistical test*		
	Human-based assignment	Algorithmic assignment	p-value of t-test	p-value of prop-test	p-value of ks-test
	(1)	(2)	(3)	(4)	(5)
	Panel A: Worker characteristics and perceived convenience				
Gender	0.40	0.44	–	0.77	–
	(0.50)	(0.51)			
Education	1.52	1.60	0.69	–	–
	(0.65)	(0.76)			
Residence	0.25	0.13	–	0.27	–
	(0.44)	(0.34)			
Age	29.36	24.88	0.08	–	0.28
	(9.17)	(8.35)			
Number of work days	2.40	1.96	0.41	–	0.91
	(2.00)	(1.37)			
Process	3.72	4.08	0.15	–	0.70
convenience	(0.98)	(0.76)			
Observations	25(24 for residence)	25(24 for residence)	–	–	–
	Panel B: Pick list characteristics				
Pick list size	20.09	21.04	0.12	–	0.12
	(20.48)	(20.62)			
Number of stocking positions	7.10	7.34	0.25	–	0.43
	(6.41)	(7.30)			
Observations	2,474	2,012	–	–	–

Note: The categorical variables in the table are defined as follows: gender = 0-male, 1-female; education = 1-middle school or under, 2-high school, 3-college or above; residence = 0-rural, 1-urban. Process convenience equals the average of a worker's responses across days if the worker showed up in our experiment for more than one day. Standard deviations are reported in the parentheses. "prop-test" refers to the two-sample proportion test, and "ks-test" refers to the Kolmogorov–Smirnov test. In addition to the tests reported in the table, we also predict pick list characteristics as a function of each worker's assignment group, following the ordinary least squares regression specification (2) described later, which further confirms that pick list characteristics do not significantly differ between conditions (p-values ≥ 0.37).

B of Table 1.2), confirming that the two groups of workers received pick lists of the same nature.

1.5 Main Results from Our Main Field Experiment

1.5.1 The Effect of Algorithmic Assignment on Perceived Fairness

We first test whether assigning pick lists via an algorithm boosts workers' perceived fairness about their task assignment process, relative to assigning pick lists via a human (Hypothesis 1). To test this hypothesis, we apply the following regression specification to worker-day level observations, with each observation representing worker i on day t:

$$\textit{Standardized perceived fairness}_{it} = \eta_0 + \eta_1 \textit{Algorithm}_i + \eta_2 X_i + \lambda_t + \epsilon_{it}, \tag{1.1}$$

where $\textit{Standardized perceived fairness}_{it}$ refers to worker i's standardized perceived fairness on day t, Algorithm_i is a binary variable equaling one if worker i was in the algorithm group and zero if worker i was in the human group, and X_i is the vector of demographics controls including worker i's gender, education, residence, and age. λ_t captures day fixed effects. We cluster standard errors at the worker level (our results are robust to clustering standard errors at the day level). We analyze fairness at the worker-day level because this is our most granular level of observation for capturing fairness, given that each worker provided their fairness perceptions once each work day.

We report results from specification (1.1) and its variants (with or without controls) in Table 1.3. In Column 1 (without control variables), a positive and significant coefficient on the indicator *Algorithm* ($p < 0.0001$) indicates that receiving pick lists from an algorithm significantly increases workers' perceived fairness about their assignment process, compared to receiving pick lists from a human distributor. Specifically, algorithmic assignment (relative to human-based assignment) increases perceived fairness by 0.94 standard deviations. This

Table 1.3: The Effects of Algorithmic (vs. Human-based) Assignment on Perceived Fairness and Productivity

Dependent variable	*Standardized perceived fairness*			*Picking efficiency*		
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm	0.94****	0.96****	1.02****	0.70***	0.68***	0.61***
	(0.20)	(0.20)	(0.23)	(0.27)	(0.23)	(0.19)
Day fixed effects	No	Yes	Yes	No	Yes	Yes
Hour fixed effects	No	No	No	No	Yes	Yes
Demographics controls	No	No	Yes	No	No	Yes
Pick list controls	No	No	No	No	No	Yes
Observations	108	108	106	4,486	4,486	4,415

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 3.92.

effect is robust when we control for day fixed effects (0.96 standard deviations, $p < 0.0001$; Column 2) as well as when we control for both day fixed effects and worker demographics (1.02 standard deviations, $p < 0.0001$; Column 3). Overall, these results support Hypothesis 1 that assigning pick lists by an algorithm (vs. a human) boosts workers' perceived fairness about their pick list assignment process.

We suspect that the positive effect of algorithmic assignment on fairness perceptions occurs because in our labor-intensive working environment where tasks are easier to be quantified, picking workers hold a strong equality motive for task assignments. To test this intuition, we distributed a survey to workers involved in our second field experiment (from December 27, 2019 to January 5, 2020) when they started their shift (see Appendix A.2). We asked workers whether they believed it is more important to ensure equality in task assignments or to customize task assignments based on workers' personal characteristics. Workers responded to this question using a five-point Likert scale from 1 ("Definitely prefer equality") to 5 ("Definitely prefer consideration of personal factors"). Workers' average response was 2.49 (95% confidence interval [2.28,2.69]), which is significantly lower than 3, the mid-point of the scale ($p < 0.0001$). This suggests that workers in our field setting on average consider it more important to ensure equality than to take into account everyone's personal characteristics in task assignment.

To further understand why workers perceive algorithmic assignment fairer than human-based assignment when they care strongly about equality, we conducted structured interviews with 13 workers after both of our field experiments ended. When asked whether a pick list assignment process run by a human distributor would be fair or unfair, more than half of workers (n = 7) indicated that a human-based assignment process might cause unfair outcomes. These workers mostly justified their judgment by mentioning that they believed human distributors are subject to personal biases. In addition, when asked whether they thought the assignment process would be more or less fair if they could receive pick lists by scanning a bar code (as opposed to from a human distributor), most workers (n = 10) believed that the process run by a machine would be fairer; and most of these workers (n = 8) explained that they believed an algorithmic assignment process does not fall prey to human distributors' personal preferences, would be able to deliver equal treatments across workers, and would not selectively favor or disadvantage certain workers. We present details about our interviews in Appendix A.5.

1.5.2 The Effect of Algorithmic Assignment on Productivity

We next test whether assigning pick lists via an algorithm (vs. a human) enhances workers' productivity (Hypothesis 2). To test this hypothesis, we apply the following specification to pick list observations:

$$\textit{Picking efficiency}_{ikt} = \delta_0 + \delta_1 \textit{Algorithm}_i + \delta_2 X_{ikt} + \lambda_t + \epsilon_{ikt}, \tag{1.2}$$

where $\textit{Picking efficiency}_{ikt}$ refers to the quantity of items worker i picked per minute for pick list k at time t, $\textit{Algorithm}_i$ is defined the same as in specification (1.1), and X_{ikt} is the vector of demographics controls (gender, education, residence, and age) and pick list controls (pick list size and the number of stocking positions). In addition to day fixed effects, λ_t also

includes hour fixed effects since pick list characteristics often change across hours within a day. We cluster standard error at the worker-hour level (our results are robust to clustering standard errors at the day-hour level). We analyze productivity at the pick list level because this is our most granular level of observation for capturing picking efficiency.

As shown in Columns 4-6 in Table 1.3, the coefficient on the indicator *Algorithm* is positive and statistically significant (all p-values ≤ 0.008) with or without controls, which means that the algorithmic assignment treatment significantly improves workers' productivity. Specifically, without control variables (Column 4), we estimate that assigning pick lists via an algorithm increases worker productivity by 0.70 items per minute, or 17.86% relative to the average picking efficiency of 3.92 in the human-based assignment group. When we add controls for time, worker demographics, and pick list characteristics, the effect remains statistically significant though decreases slightly in magnitude (17.35% in Column 5 and 15.56% in Column 6).

1.5.3 The Effect of Perceived Fairness on Productivity

Next, we estimate how workers' perceived fairness about their work assignment process affects their productivity. To causally estimate this effect, we take the instrumental variable (IV) approach and use the following specifications to explain our IV estimation:

$$\textit{Picking efficiency}_{ikt} = \alpha_0 + \alpha_1 \textit{Standardized perceived fairness}_{ikt} + \alpha_2 X_{ikt} + \lambda_t + \epsilon_{ikt} \quad (1.3)$$

and

$$\textit{Standardized perceived fairness}_{ikt} = \beta_0 + \beta_1 \textit{Algorithm}_i + \beta_2 X_{ikt} + \lambda_t + \epsilon_{ikt}. \quad (1.4)$$

Directly using specification (1.3) to estimate the effect of fairness perceptions on productivity does not yield a causal estimate because of the omitted variable bias. Unobserved variables, such as worker ability, can be correlated with both how fair workers believe they are treated and their productivity. Therefore, we use the random assignment of workers to the algorithm group as an IV for their fairness perceptions. The two-stage least squares estimate is given in specifications (1.3) and (1.4), and standard errors are clustered at the worker-hour level (our results are robust to clustering standard errors at the day-hour level). Though workers reported perceived fairness once each work day (which is why specification (1.1) has the notation *Standardized perceived fairness*$_{it}$), here we use *Standardized perceived fairness*$_{ikt}$ as the notation to indicate the level of observation (*i.e.*, pick list level) used in the two-stage least squares estimation.

To validate our IV estimation, we first check the *relevance assumption*: the IV *Algorithm*$_i$ should be correlated with the independent variable *Standardized perceived fairness*$_{ikt}$. As shown earlier in Table 1.3, the algorithmic (vs. human-based) assignment process significantly affects workers' perceived fairness at worker-day level under specification (1.1). We confirm that this effect is statistically significant at pick list level under specification (1.4) used in our IV estimation (all p-values < 0.0001 with or without control variables). Also, our IV passes the weak instrument test ($F = 1477.71$).

We next check the *exclusion restriction assumption*, which requires that the IV *Algorithm*$_i$ be independent of ϵ_{ikt} in specification (1.3). That is, assigning pick lists by an algorithm (vs. a human) should only affect productivity by altering the workers' fairness perceptions and should not be correlated with other factors that influence productivity. We think this assumption is satisfied for two reasons. First, since we randomly assigned workers to receive pick lists from either an algorithm or a human distributor, *Algorithm*$_i$ by design should not

be correlated with variables whose value was determined before the experiment (*e.g.*, worker characteristics), which we verify in Section 1.4.3.

Second, while it is impossible to statistically prove, we carefully designed our experiment to ensure that our experimental manipulation was unlikely to affect productivity via other mechanisms than fairness perceptions. During our structured interviews, we asked workers, "what factors usually influence your motivation and productivity?" The most frequently mentioned factors, brought up by 7 out of 13 workers, involve pick list characteristics including the number of items they have to collect and how many stocking positions they have to get products from. As explained in Section 1.4.1 and confirmed in Table 1.2, we ensured that human-based assignment and algorithmic assignment essentially used the same underlying rule. Another factor, which was brought up by 2 workers, is the convenience of obtaining pick lists. As explained in Section 1.4.1 and confirmed in Table 1.2, we made it similarly convenient to obtain pick lists between the algorithm and human groups. We also asked workers in the interviews, "besides pick list characteristics and the assignment process, what other factors may influence your productivity?" Workers brought up special circumstances (whether certain products are out of stocks, whether picking carts are temporarily unavailable), physical work environment (warehouse temperature, weather), and their physical well-being. All these factors should be comparable between two groups of workers since they worked in the same environment and were randomly assigned to the algorithm or human group. Furthermore, as discussed in Section 1.4.1, operational transparency is unlikely to be an alternative mechanism since workers in both groups were likely to have uncertainty about how pick lists were assigned.

Table 1.4 shows the average treatment effect of perceived fairness on productivity using IV estimation. We consistently find that workers' perceived fairness has a positive effect on productivity regardless of whether we include control variables (all p-values ≤ 0.009

Table 1.4: IV-Estimated Effect of Perceived Fairness on Productivity

Dependent variable	*Picking efficiency*		
	(1)	(2)	(3)
Standardized perceived fairness	0.71***	0.68***	0.55***
	(0.27)	(0.23)	(0.18)
Day fixed effects	No	Yes	Yes
Hour fixed effects	No	Yes	Yes
Demographics controls	No	No	Yes
Pick list controls	No	No	Yes
Observations	4,486	4,486	4,415

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency across algorithm and human groups was 4.24.

in Columns 1-3). Specifically, as perceived fairness increases by one standard deviation, worker productivity is estimated to significantly increase by 0.55-0.71 items per minute, or 12.97%-16.98% relative to the average pick efficiency of 4.24 across both algorithm and human groups.

1.6 Additional Results about Our Main Field Experiment

1.6.1 Persistence of the Treatment Effects of Algorithmic Assignment

In this section, we compare the effects of algorithmic (vs. human-based) assignment between the first day when a worker was involved in the experiment (hereafter, "the first experiment day") and later days during the experiment (hereafter, "subsequent experiment days"). This allows us to test whether our findings are primarily driven by the first experiment day when workers were new to the experiment and had little experience with either the algorithmic pick list assignment process or the revised human assignment process.

We first analyze all workers in our main field experiment, and split the sample based on whether an observation was associated with a worker's first experiment day. As shown in

Table 1.5: Effects of Algorithmic (vs. Human-based) Assignment Broken Down by First Experiment Day versus Subsequent Experiment Days

	(1)	(2)	(3)	(4)
Dependent variable:	Standardized perceived fairness		Picking efficiency	
Subsample:	First experiment day	Subsequent experiment days	First experiment day	Subsequent experiment days
Algorithm	1.52****	0.67***	1.00***	0.53**
	(0.35)	(0.23)	(0.36)	(0.25)
Day fixed effects	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes
Pick list controls	No	No	Yes	Yes
Observations	48	58	1,845	2,570

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 4.01 for workers on first experiment day and 3.87 for workers on later experiment days.

Table 1.5, the effects of algorithmic (vs. human-based) assignment on perceived fairness and productivity are positive and significant both on the first experiment day and subsequent experiment days (all p-values ≤ 0.032). This indicates that the effect of algorithmic assignment is not limited to the first day when workers might find algorithmic assignment novel and instead persists on later days after workers have gained more experience with it.

As robustness checks, we perform the same set of analyses on two additional samples. First, we focus on a filtered sample that includes only workers who participated in the experiment for more than one day, so we ensure that observations about the first experiment day came from the same set of workers as observations about subsequent experiment days. Second, we create a combined sample of workers who participated in either the first or the second field experiment for more than one day. For both samples, we consistently find that algorithmic assignment has significantly positive effects on fairness perceptions and productivity for subsequent experiment days.

1.6.2 Checking Interference Between Workers

Following [44], we use a statistical test to check whether the behavior of a particular worker depends only on her assignment process, not on the assignment process of others working

around her. This is an ex post method to detect interference between units (where each unit represents a worker in our context) in a randomized experiment. In our setting, interference between workers could occur if (1) workers in the human group viewed their assignment process as less fair and became less motivated (relative to workers in the algorithm group) because they learned that workers in the other group scanned a bar code to get their pick lists; or (2) if workers in the algorithm group perceived their assignment process fairer and became more motivated (relative to workers in the human group) because they learned that workers in the other group got their pick lists from a human distributor.

The idea of this test is to randomly draw simulations to rearrange workers into either the algorithm group or the human group, recalculate the treatment rate each day (which refers to the proportion of workers in the algorithm group that day in a simulation), and estimate the relationship between the daily treatment rate and each worker's focal outcome variables of interest (either productivity or perceived fairness). Specifically, we first randomly select 12 workers in the human group as a fixed subset and randomly re-assign the remaining 38 workers (the variant subset) in our first field experiment to be in either the human group or the algorithm group each day. For each simulation (*i.e.*, each time we re-assign the 38 workers), we compute the Spearman's rank correlation coefficient ρ between the simulated daily treatment rate and workers' productivity (or perceived fairness) across workers in the fixed subset across days they came to work. Across 1,000 simulations, we obtain 1,000 values of ρ. The distribution of the 1,000 ρs represents the approximate distribution of ρ associated with the null hypothesis that interference on productivity between workers did not occur for workers in the human group, since the 38 workers in the variant subset were purely randomly assigned to the algorithm vs. human condition in each simulation and 1,000 simulations were independent. If the productivity and perceived fairness of workers in the human group were actually affected by those in the algorithm group as a result of interference between groups,

these outcome variables should be strongly correlated with the proportion of workers in the algorithm group on a day in our actual data. However, the observed correlation coefficient ρ in our first field experiment does not significantly deviate from the center of null distribution. Therefore, we cannot reject the null hypothesis that there is no interference on productivity for workers in the human group.

To confirm the robustness of our test (and ensure that the null effect is not unique to the specific fixed subset we have drawn), we randomly re-draw 12 out of 25 workers in the human group as the fixed subset for 1,000 times, and each time a fixed subset is selected, we repeat the process described above involving 1,000 simulations. This additional step further suggests that the perceived fairness and productivity of workers in the human group are unlikely to have been affected by the interference between the algorithm and human groups. We also go through the same simulation process by treating workers in the algorithm group as the fixed subset. Similarly, we show that workers in the algorithm group are unlikely to have been affected by the interference, either. See details about our test in Appendix A.3.

1.6.3 Heterogeneous Treatment Effects Based on Task Difficulty and Sensitivity to Difficult Tasks

We have also explored how the effects vary across workers and tasks. Understanding for whom and for what type of task algorithmic assignment yields a greater impact may not only shed light on why algorithmic assignment boosts fairness perceptions and productivity in our setting but also suggest to managers who will benefit the most from algorithmic assignment processes. For example, we find that algorithmic assignment (vs. human assignment) improves productivity to a larger extent among workers with at least a high school degree than workers with a lower level of education—a finding we report in detail in Appendix A.6 in the interest

of space. In this section, we explore whether the effects of algorithmic assignment change with task difficulty and workers' sensitivity to task difficulty.

Heterogeneous Treatment Effects Based on Task Difficulty

The number of stocking positions is a key characteristic that reflects the difficulty of a pick list. A pick list with more stocking positions usually requires workers to walk more. Besides, the number of stocking positions was mentioned most frequently by workers in our interviews as a factor that could affect the difficulty of a pick list. Therefore, we use the number of stocking positions in a pick list as a measure of task difficulty (see Appendix A.6 for more details about this variable).

As mentioned in Section 1.5.1, workers in our setting on average believe that it is more important for a pick list assignment process to achieve equality than to consider individuals' characteristics. This emphasis on the equality motive is in line with prior research suggesting that people exhibit aversion to both disadvantageous inequity (when they are worse off than others) and advantageous inequity (when they are better off than others) [20, 21, 22]. Under the equality motive, receiving particularly difficult tasks from a human distributor may be perceived by workers as reflecting the distributor's bias against them, but receiving particularly easy tasks may also be attributed by workers to the distributor's intentional decisions and be viewed as reflecting the distributor's preference in favor of them. In other words, as long as pick lists assigned by a human distributor obviously deviate from the middle range of task difficulty, workers may attribute the deviations to the human distributor's personal biases and cast doubt on the fairness of human-based assignment process. Thus, workers may tend to view human-based assignment as less fair than algorithmic assignment when they receive either particularly difficult tasks or particularly easy tasks.

Table 1.6: Effects of Algorithmic Assignment Broken Down by Task Difficulty

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Standardized perceived fairness			Picking efficiency		
Workload characteristics:	Low task difficulty	Medium task difficulty	High task difficulty	Low task difficulty	Medium task difficulty	High task difficulty
Algorithm	1.57****	0.73	1.09*	0.64***	0.07	0.81**
	(0.33)	(1.06)	(0.56)	(0.20)	(0.17)	(0.33)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	No	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Pick list controls	No	No	No	Yes	Yes	Yes
Observations	52	29	25	4,102	261	52

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 3.95 for tasks with low difficulty, 3.64 for tasks with medium difficulty, and 3.16 for tasks with high difficulty.

To test this possibility, we calculate the average stocking positions across all pick lists assigned to a given worker on a given day, and we then categorize the worker-day level observations into three subsamples by splitting the range between the minimum (2.41) and maximum (15.17) of average stocking positions into three equal intervals. Specifically, if a worker's average stocking positions on a day were in the interval of [2.41,6.66], this worker on average faced *low task difficulty* that day; the interval of (6.66,10.92] represents *medium task difficulty*; and the interval of (10.92, 15.17] represents *high task difficulty*. As shown in Table 1.6, when task difficulty is either lower or higher than normal, algorithmic (vs. human-based) assignment (marginally) significantly increases fairness perceptions by 1.57 or 1.09 standard deviations ($p < 0.0001$ in Column 1 and $p = 0.078$ in Column 3) respectively. Under medium task difficulty, algorithmic assignment does not significantly increase fairness perceptions ($p = 0.51$ in Column 2), and the estimated (insignificant) treatment effect is directionally smaller than the estimated effect under low and high task difficulty. These results provide suggestive evidence for the speculation that algorithmic assignment boosts fairness perceptions (relative to human-based assignment) both when people receive particularly easy or particularly hard tasks.

To explore the further implications for productivity, we similarly categorize pick lists into three subsamples by splitting the range between the minimum (1.00) and maximum (49.00) of the number of stocking positions into three equal intervals. Specifically, if the number of stocking positions on a pick list was in the interval of [1.00,17.00], (17.00,33.00], (33.00,49.00], we deem this pick list as having a low, medium, or high task difficulty, respectively. As shown in Table 1.6, when a pick list is associated with low task difficulty, algorithmic assignment significantly boosts picking efficiency by 0.64 items per minute (or 16.20% relative to the average efficiency in the human group among low difficulty pick lists; p = 0.001 in Column 4); and when a pick list is associated with high task difficulty, algorithmic assignment significantly boosts picking efficiency by 0.81 items per minute (or 25.63% relative to the average in the human group among high difficulty pick lists; p = 0.022 in Column 6). There is not a statistically significant effect of algorithmic assignment on picking efficiency when workers handle pick lists with medium difficulty (p = 0.66 in Column 5), and the estimated (insignificant) treatment effect is directionally smaller than the estimated effect under low and high task difficulty.

These patterns are consistent with our speculation that since workers in our setting on average hold the equality motive, they view substantial deviation from the medium range of task difficulty as reflecting a human distributor's intentionality and perceive the human-based assignment process as less fair than algorithmic assignment. We acknowledge that the number of observations in each subsample is small, especially when we analyze productivity in the medium and high task difficulty subsamples. Thus, we present the results as suggestive evidence. We have also tried splitting both worker-day and pick list level observations into two subsamples based on the corresponding median value in the full sample and reported the results in Appendix A.6 for full transparency. We hope that reporting the suggestive evidence can encourage future research to more systematically integrate inequality aversion with workers' attribution of assignment outcomes.

Heterogeneous Treatment Effects Based on Worker Sensitivity to Task Difficulty

We next explore whether the effects of algorithmic assignment change with workers' sensitivity to task difficulty. In the field experiment, we asked workers how often they felt upset when they received difficult tasks. We use responses to this survey question to assess workers' sensitivity to task difficulty. For workers who worked for more than one day during our experiment, we take the average of their responses across days. The median of this measure was 2.00 across workers. We treat workers whose average response was higher than 2.00 as "high sensitivity" workers (n = 21) and workers whose average response was equal to or lower than 2.00 as "low sensitivity" workers (n = 28). For these two types of workers, we first separately estimate the effect of algorithmic (vs. human-based) assignment on perceived fairness using specification (1) and on productivity using specification (2); then we run an interaction model whereby we predict fairness perceptions and productivity as a function of the algorithmic treatment indicator, a binary variable indicating whether a worker is in the "high sensitivity" category, and the interaction of these two indicators.

As shown in Table 1.7, among high-sensitivity workers, algorithmic assignment is viewed as significantly fairer than human-based assignment by 1.59 standard deviations ($p < 0.0001$ in Column 1). The effect holds among low-sensitivity workers though with a directionally smaller size: among this subsample, algorithmic assignment significantly increases perceived fairness by 1.15 standard deviations, relative to human-based assignment ($p = 0.014$ in Column 2). The treatment effect of algorithmic assignment on fairness perception is directionally but not significantly amplified among high-sensitivity workers relative to low-sensitivity workers ($p = 0.36$ in Column 3).

Table 1.7: Heterogeneous Treatment Effect Based on Sensitivity to Task Difficulty

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Standardized perceived fairness			Picking efficiency		
Subsample of workers:	High sensitivity	Low sensitivity	All sample	High sensitivity	Low sensitivity	All sample
Algorithm	1.59****	1.15**	0.81**	1.34****	−0.33	0.15
	(0.32)	(0.44)	(0.39)	(0.26)	(0.24)	(0.24)
High Sensitivity			−0.64			−0.11
			(0.41)			(0.22)
*Algorithm * High Sensitivity*			0.49			0.84***
			(0.54)			(0.29)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	No	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Pick list controls	No	No	No	Yes	Yes	Yes
Observations	53	53	106	2,311	2,104	4,415

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 3.90 among high-sensitivity workers and 3.95 among low-sensitivity workers.

Further, as shown in Table 1.7, algorithmic assignment boosts the productivity of high-sensitivity workers by 1.34 items per minute (or 34.36% relative to the average picking efficiency of high-sensitivity workers in the human group; $p < 0.0001$ in Column 4), but does not significantly impact the productivity of low-sensitivity workers ($p = 0.16$ in Column 5). The treatment effect of algorithmic assignment on productivity is significantly amplified among high-sensitivity workers relative to low-sensitivity workers ($p = 0.003$ in Column 6).

Altogether, using all observations in the field experiment, we present suggestive evidence that the assignment process (algorithmic vs. human-based) affects high-sensitivity workers more than low-sensitivity workers. Interestingly, we also replicate this pattern when we examine only the subset of observations whose task difficulty is above the median (see Appendix A.6).

1.7 Online Experiments Assessing the Impact of Algorithmic Assignment on Fairness Perceptions

Following our field experiments, we conducted two additional online scenario-based experiments with survey respondents in the United States to replicate the effect of algorithmic (vs. human-based) assignment on perceived fairness. Online experiments have often been used to complement field studies [42, 50]. In their recent book chapter about field experiments, [51] highlighted that "lab and field should be seen as complements rather than substitutes; in particular... researchers can go back to the lab after field experiments" (p. 125).

We designed online experiments to complement our field experiments in a few ways. First, we intended to demonstrate the generalizability of our findings about fairness perceptions and the preference for equality in the warehouse context. By using a larger sample of participants under a different culture than the population in our field experiments, the online experiments help us address external validity concerns about the field results being specific to the Alibaba warehouse workforce. Second, both of our field experiments measured workers' fairness perceptions about their current assignment process by asking them to compare the algorithmic assignment process with the human-based assignment process. In our online experiments, we measured people's perceptions about their current assignment process without drawing any comparison with alternative assignment processes. Third, although we adopted the best design possible in our field experiments, workers could communicate with each other across groups. In addition to leveraging the ex post method discussed in Section 1.6.2 to detect interference between experimental groups, we seek to rule out the influence of interference in our online experiments. Since in the online experiments, participants did not know about alternative assignment processes, the contamination effect could not have played a role in the online setting. Our two online experiments followed the same design with one

exception and yielded consistent results. We report one of the experiments below and detail the other experiment in Appendix A.7.

1.7.1 Experimental Design and Analysis

We recruited study participants from an online labor marketplace, Amazon's Mechanical Turk, to complete a 4-minute study in exchange for $0.60. Only people who accessed our study on a non-mobile device, successfully completed a CAPTCHA, and passed an attention check were allowed to start our study. People who satisfied these criteria were asked to imagine themselves as a warehouse picking worker and read about descriptions of picking tasks. To ensure that participants understood this work setting and could immerse themselves into the scenario, we required that participants had to correctly answer three questions about the scenario in order to continue with the study. Those who passed our comprehension check questions and completed our study (N = 201; 41.29% female, $M_{age} = 39.47$) comprised our final sample.

Upon passing our comprehension check questions, participants were randomly assigned to either the algorithm condition (n = 100) or the human condition (n = 101). The descriptions of the two conditions mimicked the set-up in our field experiments. In the algorithm condition, participants were told that their pick list assignment process was run by a machine and that they received pick lists by scanning a bar code marked at the distribution station. In the human condition, participants were told that their pick list assignment process was run by a human and that they received hard-copy pick lists from a manager at the distribution station. Then participants in both conditions were told that the average pick list size (or the average number of items in a pick list) in the warehouse was 21 (based on the actual average pick list size in our main field experiment). Participants were also presented with their average pick list size on each of the past 10 workdays, and this information was the same between the

algorithm and human conditions. We presented this information about pick list size and kept it the same between conditions so as to control for pick list assignment outcomes and cleanly investigate people's perceptions of a given assignment process, as we did in the field. In our other online experiment, we did not provide information about pick list size and obtained similar results (see Appendix A.7).

One assumption underlying our Hypothesis 1 is that people believe humans are subject to personal biases and algorithmic assignment processes are more capable of delivering equal treatments across workers. In our field setting, most workers in our interviews did express such beliefs (as we discussed in Section 1.5.1). To test this assumption in our online experiment, we asked participants to indicate their agreement with the following statement about the assignment process they imagined getting pick lists from (either algorithmic or human-based): "I think this assignment process would treat every worker perfectly equally" (from 1 = "Strongly disagree" to 7 = "Strongly agree"). Choosing a higher (vs. lower) value indicates that the participant viewed their assignment process as more capable of preserving equality.

Then we assessed fairness perceptions about an assignment process by asking participants to indicate their agreement with four statements adapted from [52] and [10]: (1) "the way this warehouse assigns pick lists seems fair," (2) "the warehouse's process for distributing pick lists is fair," (3) "the decision regarding whether I get more difficult pick lists is fair," and (4) "the outcome of the pick list distribution is fair." The anchors on the scale ranged from 1 ("Strongly disagree") to 7 ("Strongly agree"). Choosing a higher (vs. lower) value indicates that the participant viewed their assignment process as fairer. Participants' ratings of these four statements reached a high inter-item reliability (Cronbach's alpha = 0.96) and were thus averaged to form a composite score of *Perceived Fairness*. Following the analysis in our

field experiment, we constructed *Standardized Perceived Fairness*, which equaled *Perceived Fairness* divided by its standard deviation in the whole sample.

Next, to check whether people in our online experiment held a strong equality motive for the assignment of picking tasks, we measured the perceived importance of equality and uniqueness. Specifically, we asked participants to separately rate how important they thought it was for a pick list assignment process to treat all workers equally and how important it was for a pick list assignment process to take into account individual workers' characteristics. The anchors on both scales ranged from 1 ("Not important at all") to 7 ("Very important"). Choosing a higher value indicates a higher perceived importance. In addition, we also used a single item as in our field setting and asked participants which of the two objectives they thought the warehouse should prioritize when it comes to assign picking tasks: treating all workers equality or taking into consideration personal characteristics. We obtained consistent results using these two methods to examine the relative importance of the equality versus uniqueness motive (see Appendix A.7), and focus on the former in the chapter. Finally, participants reported their gender, age, and education.

1.7.2 Results

By comparing participants' importance ratings for equality versus uniqueness, we first confirm that people on average prioritize equality over uniqueness in the warehouse task assignment setting ($M_{equality} = 5.88$, $SD = 1.27$ vs. $M_{uniqueness} = 4.53$, $SD = 1.71$; $t(200.00) = 8.84$, $p < 0.0001$ for a paired t test, Cohen's $d = 0.90$). Second, supporting the assumption underlying our Hypothesis 1, people view the assignment process run by a machine as more capable of preserving equality than the assignment process run by a human ($M_{algorithm} = 4.95$, $SD = 1.12$ vs. $M_{human} = 4.36$, $SD = 1.13$; $t(198.61) = 2.74$, $p < 0.001$, Cohen's $d = 0.73$). Further, in support of Hypothesis 1, participants in the algorithm condition perceived their

assignment process fairer than those in the human condition ($M_{algorithm} = 4.23$, $SD = 0.99$ vs. $M_{human} = 3.79$, $SD = 0.96$; $t(198.70) = 3.20$, p < 0.005, Cohen's $d = 0.45$). Note that since the variances are unequal between groups, we report degrees of freedom that have been adjusted for variance.

Chapter 2

The Value of Logistic Flexibility in E-commerce

2.1 Introduction

E-commerce has been expanding aggressively and taking over the retail world in the last several years. In the United States, the proportion of e-commerce sales in total retail sales was 13.2% in 2021. From 2017 to 2021, the total e-commerce revenue in the United States has nearly doubled from 443.2 billion dollars to 870.8 billion dollars (US Census Bureau Annual Retail Trade Survey Quarterly E-Commerce Report released in February 2022[1]). Meanwhile, we have observed the world's fastest e-commerce growth in some Southeast Asia countries. This number has even grown ten times from 9.0 billion dollars to 90.2 billion dollars during the same period in Indonesia (Statista Global Consumer Survey 2021[2]).

[1]See https://www.census.gov/retail/index.html#arts.
[2]See https://www.statista.com/outlook/dmo/ecommerce/indonesia.

This transition from offline to online retail is partially driven by the reduction of shipping time. Therefore, improving delivery speed is a core consideration in competing e-commerce platforms. According to a survey of 250 merchants across different industries, fast delivery drives more repeat customers and better online reviews—and helps beat competitors (Propeller Insights 2020[3]). Research has also shown that increasing shipping speed stimulates online store sales [53]. Therefore, platforms have been increasing their investments in reducing shipping time. Amazon, for instance, has developed a logistics system and offered same-day delivery for consumers who subscribe to its Prime membership. From 2017 to 2021, the company's annual shipping costs more than tripled from 21.7 billion dollars to 76.7 billion dollars (Amazon Annual Reports 2017-2022[4]).

However, it is becoming increasingly expensive to improve shipping speed. In the United States, a delivery option shift from five-day ground delivery to two-day express delivery could increase the shipping cost three to six times (*e.g.*, FedEx Shipping Rates[5]), which creates a decreasing marginal return of investment in shipping time. Many retailers have reached the point where further investment in reducing shipping time may mean not breaking even; they may only be able to afford to offer free shipping for ground delivery. To increase consumer purchases, retailers are exploring other ways to improve the delivery experience, especially by pursuing greater convenience, higher transparency, and better communications in logistic services.

One of the most prominent shipping strategies to provide consumers better logistic experiences is offering them the flexibility to pick up orders at a local station. In traditional e-commerce deliveries, the delivery time window is usually determined by third-party logistics providers, such as FedEx. Consumers have little freedom to decide what time of day they will receive

[3]See https://ware2go.co/faster-shipping-to-drive-a-competitive-advantage-for-merchants/.
[4]See https://ir.aboutamazon.com/annual-reports-proxies-and-shareholder-letters/default.aspx.
[5]See https://www.fedex.com/en-us/online/rating.html#.

their package. Even if they can decide on a time window to receive packages, they typically must make this decision when the item is shipped. This creates a large inconvenience cost to consumers since they may have to wait at home to sign for the package. However, if the package is first shipped to a pick-up station, a consumer would then be able to choose the pickup time according to their preference after the item has arrived at the station. Although the platform would need to pay the pick-up station for taking care of the package by piece rate, it can reduce logistic costs by batching last-mile deliveries. Meanwhile, there could also be inconvenience costs to consumers as they need to pick up the package in person from the station. Since different inconvenience costs exist in both home and pick-up station deliveries, it is unclear whether and to what extent pick-up stations would benefit consumers in various parts of a city.

In this chapter, we use pick-up stations to study an alternative core aspect of the delivery experience other than shipping speed–improving the flexibility for consumers to pick up orders at their convenience (hereafter denoted as "the logistic flexibility"). Specifically, we seek answers to the following research questions: Do pick-up stations have a positive impact on consumer purchases on the online retailing platform? If so, what is the mechanism behind this positive impact? How could these mechanisms help us to develop better pick-up station strategies, such as determining locations, as well as better shipping strategies?

To answer these questions, we collaborate with the Alibaba Group, a leading e-commerce platform in China with a nationwide logistics service. The platform has been setting up pick-up stations to extend its delivery capacity over the last several years. In 2022, it already had nearly 100 thousand stations in China that handle over 50 million packages per day on average. On a high level, we combine three data sets to conduct our research. The first data set tracks package-level logistic and transaction records for both pick-up stations and home delivery. The second data set contains consumer characteristics information. The third

data set includes the locations and opening times of pick-up stations. The combination of these three data sets enables us to trace consumers within the affected distance around the newly opened pick-up stations and observe changes in their logistic decisions and purchasing behavior.

The data sample we use in this study is based on the logistic records in Shenyang, the largest city in Northeast China. The observation period runs from June to December 2019. We observe 55 pick-up stations open in Shenyang in September 2019. Using the introduction of pick-up stations as a natural experiment for consumers, we apply the difference-in-difference (DiD) approach to compare consumers who are just inside and outside the service region of these newly opened pick-up stations. We find that the store opening has increased sales by 3.9%, and such increases mostly come from existing consumers. More importantly, we find empirical evidence that the effect of shipping speed could not be the primary driver of this increase since the average order-to-receive time for pick-up station packages is longer than that for home-delivery packages. Instead, we argue that logistic flexibility should be the primary driver of this increase.

To quantify the impact of logistic flexibility introduced by pick-up stations and conduct counterfactuals to explore better shipping strategies, we build a two-stage structural model that explicitly models how logistic flexibility may affect consumer choices. In our model, consumers' utility of receiving the package during a certain time in a day consists of a time preference that may change on a day-to-day basis. At the first stage of our model, consumers choose among home delivery, pick-up station delivery, and outside option that may reflect the utility of purchasing on other platforms, purchasing offline, or not purchasing. Facing the uncertainty of home or pick-up station delivery times (*i.e.*, the time when the packages arrive at home or pick-up station), at the second stage, consumers' true time preferences will be realized. If consumers have chosen home delivery, the delivery time and the idiosyncratic

shocks for each time preference will be realized in the second stage and in turn, consumers' utility will also be realized. If consumers have chosen the pick-up station, the delivery time to the pick-up station and the idiosyncratic shocks for each time preference will be realized, and consumers can then choose the optimal time to pick up the package from the station. Notice that our model allows two types of flexibility introduced by the pick-up stations: First, consumers may pick up at their preferred hour; second, consumers could postpone their decisions about the preferred hour until after their true time preference is realized. However, there is also a traveling cost induced by the need to go to the station, which increases with the distance.

After estimating our model, we first demonstrate that consumers generally prefer to receive the package after work hours. Specifically, when dividing a day into three intervals (10 a.m.-4 p.m., 4 p.m.-8 p.m., and other operation hours), if a consumer could get her package during after-work hours (4 p.m.-8 p.m.) instead of daytime hours (10 a.m.-4 p.m.), the consumer is willing to wait for 8.523-11.627 more hours on average. This demonstrates that consumers value the flexibility of choosing a preferred time slot. We denote this as the value of time flexibility. Consumers also gain utility from the realization of idiosyncratic shocks before picking up from stations. In other words, consumers enjoy the flexibility to make pick-up choices until the last minute to accommodate time uncertainty, which we denote as the value of choice flexibility. Running a counterfactual simulation shows that the overall improvement of opening a pick-up station can be broken down into these two benefits: 76.2% corresponds to the value of pickup time flexibility and 23.8% comes from the value of choice flexibility. Our estimation results also reveal that there are three classes of consumers. While all three classes of consumers exhibit qualitatively similar preferences, compared to a small segment (18% of consumers), the other two segments (71% and 11% of consumers, respectively) have

a higher value in choice flexibility and time flexibility. Therefore, the pick-up station is more valuable for these two segments.

Motivated by our estimation results, we conduct two counterfactual studies to improve the logistics decisions of Alibaba with the knowledge of consumers' time and flexibility preferences and their traveling costs. The first counterfactual study aims to find the optimal number of pick-up stations at optimal locations to improve the purchase rate. We run a greedy algorithm considering consumers' logistic flexibility utilities. This new location strategy can improve the purchase rate by 2.1%-6.9% and improve consumer welfare by 2.0%-7.5% using the same number of pick-up stations. We can also use our algorithm to decrease the number of pick-up stations by 56.4%-63.6% without decreasing the purchase rate. Most of the benefits from this strategy compared to the existing one is that we find new locations with a high density of consumers who are more likely to make purchases from the platform. The second counterfactual study tries to improve the shipping policy for home delivery instead of setting up pick-up stations that may be costly for Alibaba. Because different classes of consumers may have different time preferences, the shipping strategy can prioritize the consumers who feel the most pain from receiving packages during inconvenient hours (*e.g.*, during work hours). In particular, we arrange for more consumers to have their preferable delivery windows subject to the constraint of delivery capacity, and such shipping rearrangement can potentially improve the purchase rate by 8.4% and improve consumer welfare by 10.0%.

This chapter strives to make three main contributions. First, we study the impact of an offline logistic channel on the online sales channel. In particular, we demonstrate the impact of pick-up stations on consumer purchases. Second, in contrast with prior research on e-commerce logistics that mainly focused on delivery speed, we show that the impact of pick-up stations is driven by logistic flexibility, which is an alternative core aspect of the delivery experience. Third, we provide the first empirical model to study the value of logistic

flexibility, a new direction for improving logistic service. In our model, we further decompose two types of logistic flexibility—time flexibility and choice flexibility. Our work has several critical managerial implications. With the structural model, we can find a better pick-up station locations policy, which could assist the platform to achieve the same sales lift from logistic flexibility with fewer stations. We also use our model to help the platform explore a better shipping policy and put more consumers in their preferable delivery windows. Our counterfactual policies could potentially create billions in annual sales increases in CNY for the platform.

The remainder of the chapter is organized as follows. Section 2.2 reviews the literature related to our study. Section 2.3 describes the empirical setting and data. Section 2.4 shows the reduced-form evidence of the impact of pick-up stations on the e-commerce business and the potential mechanism. We develop our structural model and present the corresponding estimation results in Section 2.5 and 2.6 and conduct several counterfactual analyses in Section 2.7. Finally, we draw the conclusions in Section 2.8.

2.2 Literature Review

Our work is mainly related to three streams of prior literature: e-commerce logistics, e-commerce business strategies, and operations flexibility.

First, our study is closely connected to the growing stream of literature studying the impact of logistic services on e-commerce sales. Most of this literature has focused on delivery speed. Research in this area has found that the reduction of shipping time generates referrals in consumer acquisition [54] and increases future orders or sales for the platform [53, 55, 56, 57]. Especially, [53] uses the DiD approach to identify the causal impact of unannounced faster deliveries resulting from the opening of a new distribution center. On the contrary, the

delay of deliveries results in consumer dissatisfaction and a reduction of repurchase intentions [58]. Under the demand for faster and more reliable delivery speed, recent studies forecast delivery speed for shipping time promise policy [59] and optimize delivery speed considering the trade-off with inventory operation costs [60], assortment planning [61], and public safety [62]. Only a small proportion of this literature studies other aspects of e-commerce logistic services to encourage consumer purchases such as better word of mouth about logistic service [63] and offering a high-quality delivery option [64]. [65] examines how consumers respond to operational transparency in parcel delivery. In contrast with prior research on e-commerce logistics, we show that shipping time reduction is not the driving factor on the impact of pick-up stations. We study an alternative core aspect of delivery experience to affect sales—logistic flexibility.

Second, our research contributes to the literature studying e-commerce business strategies, such as optimizing pricing policy [66], sharing inventory information [67], decreasing search frictions [68, 69] and applying business innovations [70, 71]. This chapter extends this literature by introducing business innovations in advancing logistics functionalities. In particular, our work is related to a stream of literature studying the integration of online and offline retail channels, or omnichannel retail [72]. Research has found that retail channel integration creates a shift of consumers from online to offline channels [73], increases sales dispersion [74], and expands fulfillment flexibility [75]. Research has shown that this business implementation does not always benefit the retailer. The shift of consumers across channels may hurt profits when the store fulfillment is less cost-effective than the online fulfillment [76] or when both store visiting and online waiting costs are high [77]. While prior research in this area has focused on the shift of consumers and sales across offline and online sales channels, we complement prior research by studying the impact of the offline logistic channel

(pick-up stations) on the online sales channel (e-commerce platform). We explicitly model how sales on the online channel are lifted by the offline logistic channel.

Third, this chapter relates to the literature on store entry, in particular, spatial treatment and location selection. Building on the classic literature of game theoretic entry models [78, 79], recent literature incorporates spatial correlations and studies choice of locations in developing store networks, considering the local competition, consumer transportation, and distribution cost [80, 81, 82, 83]. [82] finds there is a trade-off between population density and income level in selecting store locations. [83] combine demographics and the retailer's historical sales to predict demand at potential locations. Furthermore, [84] quantify the impact of e-commerce access on domestic consumer welfare in reducing spatial inequality. We extend this literature by examining the impact of location selection of pick-up stations on sales in the online channel.

Finally, our work builds on the vast literature studying operational flexibility in using multipurpose resources, [85, 86], and especially, the large body of literature studying supply chain and manufacturing flexibility. This literature studies the flexibility of the manufacturing process [87, 88], the flexibility of the manufacturing resource [89, 90] and the flexibility of distribution. Specifically, an emerging stream of literature considers the flexibility of the distribution process in e-commerce [91, 92, 93]. We contribute to this literature by providing the first research on the flexibility in the last-mile delivery of the e-commerce supply chain.

2.3 Empirical Setting and Data

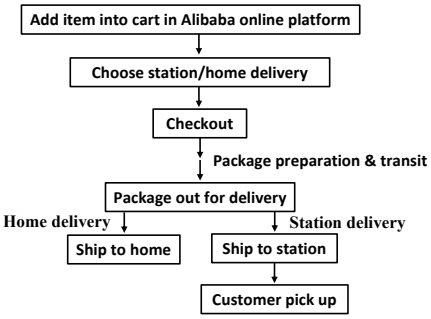
2.3.1 Empirical Settings

We collaborate with Alibaba (hereafter denoted as "the platform"), a leading e-commerce platform in China that has a nationwide logistics service. The platform handles a gigantic number of packages in China: The daily average package amount is more than 200 million; during annual promotion days, such as the double eleven holiday (*i.e.*, November 11th), the daily number of packages could even reach more than 3,000 in a single residential community. Therefore, the platform is establishing pick-up stations to stretch its "last-mile" package delivery network and extend the delivery capacity. Moreover, pick-up stations offer consumers more time flexibility, privacy, and safety in receiving packages. From 2019 to 2022, the daily number of packages handled by pick-up stations showed approximately 100% year-on-year growth. In 2022, Alibaba already had nearly $100,000$ pick-up stations in China that handle over 50 million packages per day on average, reaching more than 100 million consumers in more than 200 cities.

The area of a pick-up station is between $20m^2$-$200m^2$. While establishing some of its own spots, the platform mostly enrolls pick-up stations by franchising. Pick-up stations could be set up either as independent stations or in existing businesses such as convenience stores, newsstands, or community offices. The franchisee, in return, gets paid by piece rate and drives store footfall. Figure 2.1(a) shows a typical pick-up station in Alibaba. In this figure, the worker is retrieving the packages from the shelf for consumers. When placing an order, consumers can choose to ship their items to nearby pick-up stations. In this way, instead of waiting for the delivery to arrive at home, consumers are free to collect packages that have arrived at the selected pick-up station anytime during operation hours. When they arrive at

Figure 2.1: Pick-up Station and Logistic Timeline

(a) In a Typical Pick-up Station



(b) Logistic Timeline

a pick-up station, consumers normally present the worker with the shipping code or a QR code to identify their packages. In the U.S., Amazon also adopted this business practice in recent years (*e.g.*, offering a pick-up option in Whole Foods Market).

Figure 2.1(b) shows a brief timeline of how pick-up stations function for online orders. When a consumer shops at a platform in Alibaba, such as Taobao and Tmall, before checking out, she can change the package delivery option to pick-up station delivery. After she checks out, her package is prepared by online stores and sent on its way to the distribution center. When the package leaves the distribution center, it will shipped either directly to the consumer's home, or to the consumer's selected pick-up station. A message will be sent to the consumer when the package arrives at the station. Then the the consumer can choose a time to pick up the package at her convenience.

2.3.2 Our Data

We conduct our empirical study based on the logistic records in Shenyang, the largest city in Northeast China[6]. The observation period runs from June 2019 to December 2019. We

[6]We also provide a robustness check on our reduced-form evidence based on data from another city in Appendix B.1.

combine three data sets to conduct our research. The first data set tracks 169,319,246 package-level logistic and transaction records from $7,584,151$ consumers, where $24,179,337$ (16.7%) records are from pick-up stations delivery and $145,139,909$ (83.3%) records are from home delivery. All consumers' transactions and corresponding logistic timelines were recorded during this period, including the time that packages were purchased (hereafter denoted as "checkout time"), attempted to deliver (hereafter denoted as "delivery time"; for station delivery this is the time packages were delivered to the station), and signed or received (hereafter denoted as "receiving time"). We also observe the item value for each package.

The second data set contains consumers' location information. For the purpose of protecting consumer privacy, the platform helps us to convert the locations of consumers into an equivalent discrete space according to the longitudes and latitudes of their addresses. Specifically, the city of Shenyang is divided into 1.3 million locations of discrete grids (0.1 km $\times$ 0.1 km), and each consumer is put into the corresponding grid based on her location. Among these locations, we observe consumers from $35,766$ locations during the observation period. The third data set includes the locations (longitudes and latitudes) and opening times of pick-up stations. We focus on the introduction of 55 pick-up stations opened in Shenyang within September 2019 and, in Appendix B.1, we conduct a robustness check on a sample involving pick-up stations opened during another period.

Table 2.1 summarizes the important variables in our data set at the package level for the full sample, the sub-sample of station delivery, and the sub-sample of home delivery. The checkout hour, delivery hour, and receiving hour are recorded in a 24-hour format, and are the corresponding hours of checkout time, delivery time, and receiving time. The order-to-deliver time is the time duration that a package is processed on its way to shipping, defined by the time difference between the delivery time and checkout time in days. The order-to-receive time is the actual waiting time for consumers to receive a package after placing the order,

Table 2.1: Summary Statistics of Package Attributes

	Full sample	Station delivery	Home delivery
Checkout hour (24-hour)	13.754	13.622	13.776
	(6.251)	(6.450)	(6.217)
Delivery hour (24-hour)	12.998	12.008	13.182
	(3.457)	(3.067)	(3.495)
Receiving hour (24-hour)	14.132	15.172	13.938
	(3.645)	(3.397)	(3.657)
Order-to-deliver time (days)	3.801	3.718	3.817
	(3.449)	(2.978)	(3.529)
Order-to-receive time (days)	4.008	4.469	3.922
	(3.531)	(3.238)	(3.576)
Items value (CNY)	111.766	72.720	118.271
	(495.098)	(215.053)	(527.216)
Observations	169,319,246	24,179,337	145,139,909

Note: Standard deviations are reported in parentheses. The mean receiving hour and the mean delivery hour for home delivery packages are not equal because home delivery orders can also have some deliver-pending such as multiple delivery attempts and shipping box delivery.

calculated by the time difference between the receiving time and checkout time in days. On average, an order takes 4.008 days to arrive in the hands of consumers and each order on average costs 111.80 CNY, equivalent to 16.69 USD at the time of writing. Interestingly, while the order-to-deliver time is longer for home delivery orders compared to station orders (3.817 days vs. 3.718 days), the actual time to receive the order, *i.e.*, order-to-receive time, is longer for station orders compared to home delivery orders (4.469 days vs 3.922 days). This signals that the consumers are willing to wait longer to take the station orders at specific times, which will be important later for our empirical analyses.

2.4 Empirical Evidence

In this section, we first analyze the impact of introducing pick-up stations and the potential mechanism, which later motivates our structural model and estimation in Section 2.5.

2.4.1 The Impact of Pick-up Stations

In order to estimate the impact of the pick-up stations, we aggregate our data at the location-day level from June 2019 to December 2019. Similar to our settings, [74] also compare

consumer purchases across different locations and study the aggregated impact of opening a retail store. However, in contrast with their settings where locations are divided by districts, we are able to refine locations into 0.1 km $\times$ 0.1 km grid units. Smaller location units are crucial in studying the impact of pick-up stations in our setting since the effective distance of a pick-up station is much shorter than a retail store—most consumers pick up packages from a station within their community.

Our empirical strategy is to compare consumers who are affected by an opening of a new station to those who are not to understand the impact of opening pick-up stations on consumer behaviors. Since the sales of e-commerce are largely influenced by promotions and seasonality, we use a DiD approach to rule out the influence of time trends between the pre-treatment and post-treatment periods. This strategy relies on the assumptions that (a) opening a pick-up station is an exogenous decision to consumers, which is confirmed by the platform and (b) consumers who are affected by the opening are parallel or comparable to those who are not, which we will explain later. According to the platform, 90% of consumers who choose to send purchases to pick-up stations in the city are within 0.4 km of this station. Therefore, we define 0.4 km as the treatment distance: The consumers who live within 0.4 km of a newly opened station are the affected consumers, and those who live outside this circle are consumers unaffected by this station[7].

Moreover, because we want consumers who are affected by a station to have parallel trends with those who are not, we would like to compare customers who are just inside and outside of the treatment distance since these consumers are more likely to have parallel trends due to their proximity. Therefore, when doing our analyses, we define our observation distance to be 0.4 km, which means we are comparing consumers who live within 0 to 0.4 km of a station

[7]If consumers who live outside the circle are still affected by the opening of the stations, our comparison results will only underestimate the true positive effects of the stations.

to those who are within 0.4 to 0.8 km. We confirm that the parallel trend assumption is satisfied using such treatment and observation distances in Appendix B.2. Later we conduct various robustness analyses to vary the treatment distance as well as the observation distance to demonstrate that our empirical results are not affected by the choices of these parameters.

In total, 6,655 grids are left in our observations. They span over 33 residential districts, and 2,102 grids are within the affected distance of stations opened in September. In this analysis, each observation is a location-day pair, representing location (grid) i on day t, and is either identified in the treatment or control group based on whether the distance between the location and the pick-up station closest to her is within a treatment distance of 0.40 km. Our main analysis covers an observation period of 180 days for each geographical location. These include 90 days of the pre-treatment period and 90 days of the post-treatment period, which are counted relative to the closest station's open date. In total, there are 1,197,900 location-day observations. The following DiD specification is used to test the effect of the pick-up stations:

$$Outcome\ Variable_{it} = \beta_1 After_t + \beta_2 Station_i + \beta_3 Station_i * After_t + \mu_i + \lambda_t + \varepsilon_{it},$$

where $Outcome\ Variable_{it} \in \{GMV_{it}, Items\ value_{it}, Orders_{it}, New\ comers_{it}\}$ represents the outcome consumer behaviors that we are interested in. GMV_{it}, $Items\ value_{it}$, $Orders_{it}$ and $New\ comers_{it}$ are the gross merchandise volume, *i.e.*, total sales, the average items value per order, number of orders and number of new customers at location i on day t. $After_t$ is a binary variable equaling one if the closest treatment station is already open on day t and zero otherwise, $Station_i$ is a binary variable equaling one if the closest station is within 0.40 km from the center of the grid and zero otherwise. μ_i and λ_t are area and time fixed effects. Time fixed effects include weekday fixed effects and year-week fixed effects. In area

Table 2.2: Impact of Pick-up Stations

	Dependent variable:			
	GMV	*Items value*	*Orders*	*New comers*
	(1)	(2)	(3)	(4)
After	211.416***	−1.491	1.780***	0.013
	(57.868)	(1.488)	(0.338)	(0.012)
Station	233.033***	−1.884**	2.471***	0.022***
	(28.023)	(0.719)	(0.164)	(0.006)
*Station*After*	106.989**	0.438	0.708**	−0.010
	(39.441)	(1.012)	(0.230)	(0.008)
Relative Effect Size	3.9%	0.4%	2.9%	-5.4%
Time fixed effects	Yes	Yes	Yes	Yes
Area fixed effects	Yes	Yes	Yes	Yes
Observations	1,197,900	957,018	1,197,900	1,197,900

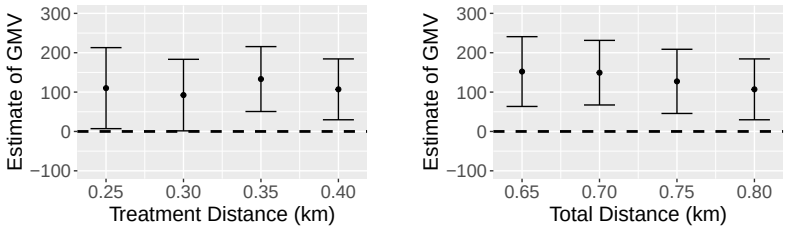
Note: *p<0.05; **p<0.01; ***p<0.001.
Note: Average of *GMV* was 2774.850, average of *Items value* was 109.744, average of *Orders* was 24.108, average of *New comers* was 0.185.

fixed effects, we control for the fixed effect of each residential district, which is the level of residence and logistic distributions of the platform.

The coefficient β_3 captures the impact of a pick-up station on those locations that are close to a station, compared with other locations. Table 2.2 shows the results of these regressions. Column 1 in Table 2.2 demonstrates that opening a station can increase the total sales of a grid in a day by 106.99 CNY (15.81 USD), which represents a 3.9% sales lift. Moreover, Column 2 and Column 3 in Table 2.2 show that the sales gain comes mainly from the increasing number of orders rather than the average order size. Last but not least, Column 4 demonstrates that the number of new customers does not change significantly after the introduction of the stations, which means the sales lift is predominantly driven by existing customers' behavior changes.

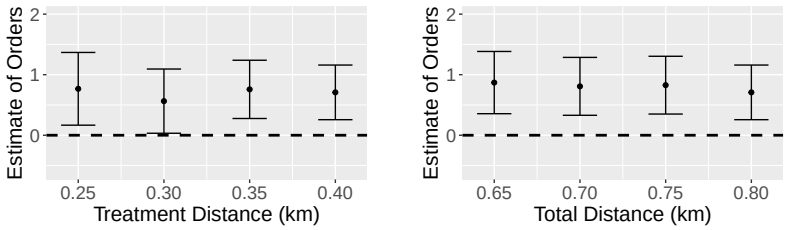
Furthermore, we show that the impact of pick-up stations is qualitatively the same across different choices of treatment and observation distances. Figure 2.2(a) shows that the impact of opening a store on GMV is robust if we change the treatment distances from 0.4 km to 0.35 km, 0.3 km, and 0.25 km. Figure 2.2(b) shows that if we change the observation

Figure 2.2: Robustness of Estimates on GMV (error bar: 95% confidence interval)



(a) Treatment Distance (b) Total Observation Distance

Figure 2.3: Robustness of Estimates on Number of Orders (error bar: 95% confidence interval)



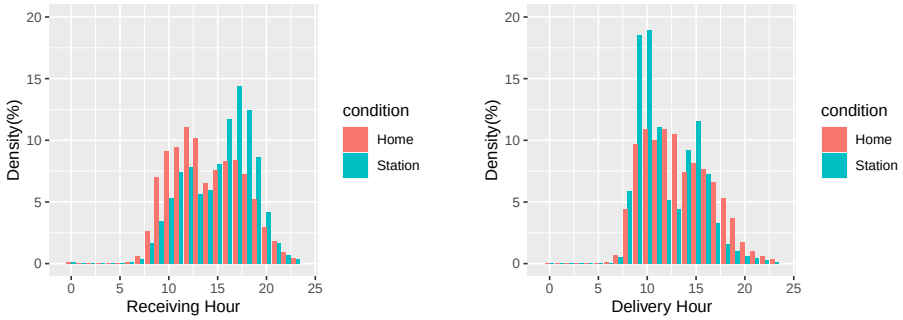
(a) Treatment Distance (b) Total Observation Distance

distance from 0.8 km to 0.75 km, 0.7 km, and 0.65 km, the results on GMV remain consistent. Moreover, Figures 2.3(a) and 2.3(b) confirm that the estimates on the number of orders also remain qualitatively similar when we vary the treatment and observation distances.

2.4.2 Possible Mechanism

Although past empirical work in e-commerce logistics documents higher shipping speed is an important driver of platform profits (*e.g.*, [53]), we show that the speed effect is not the driving factor on the impact of pick-up stations in our setting. From Table 2.1, the average time-to-receive an order in station delivery is 8.1% longer than that in home delivery, and the difference is statistically significant ($p < 0.0001$). This means that the sales lift after opening a pick-up station cannot be accredited to the shipping speed reduction. Moreover, Table 2.1 also shows that the average delivery time to the station is 9.3% shorter than the delivery time to the home, and such difference is statistically significant ($p < 0.0001$). Combing these two

Figure 2.4: Distribution of Receiving Hour and Delivery Hour



(a) Receiving Hour (b) Delivery Hour

pieces of empirical evidence, we can see that consumers receive their orders later after using pick-up stations because they strategically wait longer to pick them up. This motivates us to hypothesize that pickup time flexibility is the driving force behind the impact of stations.

To further validate our hypothesis, we compare logistics time records for station delivery orders and home delivery orders in Figure 2.4(a)–2.4(b). Figure 2.4(a) depicts the distribution of package receive hours, *i.e.*, the time when consumers sign the packages at home or the time when they pick them at the station. Both distributions for station pickups and home deliveries are bimodal but their highest peaks differ. The highest peak for station orders occurs from 4 p.m. to 7 p.m., which is commonly perceived as after-work hours. While the two peaks are more similar for home delivery orders, the slightly higher peak is between 10 a.m. to 2 p.m., which is considered as working hours. While Figure 2.4(a) shows that the peak of delivery pick-up time is during after-work hours, it cannot refute the possibility that consumers can only pick up packages during after-work hours because these packages are delivered to the stations during this time frame. Figure 2.4(b) rules out this possibility. It shows the distribution of package delivery time, and it can be seen that most station orders were delivered in the morning, *i.e.*, during working hours. Combining these two figures, we can again infer that although orders mostly arrive at pick-up stations in the morning,

consumers typically wait until after-work hours to pick up their packages. This suggests that they may have different time preferences than the assigned delivery time in home delivery. In the next section, we will propose a structural model that explicitly accounts for such time preferences and estimate consumers' time preferences from their choices between home and station deliveries in our data.

2.5 Structural Model and Results

To study how logistic flexibility explicitly affects consumer purchases, in this section we present a structural model considering consumers' decisions about purchasing and delivery options, demonstrate how we estimate the model, and provide estimation results.

2.5.1 Consumer Choice Model

In our model, a consumer has two decision stages. In the first stage, the consumer chooses among home delivery, pick-up station delivery, and the outside option, which may reflect the utility of purchasing on other platforms or offline. If a consumer chooses the home delivery or the outside option, she does not need to make a decision in the second stage and the delivery time and the final utility will be realized in the second stage. However, if station delivery is chosen, after the package is delivered to the station, the consumer will receive a message and choose a time to pick up her package in the second stage. The time that a package arrives at the station determines the consumer's second-stage choice set of pick-up time slots.

Given the complexity of station delivery, let us focus on the second-stage decision in station delivery first. Specifically, consider a potential order i in the second stage. The consumer j would optimally choose her pick-up time from the feasible pickup time choice set. We denote the time a package shipped to the pick-up station as $T_{ij}^S = \delta$ ($\delta \in \mathbb{O}$), where $\mathbb{O}$

is the set of operating hours in a day. We define x as the hour of a day the consumer chooses to pick up from the station, where $x \in \mathbb{V}(\delta)$, and $\mathbb{V}(\delta)$ is the feasible pickup time choice set (*i.e.*, x must be later than δ). Since in our data set, nearly all packages were picked up within three days of the delivery time, for the reduction of estimation dimensionality, we focus on packages that were picked up within three days during the operation hours from 8:00 a.m. to 21:59 p.m., this includes 95.4% of all packages in station deliveries. We denote 0:00-0:59 a.m. in the ship-to-station day as hour 0. As discussed in Section 2.5.2, the set of operation hours is $\mathbb{O} = \{8, 9, ..., 21\}$. The pick-up choice set is $\mathbb{V} = \{8, 9, ..., 21\} \cup \{32, 33, ..., 45\} \cup \{56, 57, ..., 69\}$. Therefore, for a package shipped to the pick-up station at hour $T_{ij}^S = \delta$ ($\delta \in \mathbb{O}$), the consumer can choose a pickup hour, x, in the subsequent operation hours within the choice set. This means that the feasible pickup choice set is $\mathbb{V}(\delta) = \{\delta, \delta + 1, ..., 21\} \cup \{32, 33, ..., 45\} \cup \{56, 57, ..., 69\}$. For example, if the order arrives at 11 a.m. on Monday, the feasible choice set for this consumer will be from 11 a.m. to 21 p.m. on Monday ($\{11, 12, ..., 21\}$), 8 a.m. to 21 p.m. on Tuesday ($\{32, 33, ..., 45\}$), and 8 a.m. to 21 p.m. on Wednesday ($\{56, 57, ..., 69\}$).

Let $L_{ij}^S = l$ ($l \in \mathbb{L}$) be the in-transit period (in hours) that package i takes from the time the consumer made the order to 0 a.m. of the day when the package was out for delivery, where $\mathbb{L}$ is the feasible set of the in-transit period. d_j denotes the consumer's distance to the closest pick-up station. β_{jx}, γ_j, and η_j represent the individual specific utility of time preference at hour x, sensitivity on shipping time, and sensitivity on pickup travel distance, respectively. θ_j^S is the constant utility term for consumer j in choosing a pick-up station compared to home delivery. The stochastic component ε_{ijx} represents the individual idiosyncratic time preferences over pickup hours and is realized before consumers make decisions in the second stage. We assume that ε_{ijx} follows a Gumbel distribution $\varepsilon_{ijx} \sim Gumbel(0, \sigma_j)$ and is independent and identically distributed over pickup hours x. Since ε_{ijx} is unknown in the

first stage, we call the scale parameter σ_j as the uncertainty parameter, where a larger σ_j represents a larger uncertainty of her time preference for hour x at the time of first-stage decision-making. With these notations, if order i is shipped to a pick-up station with in-transit time l, and at the second stage consumer j picks up at hour x, her utility is then

$$U_{ij}^S(\varepsilon_{ijx}, x, \delta, l) = \beta_{jx} + \gamma_j(l + x) + \eta_j d_j + \theta_j^S + \varepsilon_{ijx},\ x \in \mathbb{V}(\delta).$$

Given consumers will maximize their utility in choosing the pick-up time, the utility in the second stage given the feasible pickup time choice set is:

$$\tilde{U}_{ij}^S(\boldsymbol{\varepsilon}_{ij}, \delta, l) = \max_{x \in \mathbb{V}(\delta)} U_{ij}^S(\varepsilon_{ijx}, x, \delta, l) = \max_{x \in \mathbb{V}(\delta)} \beta_{jx} + \gamma_j(l + x) + \eta_j d_j + \theta_j^S + \varepsilon_{ijx}.$$

In the first stage, the consumer makes the purchase and delivery decisions. There is uncertainty regarding when the package will arrive at the pick-up station (*i.e.*, δ) as well as how her time preferences (*i.e.*, ε_{ijx}'s) will evolve. The consumer would form an expectation on her utility for choosing the pick-up delivery option. Let $Pr(T_{ij}^S = \delta)$ denote the probability of the package arriving at the pick-up station at hour $T_{ij}^S = \delta$, $Pr(L_{ij}^S = l)$ denote the probability for the in-transit time to be $L_{ij}^S = l$. Therefore, facing the uncertainty of package arrival time to the station, package in-transit period, and idiosyncratic time preferences ($\boldsymbol{\varepsilon}_{ij}$), when the consumer makes the delivery option decision in the first stage, her expected utility in choosing the pick-up station is:

$$v_{ij}^S = \sum_{l \in \mathbb{L}} \sum_{\delta \in \mathbb{O}} \int \tilde{U}_{ij}^S(\boldsymbol{\varepsilon}_{ij}, \delta, l) d\boldsymbol{\varepsilon}_{ij} Pr(T_{ij}^S = \delta) Pr(L_{ij}^S = l).$$

In the first stage, we assume that there is an idiosyncratic utility shock, e_{ij}^S, where $e_{ij}^S \sim Gumbel(0,1)$, that will affect the preference for the option. For example, there may be an advertising effort from Alibaba's platform to encourage consumers to choose the pick-up station option which researchers do not observe. For model identification, we normalize the scale parameter of e_{ij}^S to 1. Therefore, consumer j's utility on order i conditional on choosing the pick-up station is:

$$u_{ij}^S(e_{ij}^S) = v_{ij}^S + e_{ij}^S. \tag{2.1}$$

The following proposition provides a closed-form representation for the expected utility v_{ij}^S that is later used in estimation:

At the first stage, if station delivery is chosen, consumer j has a utility upon purchasing order i:

$$u_{ij}^S(e_{ij}^S) = v_{ij}^S + e_{ij}^S, \tag{2.2}$$

where the first term is

$$v_{ij}^S = \sigma_j \sum_{\delta \in \mathbb{O}} Pr(T_{ij}^S = \delta) ln[\sum_{k \in \mathbb{V}(\delta)} exp(\frac{\beta_{jk} + \gamma_j k}{\sigma_j})] + \sigma_j C + \gamma_j \sum_{l \in \mathbb{L}} l Pr(L_{ij}^S = l) + \eta_j d_j + \theta_j^S,$$

and C is the Euler constant.

Proof.

In deriving $v_{ij}^S = \sum_{l \in \mathbb{L}} \sum_{\delta \in \mathbb{O}} \int \tilde{U}_{ij}^S(\boldsymbol{\varepsilon}_{ij}, \delta, l) d\boldsymbol{\varepsilon}_{ij} Pr(T_{ij}^S = \delta) Pr(L_{ij}^S = l)$, we rewrite $\int \tilde{U}_{ij}^S(\boldsymbol{\varepsilon}_{ij}, \delta, l) d\boldsymbol{\varepsilon}_{ij}$ as $\int [\max_{x \in \mathbb{V}(\delta)} (\beta_{jx} + \gamma_j x + \varepsilon_{ijx})] d\boldsymbol{\varepsilon}_{ij} + \gamma_j l + \eta_j d_j + \theta_j^S$. Let $Y = \max_{x \in \mathbb{V}(\delta)} (\beta_{jx} + \gamma_j x + \varepsilon_{ijx})$. We first show that Y follows a Gumbel distribution:

Since $Pr(Y \leq y) = \prod_{k \in \mathbb{V}(\delta)} Pr(\beta_{jx} + \gamma_j x + \varepsilon_{ijx} \leq y)$, we have

$$lnPr(Y \leq y) = \sum_{k \in \mathbb{V}(\delta)} lnPr(\varepsilon_{ijx} \leq y - \beta_{jx} - \gamma_j x) = - \sum_{k \in \mathbb{V}(\delta)} exp[-(y - \beta_{jx} - \gamma_j x)/\sigma]$$
$$= -exp(-y/\sigma + ln \sum_{k \in \mathbb{V}(\delta)} exp[(\beta_{jx} + \gamma_j x)/\sigma_j]) = lnPr(Z \leq y'),$$

where $y' = y/\sigma_j$. Since $Pr(Z \leq y')$ is the cumulative density function of Gumbel distribution with location parameter $ln \sum_{k \in \mathbb{V}(\delta)} exp[(\beta_{jx} + \gamma_j x)/\sigma_j]$ and scale parameter 1, $Y = \sigma_j Z$ follows a Gumbel distribution with location parameter $\sigma_j ln \sum_{k \in \mathbb{V}(\delta)} exp[(\beta_{jx} + \gamma_j x)/\sigma_j]$ and scale parameter σ_j.

Therefore $\int [\max_{x \in \mathbb{V}(\delta)} (\beta_{jx} + \gamma_j x + \varepsilon_{ijx})] d\boldsymbol{\varepsilon}_{ij} = E[Y] = \sigma_j ln \sum_{k \in \mathbb{V}(\delta)} exp[(\beta_{jx} + \gamma_j x)/\sigma_j] + \sigma_j C$, where C is the Euler constant. Then we have $v_{ij}^S = \sigma_j \sum_{\delta \in \mathbb{O}} Pr(T_{ij}^S = \delta) ln[\sum_{k \in \mathbb{V}(\delta)} exp(\frac{\beta_{jk} + \gamma_j k}{\sigma_j})] + \sigma_j C + \gamma_j l + \eta_j d_j + \theta_j^S$.
□

If order i is directly shipped to home and arrived at hour $T^H = \delta$ ($\delta \in \mathbb{O}$), consumer j's utility is defined as

$$U_{ij}^H(\varepsilon_{ij\delta}, \delta, l) = \beta_{j\delta} + \gamma_j(l + \delta) + \theta_j^H + \varepsilon_{ij\delta},$$

where $L_{ij}^H = l$ ($l \in L^S$) is the in-transit period (in hours) that package i takes from the time that the consumer made the order to 0 a.m. of the date the package was out for delivery, and θ_j^H is the difference in the utility of choosing the home delivery option over the pick-up station option (after controlling for the time preference, delivery time, and traveling distance differences). We allow θ_j^H to exist because, as home delivery was the only delivery option in the past, some consumers may choose this option out of habit or inertia in making choices.

Notice that consumers share the same set of idiosyncratic time preference shocks, *i.e.*, $\boldsymbol{\varepsilon}_{ij}$, regardless of whether she chooses home or pick-up station deliveries. This makes sense empirically since consumers' time preferences during the pickup day should be correlated with their other activities instead of their mode of delivery. This means the scale parameter

for the idiosyncratic shocks in home and station delivery options is the same. Similarly, at the first stage, we denote the idiosyncratic utility shock for choosing the home delivery for consumer j and order i as $e_{ij}^H \sim Gumbel(0,1)$. The following lemma provides the closed-form utility for consumer j on order i conditional on choosing home delivery: At the first stage, if home delivery is chosen, consumer j has a utility upon purchasing order i:

$$u_{ij}^H(e_{ij}^H) = \sum_{l\in\mathbb{L}^{\mathbb{H}}}\sum_{\delta\in\mathbb{O}}\int U_{ij}^H(\varepsilon_{ij\delta},\delta,l)d\varepsilon_{ij\delta}Pr(T^H=\delta)Pr(L^H=l) + e_{ij}^H = v_{ij}^H + e_{ij}^H, \tag{2.3}$$

where the term

$$v_{ij}^H = \sum_{\delta\in\mathbb{O}}(\beta_{j\delta}+\gamma_j\delta)Pr(T^H=\delta) + \sigma_j C + \gamma_j\sum_{l\in\mathbb{L}^{\mathbb{H}}} lPr(L^H=l) + \theta_j^H.$$

Finally, we normalize the expected utility of the outside option to 0. If the consumer chooses the outside option in the first stage, her utility will be represented by an idiosyncratic shock $e_{ij}^0 \sim Gumbel(0,1)$:

$$u_{ij}^0(e_{ij}^0) = e_{ij}^0. \tag{2.4}$$

Combining the above utilities, in the first stage, a consumer decides between purchasing on the platform with station delivery or home delivery, and the outside options to purchase on other platforms. The consumer solves the maximization problem:

$$\hat{u}_{ij} = \max\{u_{ij}^S, u_{ij}^H, u_{ij}^0\}.$$

Let c_{ij} be the choice the consumer j makes on order i, which equals 0 if it is an outside option, 1 if it is a station delivery, and 2 if it is a home delivery. Let $\mathbf{\Theta}_j$ be the vector of all model

parameters, $P_i(\mathbf{\Theta}_j)$ denote the likelihood for the consumer to make the choice. Especially, the likelihood to choose station delivery and pick up at hour $x = k$ is a joint probability $P_i(c_{ij} = 1, x = k) = P_i(c_{ij} = 1)P(x = k|c_{ij} = 1)$, where $P_i(c_{ij} = 1) = \frac{exp(v_{ij}^S)}{1+exp(v_{ij}^H)+exp(v_{ij}^S)}$ and $P(x = k|c_{ij} = 1) = \frac{exp\{[\beta_k+\gamma(t_{ij}^S+k)+\eta_j d_j+\theta_j^S]/\sigma_j\}}{\sum_{m\in\mathbb{V}(\delta)} exp\{[\beta_m+\gamma(t_{ij}^S+m)+\eta_j d_j+\theta_j^S]/\sigma_j\}} = \frac{exp[(\beta_k+\gamma k)/\sigma_j]}{\sum_{m\in\mathbb{V}(\delta)} exp[(\beta_m+\gamma m)/\sigma_j]}$.

Therefore, given the idiosyncratic utility shock $e_{ij}^S, e_{ij}^H, e_{ij}^0 \sim Gumbel(0, 1)$, the probability of choice for each potential order i is then

$$P_i(c_{ij}, \mathbf{\Theta}_j) = \begin{cases} \frac{exp(v_{ij}^S)}{1+exp(v_{ij}^H)+exp(v_{ij}^S)} \frac{exp[(\beta_k+\gamma k)/\sigma_j]}{\sum_{m\in\mathbb{V}(\delta)} exp[(\beta_m+\gamma m)/\sigma_j]}, & \text{if } c_{ij} = 1, \text{ and } x = k,\ T_{ij}^S = \delta; \\ \frac{exp(v_{ij}^H)}{1+exp(v_{ij}^H)+exp(v_{ij}^S)}, & \text{if } c_{ij} = 2; \\ \frac{1}{1+exp(v_{ij}^H)+exp(v_{ij}^S)}, & \text{if } c_{ij} = 0. \end{cases} \tag{2.5}$$

We allow consumers to be heterogeneous in their preferences using latent class estimation. We assume there are multiple latent types of consumers. Let $Pr(Type = k)$ be the probability of consumers in type k, $k = 1, 2, ..., K$, the log-likelihood function for K types of consumers is

$$L = \sum_i log[\sum_k Pr(Type = k)P_i(c_{ik}, \mathbf{\Theta}_k)]. \tag{2.6}$$

We estimate $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, ... $\mathbf{\Theta}_K$ by maximizing the log likelihood function in Equation (2.6).

2.5.2 Model Estimation and Identification

In our structural estimation, observations are at the order level. We use the logistic records of $5,486,243$ orders from $893,126$ consumers within 1.00 km distance to the 55 pick-up stations in the next month after the stations were opened (*i.e.*, October 2019). Since we do not observe outside options in the data, we need a proxy for market potential; similar to the past

literature (see [94]), we assume consumers consider products offered by the platform each day. In other words, on a calendar date, a consumer either purchases on the platform or chooses the outside option to purchase elsewhere, online or offline[8]. To estimate the time preferences for different periods, we model the logistic timeline in discrete hours. Consumers form rational expectations on the probabilities of delivery hours of station delivery and home delivery. We derived the delivery probabilities based on the empirical distribution (see Figure 2.5(b)) of delivery hours in the city in our whole observation scope.

Our model can be essentially reduced to a two-stage multinomial choice model; following the classic literature on the multinomial choice model [95], the parameters are identifiable in our model. Specifically, the time preference is identified by the choices of consumers in their pick-up hours. The utility parameters related to station versus home delivery are identified by consumers' first-stage decisions in choosing between station and home delivery when the station becomes available in our data. Station distance sensitivity is identified by consumers' first-stage decisions because consumers with different traveling distances to the station would be affected differently in their decisions between station and home delivery choices. Moreover, package waiting sensitivity is identified by the choices of consumers in both their pick-up hours considering the realized waiting length in the second stage when choosing station delivery, and their first stage decisions between station and home delivery choices considering the expected waiting length in station delivery and home delivery. Last but not least, the scale parameter of the second-stage time idiosyncratic shock is identified by, at the individual level, the variance of the time an individual picks up the order from the station.

Last but not least, we will briefly introduce our numerical method choices in estimating our likelihood function. We use the gradient-free Nelder-Mead method to optimize the

[8]For the consumers who have one or more orders in a day, we consider it as a purchase decision in the day and save the first logistic timeline in model estimation.

Table 2.3: Estimation Results

	Low type	Medium type	High type
Daytime receiving value	2.291	0.618	10.218
	(0.001)	(0.001)	(0.003)
After-work receiving value	4.916	11.536	22.122
	(0.003)	(0.002)	(0.003)
Scale parameter (uncertainty)	5.361	14.802	16.642
	(0.001)	(<0.001)	(<0.001)
Package waiting sensitivity	−0.308	−0.939	−1.032
	(<0.001)	(<0.001)	(<0.001)
Station distance sensitivity	−2.596	−0.991	−0.106
	(0.009)	(0.002)	(0.001)
Home constant	−2.422	−1.968	−10.254
	(0.001)	(0.001)	(0.001)
Station constant	−18.996	−40.914	−51.081
	(0.005)	(0.001)	(0.002)
Type probability	0.184	0.711	0.105
	(<0.001)	–	(<0.001)
Observations		893,126	

Note: Standard errors are presented in parentheses. Daytime receiving is between 10:00 a.m.-3:59 p.m., after-work receiving is between 4:00 p.m.-7:59 p.m..

log-likelihood function (2.6). We bootstrap the standard errors by random drawing the sample at the consumer level 100 times, and repeatedly estimating the model[9]. We use BIC criteria to select the optimal number of latent classes. With the number of latent classes increasing from one to two, three, and four, the BIC would reduce by 4.3%, 4.4%, and 0.7%, respectively. Since with the number of classes increases from three to four, the marginal BIC reduction is smaller than 1%, and the additional class only takes a small segment of 5.5%, so we choose to have only three latent classes of consumers[10].

2.5.3 Results

Table 2.3 reports the estimates of the consumer choice model. Our latent class model shows that there are three types of consumers with different preferences and they account for 18.4%,

[9]We do not use the closed-form asymptotic distributions to derive standard deviations since we find the calculated Hessian matrix imprecise depending on the absolute error we set in function convergence. Therefore, we use bootstrapping which has been shown in the literature to be more robust.

[10]We report the estimation results from one to four latent classes in Table B.3 Appendix B.3, which demonstrate that using four latent classes will generate qualitatively similar results compared to using three latent classes.

71.1%, and 10.5% of the total consumers respectively (Row 8). The estimates of daytime (*i.e.*, 10 a.m.-4 p.m.) receiving values for these three types of consumers are 2.291, 0.618, and 10.218, respectively (Row 1). The estimates of after-work (*i.e.*, 4 p.m.-8 p.m.) receiving values are 4.916, 11.536, and 22.122, respectively (Row 2). For each type of consumer, the after-work receiving value is statistically significantly greater than the daytime receiving value. This result shows that consumers would prefer to receive/pick up their packages at the end of a day rather than in the middle of a day, since most people have more flexibility to handle packages during the after-work period. This is also consistent with our reduced-form evidence from Figure 2.4(b) that consumers wait strategically until after-work hours to pick up their packages. Note that the difference between daytime receiving value and after-work receiving value is an indicator of the consumer's value on flexibility of being able to choose a preferred pickup time before the last-minute idiosyncratic shock is realized, *i.e.*, the time flexibility. We sort three types of consumers in ascending order by how much they value time flexibility, and thereafter, we label them as low type, medium type, and high type, respectively.

The scale parameter (uncertainty) captures the consumer's variation of receiving value in each hour (Row 3). For low-type consumers, the uncertainty parameter is estimated to be 5.361 (Column 1). The uncertainty parameter is much larger in medium-type consumers (14.802, Column 2), and even slightly larger in high-type consumers (16.642, Column 3). These results imply that, when consumers make the choice in the first stage, there is a larger uncertainty about their actual preferences for when to receive their packages for the home delivery option. For the pick-up station option, however, consumers can choose the pick-up time after the idiosyncratic shocks are realized. We label such benefits as choice flexibility. Our estimation results show that the last two types of consumers place a higher value on choice flexibility. Since they also place a higher value on time flexibility, pick-up stations bring these two types higher value than the low type.

All consumers have negative waiting sensitivity ranging from -0.308 to -1.032 per hour (Row 4). Comparing this parameter to the time flexibility, for low-type consumers, the time flexibility is worth around $\frac{4.916-2.291}{0.308} = 8.523$ hours of waiting. In other words, a low-type consumer is willing to wait another 8.523 hours to receive her package if she can get it during after-work hours compared to daytime hours. Such cutoffs for medium-type and high-type consumers are 11.627 hours and 11.535 hours, respectively. Moreover, in general, consumers experience disutility if they need to travel longer to the pick-up station: The low-type, medium-type, and high-type consumers experience -2.596, -0.991, and -0.106 (Row 5) disutility, respectively, if they travel one more km to the pick-up station. A simple comparison shows that each kilometer of traveling distance is worth 0.103 to 8.429 hours of waiting depending on the type of consumers. Last but not least, we see that all consumers have higher utility for home delivery compared to station delivery holding everything else equal. Specifically, low-type, medium-type and high-type consumers have 16.574, 38.946, and 40.827 utility differentials between home delivery and station delivery (calculated by the difference between Row 6 and Row 7). Since we have controlled for the traveling distance in the pick-up option, these differences do not reflect such costs. Rather, they may indicate that consumers have inertia in switching options (since many are used to home delivery), or there are other inconveniences related to this option (*e.g.*, they must leave home for the package and wait in the station during peak hours). Besides, these differences also include the difference in expected waiting disutilities for the in-transit period between home delivery and station delivery.

2.6 Counterfactual Policies

One reason we estimate the structural model is to use counterfactuals to explore better logistic solutions for the platform. In this section, we will use our estimated structural model

to perform several counterfactual studies. We will first present a counterfactual study to break down the value of pick-up stations into the value of time flexibility and the value of choice flexibility. We will then develop a strategy for selecting new pick-up station locations based on our estimates. Last, we will also demonstrate how to rearrange the shipping time across different types of consumers to improve sales based on our estimated logistic flexibility.

2.6.1 Time Flexibility and Choice Flexibility

We first use a counterfactual study to quantify the effect of pickup time flexibility and choice flexibility in driving total sales. Using our estimated structural model, we first simulate the sales before and after the introduction of pick-up stations to quantify the impact of pick-up stations in driving sales. We then repeat the same exercise but restrict the average time preference for daytime and after-work hours to be the same and equal to the weighted average time preference, which then restricts the value of the time flexibility to zero in the post-station simulation. This shows us the impact of pick-up stations in driving sales through only choice flexibility. We then compare these two simulation results to break down the impact of pick-up stations on sales into the impact of time flexibility and choice flexibility.

Specifically, in the first simulation, we randomly draw consumers' uncertainty in Stage 1 based on Equations (2.2) and (2.3). For each consumer, we first determine the consumer's type based on the estimated type probability in Table 2.3. We then simulate the consumer's decisions based on her type and corresponding preference parameters in Table 2.3 for both prior-station and post-station scenarios. In the prior-station scenario, consumers are only allowed to choose between home delivery and the outside option since the station is not available to them. In the post-station scenario, consumers are allowed to choose between not only home delivery and the outside option, but also station delivery. In our second simulation, we repeat the same procedure except for one change–we assume that consumers have the

Table 2.4: Simulations from the Structural Model

	Simulation 1: overall performance				Simulation 2: value of flexibility			
	All	Low type	Medium type	High type	All	Low type	Medium type	High type
Panel A: Prior-station and post-station purchase rate								
Prior-station purchase rate	0.115	0.315	0.067	0.085	0.115	0.315	0.067	0.085
Post-station purchase rate	0.136	0.318	0.073	0.242	0.120	0.317	0.068	0.126
Purchase rate increase	0.021	0.002	0.006	0.157	0.005	0.002	0.001	0.041
Panel B: Decomposition of post-station purchase rate								
Post-station purchase rate with station delivery	0.023	0.003	0.006	0.171	0.006	0.003	0.001	0.045
Post-station purchase rate with home delivery	0.113	0.314	0.067	0.070	0.114	0.315	0.067	0.081

same time preference for daytime and after-work hours, which equals the weighted average of those two parameters in Table 2.3 for each corresponding type.

Table 2.4 reports the simulation results. From Panel A, the purchase probability from the Alibaba platform prior to the introduction of the station is 11.5% and this rate has increased by 2.1 percentage points (*i.e.*, a 18.3% increase) to 13.6% (Column 1) after the pick-up station is introduced. Moreover, if we force the time preference to be the same across daytime and after-work hours after the station is introduced, and equal to the expectation of time preference given the distribution of delivery hours, the purchase probability for an average consumer is 12.0%. Since prior to the introduction of the station, the purchase probability is 11.5%, the choice flexibility on average increases the purchase probability by $12.0 - 11.5 = 0.5$ percentage points (Column 5, *i.e.*, a 4.3% increase), while the time flexibility increases the purchase probability by $2.1 - 0.5 = 1.6$ percentage points. In other words, $\frac{1.6}{2.1} = 76.2\%$ of the purchase rate increase is from the value of pickup time flexibility, and the rest $1 - 76.2\% = 23.8\%$ is from the value of choice flexibility[11]. As shown in Panel B, when the value of pickup time flexibility is zero, the purchase rate with the station delivery option would be largely reduced from 2.3% to 0.6% (Column 1 and Column 5, *i.e.*, a 73.9% decrease).

[11]Note that choice and time flexibility may interact in the utility function. Here we treat the impact of choice flexibility as the impact of knowing the realization of idiosyncratic shocks prior to making the decision and everything else as the time flexibility.

Similarly, from the prior-station and post-station purchase probabilities for each type of consumer in Columns 2-4 and Columns 6-8 of Panel A, we derive that the value of logistic flexibility decomposition differs across different types of consumers: pickup time flexibility accounts for 33.3%, 83.3% and 73.9% of total sales lift for low-type, medium-type, and high-type consumers respectively. That is, in medium-type and high-type consumers, the logistic flexibility is mainly driven by pickup time flexibility. In low-type consumers, since the difference between daytime receiving value and after-work receiving value is much smaller, the logistic flexibility is mainly driven by the flexibility to delay the pick-up time decisions after packages arrive, or choice flexibility. We also find that the purchase rate increase after the pick-up station is introduced mainly comes from high-type consumers since, for this type of consumer, the absolute value of logistic flexibility is the highest.

2.6.2 Pick-up Station Location Strategy

Since consumers need to pick up a package in person for station deliveries, those who are closer to pick-up stations will benefit more. However, Alibaba mostly enrolls pick-up stations by franchising without giving much consideration to locations. With the consideration of consumer-type distribution and consumer utilities of logistic flexibility, Alibaba should be able to better serve consumers by placing pick-up stations near those who are more sensitive to the opening of these stations (*i.e.*, medium-type and high-type consumers). With this idea in mind, we use the estimated structural model to run a counterfactual logistics strategy for selecting better locations to set up new pick-up stations among potential locations in the whole city. Figure 2.5(a) shows the citywide population density map, as well as the locations of the 55 pick-up stations in the city (in blue circles). On the density plot, each bin represents the associated consumer population, and a darker bin suggests a higher population (see the color bar on the right side)

To run the counterfactual analysis of pick-up station locations, we need to know each customer's type in the entire city in addition to the type of each customer in our estimated sample. Therefore, we first predict each consumer's type from their purchasing records within the estimation sample based on a Bayesian method. Specifically, let $\mathbb{U}_j$ be the set of order decisions that consumer j chooses. If the consumer j is in type k, the probability of her logistic choices can be written as

$$Pr_j(\mathbf{\Theta}_k) = \prod_{i \in \mathbb{U}_j} Pr_i(c_{ij}, \mathbf{\Theta}_k).$$

Therefore, based on her purchase history $\mathbb{U}_j$, the posterior probability for consumer j to be in type k can be written as:

$$Pr_j(Type = k) = \frac{Pr_j(\mathbf{\Theta}_k)}{\sum_{m=1}^{3} Pr_j(\mathbf{\Theta}_m)},\ k = 1, 2, 3.$$

For each consumer in our estimation sample, we assign her to be low-type, medium-type, or high-type by comparing and selecting the type associated with the largest posterior probability. From our prediction results, 19.5% of consumers are predicted as low type, 67.7% are predicted as medium type, and 12.9% are predicted as high type, which is consistent with our estimation in Table 3.

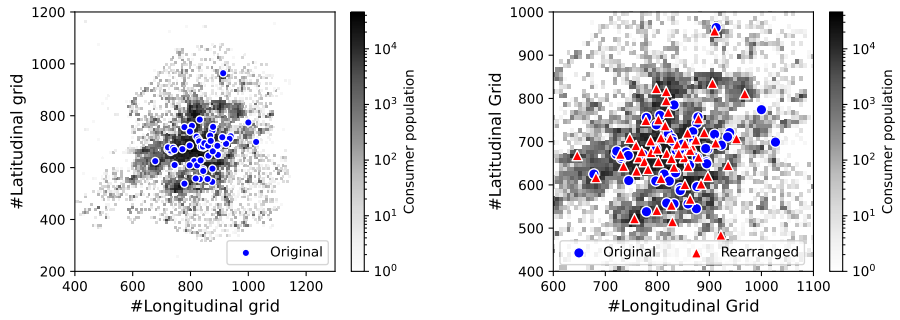
Moreover, we also need to determine the type for all consumers in the city of Shenyang who are not in our estimation sample. Since we don't observe logistic choices for consumers outside our estimation sample, we are not able to estimate consumer type using the above method. Therefore, we use the following logistic regression to extrapolate consumer type from the estimation sample for the whole city given consumers' previous aggregated purchasing

information from Alibaba:

$$Pr_i(Type = k) = \frac{exp(\beta_{\mathbf{k}}\mathbf{Z_i})}{1 + \sum_m exp(\beta_{\mathbf{m}}\mathbf{Z_i})},$$

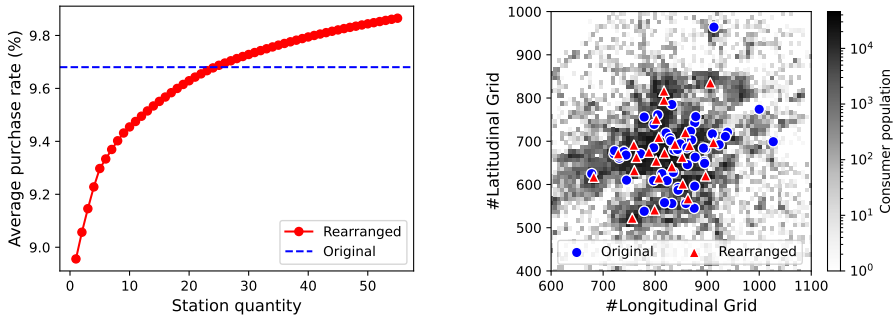
where $\mathbf{Z_i}$ is the vector of demographics including the consumer's average total value in a package, the average value per item in a package, the average value for the most expensive item in a package, and the average total monthly purchases. The average values are measured during the observation period from June 2019 to December 2019. Using the logistic regression, we extrapolate consumer type for the entire city including 47,920 locations and 7,141,676 consumers. After the extrapolation, 5.2% consumers are assigned to the low type, 85.7% are assigned to the medium type, and 9.2% are assigned to the high type.

Figure 2.5: Picking Better Locations of Pick-up Stations (length/height of one grid: 0.1 km)



(a) Map: Population and Stations in the City

(b) Map: Improve Value of Logistic Flexibility



(c) Purchase Rate and Picked Locations

(d) Map: Run Fewer Locations

With each consumer's predicted type and corresponding preference parameter, we then use a greedy algorithm to rearrange the locations of these newly introduced 55 pick-up stations over potential locations in the whole city to maximize the overall consumer purchase probability after the introduction of stations. We fix the locations of all other stations and only move the locations of the 55 new stations. In this greedy algorithm, we select the locations of pick-up stations sequentially. In selecting the location of the first pick-up station, we search through each possible location in our grid map and calculate consumers' corresponding travel distances to that location. For each location, we then compute each consumer's utility by Equation (2.2)-(2.4) and in turn her overall purchase rate across home and station delivery by Equation (2.5). Finally, we calculate the average purchase rate among all consumers and select the best location of the first pick-up station by searching through each possible location and choosing the one with the largest average purchase rate. In selecting the location of the second pick-up station, we again search among all potential locations. Since we have already selected a location, this time consumers' corresponding travel distances to the closest station under each potential location are calculated by the minimum value between the distance to this potential location and the distance to the previously selected locations. Then we repeat the same process as the previous step to get the best location of the third pick-up station. In choosing the locations of the 3^{rd} to the 55^{th} pick-up stations, we repeat the same step 53 times.

Figure 2.5(b) shows our rearranged pick-up station locations (in red triangles) along with their original locations (in blue circles) and the population density map. Our location strategy concentrates more on grids with higher consumer density compared to the original station distribution of Alibaba. For example, the locations in eastern rural areas are moved to the northwest where consumer density is much higher. We also demonstrate the marginal sales lift as new stations are gradually introduced using our greedy algorithm in Figure 2.5(c).

Table 2.5: Performance of Location Counterfactuals

	Before	After	Improvement
Panel A: Without effective distance			
Purchase rate	0.097	0.099	0.002
Annual GMV (million)	33,247	33,945	698
Daily consumer welfare	262.88	268.24	5.36
95% group daily welfare	41.23	44.00	2.77
75% group daily welfare	0	0	0
Panel B: 1 km effective distance			
Purchase rate	0.087	0.093	0.006
Annual GMV (million)	33,247	35,535	2,288
Daily consumer welfare	234.59	252.18	17.59
95% group daily welfare	27.71	35.90	8.19
75% group daily welfare	0	0	0

Note: Annual GMV is measured in CNY. Consumer welfare is measured in CNY per consumer per day.

Specifically, the blue dashed line shows the average purchase rate among all consumers (*i.e.*, consumers inside and outside our estimation sample) with the original 55 pick-up station locations (0.097), and the red dots mark the average purchase rate when we gradually add pick-up stations to locations selected by our greedy algorithm. Panel A of Table 2.5 summarizes the purchase rate before and after the rearrangement. After rearranging the locations of these pick-up stations, the average purchase rate would increase to 0.099, representing a 2.1% sales lift compared to the original locations (Row 1 of Panel A). Considering the annual GMV in the city (33,247 million CNY), our back-of-envelope calculation shows that the rearrangement could potentially increase annual sales by $33,247 \times 2.1\% = 698$ million in CNY (Row 2 of Panel A).

More importantly, as shown in Figure 2.5(c), there is diminishing marginal return in adding new pick-up stations and we can reach the original purchase probability using just 24 new locations based on our rearrangement strategy. This represents a 56.4 percentage reduction in the number of pick-up stations required. These 24 station locations (in red circles) as well as the locations of the original pick-up stations are shown in Figure 2.5(d). Compared with the station locations in Figure 2.5(b), Figure 2.5(d) focuses on locations in the center of the city where more population is covered. As a robustness analysis, we also conduct

the counterfactual exercise under the assumption that stations have an effective distance of 1.00 km, which matches the distance scope of our structural estimation. The performance is summarized in Panel B of Table 2.5. After the rearrangement, the increase in average purchase rate would be even higher (from 0.087 to 0.093, or a 6.9% sales lift, Row 1 of Panel B). Besides, we can reach the original purchase probability using just 20 new locations based on our strategy (or a 63.6% reduction).

Last but not least, we calculate the welfare implications of our counterfactual policy. For each consumer, we draw a simulation to estimate the average daily consumer welfare, where we randomly draw the consumer's idiosyncratic shock for station delivery, home delivery, and the outside option. The utilities are calculated from Equation (2.2), (2.3), and (2.4). Since we do not have price data in our sample, we then use the price coefficient of -0.040 (for 100 CNY) estimated by [96] to calculate consumer welfare[12]. We divide the utilities by this price coefficient and find the increase in consumer welfare to be 5.36 CNY per consumer per day (or equivalently, a 2.0% increase, Row 3 of Panel A). We also include different majority groups in quantiles for consumer welfare by computing the average daily consumer welfare for each group. For a 95% (with the top 5% outlier excluded) majority group, the welfare improvement would be 2.77 CNY (or 6.7%, Row 4 of Panel A) per consumer per day, and the average improvement is zero for a 75% majority group (Row 5 of Panel A).

2.6.3 Shipping Window Strategy

Our estimated structural parameters also provide us with consumers' time preferences, which allows us to improve the existing shipping strategy. In this counterfactual analysis, we will redesign the shipping strategy by reassigning the delivery time across consumers under the

[12]This number is also consistent with the range of price coefficient estimated in [97], which is between -0.28 and -0.12 for 100 USD, or between -0.045 and -0.019 for 100 CNY.

constraint of delivery capacity. We assume the platform can only decide the time windows but not the exact delivery time of each consumer because the latter depends on many operational details, such as routing and traffic, over which the platform has no control. Therefore, we divide the delivery time into two delivery windows since the distribution of delivery time is bimodal with one peak in the morning and another in the afternoon. In order to do so, we estimate a Gaussian mixture model under two classes (see Appendix B.4 for the distribution of the estimated Gaussian mixture model), and each class represents the normal distribution of delivery time in one time window.

The number of deliveries in the first delivery window takes a proportion of 53.0% among all deliveries. The delivery time in this time window follows a Gaussian distribution with the mean time of delivery at 11:19 a.m. and a standard deviation of 1.855 hours. Thereafter, we call this time window "the morning delivery window". The second delivery window takes a proportion of 47.0%. It follows a distribution with the mean time of delivery at 5:02 p.m. and a standard deviation of 1.815 hours. Thereafter, we call this delivery window "the afternoon delivery window". We then fix the delivery capacity by assuming that 53.0% and 47.0% as the capacity constraints for the morning delivery window and the afternoon delivery window. In other words, 53.0% deliveries should be in the morning and 47.0% deliveries should be in the afternoon. If a consumer is assigned to the morning delivery window, the exact delivery time follows the first normal distribution. If the consumer is assigned to the afternoon window, the exact delivery time follows the second normal distribution. Note that compared to previous sections where we assume the distribution of delivery time is homogeneous to all consumers, this exercise will assume delivery time is homogeneous to customers within their fixed time window.

Under this new delivery time distributions and the corresponding capacity constraints, we then try to optimally rearrange consumers' delivery windows between morning and afternoon

delivery windows. Let $\mathbf{P^L} = (P_0^L, P_1^L)$, $\mathbf{P^M} = (P_0^M, P_1^M)$, and $\mathbf{P^{HH}} = (P_0^{HH}, P_1^{HH})$ be vectors representing the purchase probabilities for different types of consumers, where P_i^k is the purchase probability for type k consumers who are assigned to the delivery window i, $i = 0, 1$ represents a delivery window in morning or afternoon, and $k = L, M, HH$ represents a consumer type as low-type, medium-type, or high-type, respectively. Similarly, let $\mathbf{w^L} = (w_0^L, w_1^L)$, $\mathbf{w^M} = (w_0^M, w_1^M)$, and $\mathbf{w^H} = (w_0^{HH}, w_1^{HH})$ be vectors representing the proportions of the population allocated to morning and afternoon delivery windows for low type, medium type, and high type consumers, respectively. Let $\tau_0 (= 53.0\%)$ and $\tau_1 (= 47.0\%)$ be the capacities for the morning delivery window and the afternoon delivery window. Let $\eta^L (= 5.2\%)$, $\eta^M = (85.7\%)$, and $\eta^{HH} (= 9.2\%)$ denote the proportions of consumers in low type, medium type, and high type in the entire city, where the proportions of consumer types have been calculated in Section 2.6.2. The objective of the optimization problem for shipping window rearrangement is to maximize the purchase rate in the entire city and can be written as:

$$
\begin{aligned}
\max_{\mathbf{w^L}, \mathbf{w^M}, \mathbf{w^{HH}}} \quad & (\mathbf{P^L}, \mathbf{P^M}, \mathbf{P^{HH}})^T (\mathbf{w^L}, \mathbf{w^M}, \mathbf{w^{HH}}) \\
\text{subject to} \quad & \sum_{k \in \{L,M,HH\}} w_i^k \leq \tau_i, && i \in \{0, 1\} \\
& \sum_{i \in \{0,1\}} w_i^k = \eta^k, && k = \{L, M, HH\} \\
& w_i^k \in [0, 1], && i \in \{0, 1\},\ k = \{L, M, HH\}
\end{aligned}
$$

The objective function is the purchase rate in the city, which equals the product of calculated purchase probabilities associated with each delivery window per consumer type, and the proportion of the population allocated to each delivery window for different types of consumers. The first row in the optimization constraints represents the set of delivery capacity constraints. The second row represents the constraints that after allocation, the population for morning

Table 2.6: Delivery Windows Counterfactuals

	Panel A: Parameters and results			
	Purchase probability P_i^k		Delivery window allocation w_i^k	
	Morning delivery window ($i=0$)	Afternoon delivery window ($i=1$)	Morning delivery window ($i=0$)	Afternoon delivery window ($i=1$)
Low type ($k=L$)	0.262	0.345	5.2%	0.0%
Medium type ($k=M$)	0.032	0.182	47.8%	37.9%
High type ($k=HH$)	0.012	0.348	0.0%	9.2%
	Panel B: Comparison of Performance			
	Before	After	Improvement	
Purchase rate	0.119	0.129	0.01	
Annual GMV (million)	33,247	36,040	2,793	
Daily consumer welfare	327.22	359.99	32.77	
95% group daily welfare	138.23	160.09	21.14	
75% group daily welfare	5.07	7.04	1.97	

Note: Annual GMV is measured in CNY. Consumer welfare is measured in CNY per consumer per day.

and afternoon delivery windows should be equal to the total population of each consumer type. Finally, the third row shows the constraints for the feasible sets of $\mathbf{w^L}$,$\mathbf{w^M}$, and $\mathbf{w^{HH}}$[13].

In Column 1-2 of Panel A of Table 2.6, we first report the calculated value of $\mathbf{P^L}$, $\mathbf{P^M}$, and $\mathbf{P^{HH}}$. The purchase probabilities are slightly different with Equation (2.5) in Section 2.5.1, since consumer j chooses only between home delivery option and the outside option, and can be written as:

$$P_{ij} = \frac{exp(v_{ij}^H)}{1 + exp(v_{ij}^H)}.$$

We then report the optimization results for the optimized delivery window allocation $\mathbf{w^L}$, $\mathbf{w^M}$, and $\mathbf{w^{HH}}$ in Column 3-4 of Table 2.6. Under the Gaussian mixture model, the purchase rate is 0.119. If we rearrange delivery windows, the purchase rate would be increased to 0.129 (or equivalently, increased by 8.4%, Row 1 of Panel B). Considering the annual platform GMV in the city (33, 247 million CNY), this strategy could potentially increase annual sales

[13]Note that $\mathbf{P^L}, \mathbf{P^M}, \mathbf{P^{HH}}$ is not a function of delivery time rearrangement ($\mathbf{w^L}, \mathbf{w^M}, \mathbf{w^{HH}}$) since we have assumed that the delivery time capacities in each time window do not change. If our strategy allows not only re-arranging customers but also changing delivery time capacities in each time window, the optimization problem will become much harder since the purchase rate will then depend on the delivery capacity rearrangement.

by $33,247 \times 8.4\% = 2,793$ million (Row 2 of Panel B) in CNY. Similar with Section 2.6.2, we draw 10,000 simulations for each consumer type in each delivery window to estimate the average daily consumer welfare from the utilities of home deliveries in Equation (2.3). We also show the consumer welfare for the whole population and each majority group. After the rearrangement, the average daily consumer welfare would be increased by 32.77 CNY per consumer for the whole population (or 10.0%, Row 3 of Panel B). For a 95% majority group and a 75% majority group, the welfare improvement would be 21.14 CNY (or 15.3%, Row 4 of Panel B) and 1.97 CNY (or 38.9%, Row 5 of Panel B) per consumer per day.

Chapter 3

Conclusion

This dissertation contributes to the emerging operations issues on data-driven platforms and digital operations by examining in-warehouse worker behavior and out-of-warehouse customer behavior in the last mile of logistics. In Chapter 1, we study how algorithmic (vs. human-based) task assignment processes change workers' fairness perceptions and productivity. In Chapter 2, we use the introduction of local pick-up stations to study the impact of improving logistic flexibility on customer behavior in online retailing.

We study the impact of algorithmic (vs. human-based) work assignment on assignment recipients' fairness perceptions and productivity in Chapter 1. In two randomized field experiments, we randomly assigned picking workers in one of Alibaba's warehouses to receive tasks either from an algorithm or from a human distributor. Combining performance data from Alibaba's digital labor system with survey responses, we present several findings.

First, we find that assignment recipients' fairness perceptions change with the framing of how their tasks are determined. In our field setting where workers believe that task assignments should prioritize equality over consideration of personal characteristics, receiving tasks from

an algorithm increases workers' perceived fairness by 0.94-1.02 standard deviations (depending on the inclusion of control variables), relative to receiving tasks of an identical nature from a human. While we sought to ensure that tasks were distributed to workers in both groups using the same underlying rule, workers may believe that algorithms can apply rules consistently across workers and treat everyone equally but human distributors have the discretion to favor some workers, as our interviews suggest.

Second, we find that the two types of task assignment methods have an economically meaningful difference in productivity: workers' picking efficiency increases by 15.56%-17.86% when pick lists are assigned by an algorithm than when pick lists are assigned by a human distributor. This is driven by the positive effect of fairness perceptions on productivity. Using the IV approach, we estimate that a one-standard-deviation increase in perceived fairness is associated with a boost of picking efficiency by 12.97%-16.98%.

In addition, we find that the effects of algorithmic assignment on fairness perceptions and productivity are not limited to the first experiment day when workers might find algorithmic assignment novel but persist on later days after workers have gained more experience with it. Our analysis of heterogeneous treatment effects provides suggestive evidence that the positive effects of algorithmic assignment (relative to human-based assignment) on fairness perceptions and productivity hold when workers experience particularly high or low task difficulty, and are stronger among workers who reported being more (vs. less) upset about receiving difficult tasks.

We conducted two auxiliary online experiments. Similar to Chinese picking workers in our collaborating warehouse, U.S. survey respondents also believe that it is more important for a task assignment process to maintain equality than to consider task recipients' unique characteristics. They expect algorithms to be better at delivering equal treatments across

workers than human task distributors. As a generalization of our finding about perceived fairness in the field experiments, our online experiments reveal that people in Western culture also perceive an algorithmic assignment process to be fairer than a human-based assignment process, even when assignment outcomes are the same between these two processes.

Our research has important practical implications. First, our results highlight that when algorithms are applied to solve operational problems, they can have broader impacts beyond offering greater efficiency and accuracy than humans. Managers may want to introduce algorithmic assignment processes to reap their benefits on fairness perceptions and productivity. Second, our ability to observe productivity differences between groups even when the algorithm and human distributors assigned objectively comparable pick lists suggests that the framing and people's beliefs about an assignment process also matter to productivity. This insight encourages managers to find the most motivating framing of task assignment processes for their workers and to understand their workers' beliefs about different allocation processes. Third, our results underscore the important role of psychological factors such as fairness perceptions in driving workers' motivation and productivity. Simple strategies like drawing employees' attention to algorithmic task assignment processes that are used in their organizations may lead employees to perceive their managers as caring about fairness, which could be beneficial for employees' performance.

In contrast with past literature in e-commerce logistics that generally focuses on shipping speed improvement, in Chapter 2, we study an alternative core aspect of delivery experience to lift sales—improving logistic flexibility. Using the introduction of pick-up stations as a natural experiment to consumers, we demonstrate via a DiD approach that pick-up station opening has increased the sale by 3.8%, and this increases come mostly from existing consumers. We also provide suggestive evidence that shipping speed is not the primary driver of this sales lift, and instead, logistic flexibility could be the main mechanism.

To further study the impact of logistics flexibility on consumer behaviors, we develop and estimate a two-stage structural consumer choice model that describes how logistic flexibility may affect consumers' purchase decisions. From the structural estimation, we find that there are three classes of consumers. For all three classes of consumers, the after-work period is generally preferred over the daytime period. Moreover, our model estimation shows that consumers differently value two types of logistic flexibility—pickup time flexibility and choice flexibility: 76.2% of purchase rate improvement brought by pick-up stations is from the value of pickup time flexibility, and the rest is from the value of choice flexibility. Moreover, compared to the majority segment (71% of consumers), the two smaller segments (19% and 11% of consumers, respectively) place a higher value on choice flexibility and time flexibility. That is to say, the pick-up station is most valuable for these two segments.

In addition, we use two counterfactual studies to improve the logistic decisions based on our estimation results. First, we demonstrate that better relocating pick-up stations using our estimates can improve the purchase rate by 2.1%-6.9% with the same number of pick-up stations or reduce the number of pick-up stations by 56.4%-63.6% while maintaining the purchase rate. The consumer welfare would be increased by 2.0%-7.5%. Second, we also re-design a counterfactual shipping policy without changing the delivery capacity based on our estimated consumers' time preferences. We demonstrate that this counterfactual policy could improve sales by 8.4% and improve consumer welfare by 10.0%. Our back-of-envelope calculations show that these counterfactual policies could potentially create billions in annual sales increases in CNY for the platform.

Chapter 1 opens up avenues for future research. First, previous work in algorithmic bias suggests that algorithms at times exhibit discrimination towards females and underrepresented racial minorities [12, 14, 16, 17]. It would be interesting to examine whether females and racial minorities express different sentiments towards algorithmic assignment compared to

males and those who are not racial minorities. Prior research suggests that compared with males, females care more about procedural fairness [98], put greater emphasis on equal opportunity for all [99], and are more inequality averse in economic experiments [100]. In our context, given that we propose people view algorithmic assignment as fairer than human-based assignment when they particularly care about equality (over other values), the positive effect of algorithmic assignment on perceived fairness may be larger among females than males. Consistent with this speculation, we find that algorithmic assignment has a directionally larger impact on fairness perceptions among females than among males, but the difference is not statistically significant (see Appendix A.8).

Second, while we deliberately tone down the impact of operational transparency in our field experiments so as to focus on perceived fairness as the underlying mechanism (as discussed in Section 1.4.1), it would be valuable for future research to vary the level of transparency in algorithmic and human-based assignment processes and understand the role of transparency in driving people's responses to algorithmic decision making in other settings. Third, since most workers only participate in our main field experiment for no more than two days, we do not have enough samples to examine how the effects of algorithmic (vs. human-based) assignment on fairness and productivity change in the long term. Besides, our field experiments only involve temporary workers. We encourage future research to study how workers' responses to algorithmic decision-making processes change over time as well as whether regular workers' prior work experiences moderate such responses.

Although Chapter 2 makes significant progress in modeling the value of logistic flexibility, it also has certain limitations that point to promising future research directions. First, it would be interesting to see how consumer demographics, such as gender, age, education level, and where and when they work, could affect their value of logistic flexibility. Understanding this heterogeneity would help the platform better customize logistic policy. Second, we

assume that consumers have same value of logistic flexibility across days and value of items, ignoring the difference between workdays and holidays as well as low-value and high-value items. Future research may investigate how the differences in time and purchased items affect consumers' value of logistic flexibility. Third, while our research focuses on the additional logistics flexibility introduced by pick-up stations, these stations could also change consumers' shipping experiences in other dimensions. For example, pick-up stations may increase the safety of packages, and it would be interesting to study how consumers react to changes brought by pick-up stations in other dimensions. Finally, our study focuses on the city of Shengyang, which is one of the largest cities in China. It is important to also study the value of logistic flexibility in rural areas, and in cities in other countries such as the U.S., where it is easier to leave the package in front of the door when no one is at home.

References

[1] Atul Bhandari, Alan Scheller-Wolf, and Mor Harchol-Balter. "An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers." In: *Management Science* 54.2 (2008), pp. 339–353.

[2] Yu Zhang and Vidyadhar Kulkarni. "Automated teller machine replenishment policies with submodular costs." In: *Manufacturing & Service Operations Management* 20.3 (2018), pp. 517–530.

[3] Jiankun Sun, Dennis J Zhang, Haoyuan Hu, and Jan A Van Mieghem. "Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations." In: *Management Science* (2021).

[4] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. "Persuasiveness of expert systems." In: *Behaviour & Information Technology* 17.3 (1998), pp. 155–163.

[5] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114.

[6] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them." In: *Management Science* 64.3 (2018), pp. 1155–1170.

[7] Eugina Leung, Gabriele Paolacci, and Stefano Puntoni. "Man versus machine: Resisting automation in identity-based consumer behavior." In: *Journal of Marketing Research* 55.6 (2018), pp. 818–831.

[8] Jennifer M Logg, Julia A Minson, and Don A Moore. "Algorithm appreciation: People prefer algorithmic to human judgment." In: *Organizational Behavior and Human Decision Processes* 151 (2019), pp. 90–103.

[9] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases." In: *Marketing Science* 38.6 (2019), pp. 937–947.

[10] David T Newman, Nathanael J Fast, and Derek J Harmon. "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions." In: *Organizational Behavior and Human Decision Processes* 160 (2020), pp. 149–167.

[11] Bo Cowgill and Catherine E Tucker. "Economics, fairness and algorithmic bias." In: *Preparation for: Journal of Economic Perspectives* (2019).

[12] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. "Algorithmic fairness." In: *AEA Papers and Proceedings*. Vol. 108. 2018, pp. 22–27.

[13] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." In: *arXiv preprint arXiv:1609.05807* (2016).

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." In: *Science* 356.6334 (2017), pp. 183–186.

[15] Bo Cowgill. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." In: *Columbia Business School, Columbia University* (2018).

[16] Anja Lambrecht and Catherine Tucker. "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads." In: *Management Science* 65.7 (2019), pp. 2966–2981.

[17] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations." In: *Science* 366.6464 (2019), pp. 447–453.

[18] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. "Resistance to medical artificial intelligence." In: *Journal of Consumer Research* 46.4 (2019), pp. 629–650.

[19] Noah Castelo, Maarten W Bos, and Donald R Lehmann. "Task-Dependent Algorithm Aversion." In: *Journal of Marketing Research* 56.5 (2019), pp. 809–825.

[20] Ernst Fehr and Klaus M Schmidt. "A theory of fairness, competition, and cooperation." In: *The Quarterly Journal of Economics* 114.3 (1999), pp. 817–868.

[21] Christopher T Dawes, James H Fowler, Tim Johnson, Richard McElreath, and Oleg Smirnov. "Egalitarian motives in humans." In: *Nature* 446.7137 (2007), pp. 794–796.

[22] Peter R Blake and Katherine McAuliffe. ""I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity." In: *Cognition* 120.2 (2011), pp. 215–224.

[23] Karel H Van Donselaar, Vishal Gaur, Tom Van Woensel, Rob ACM Broekmeulen, and Jan C Fransoo. "Ordering behavior in retail stores and implications for automated replenishment." In: *Management Science* 56.5 (2010), pp. 766–784.

[24] Nil Karacaoglu, Antonio Moreno, and Can Ozkan. "Strategically Giving Service: The Effect of Real-Time Information on Service Efficiency." In: *Available at SSRN 3260035* (2018).

[25] Yao Li, Lauren Xiaoyuan Lu, Susan F Lu, and Jian Chen. "The value of health it interoperability: Evidence from interhospital transfer of heart attack patients." In: *Available at SSRN 3557010* (2021).

[26] Robert S Huckman, Bradley R Staats, and David M Upton. "Team familiarity, role experience, and performance: Evidence from Indian software services." In: *Management science* 55.1 (2009), pp. 85–100.

[27] Tom Fangyun Tan and Serguei Netessine. "When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance." In: *Management Science* 65.8 (2019), pp. 3495–3517.

[28] Maria R Ibanez and Michael W Toffel. "How scheduling can bias quality assessment: Evidence from food-safety inspections." In: *Management Science* 66.6 (2020), pp. 2396–2416.

[29] Yuqian Xu and Lingjiong Zhu. "Operational Risk Management: Team-Based Effort and Incentive Bonus." In: *Available at SSRN 3191887* (2020).

[30] Zeynep Akşin, Sarang Deo, Jónas Oddur Jónasson, and Kamalini Ramdas. "Learning from many: Partner exposure and team familiarity in fluid teams." In: *Management Science* 67.2 (2021), pp. 854–874.

[31] Anqi Angie Wu and Yixin Iris Wang. "The More Monitoring, the Better Quality? Empirical Evidence from the Generic Drug Industry." In: *Empirical Evidence from the Generic Drug Industry (January 27, 2021)* (2021).

[32] Ryan W Buell and Michael I Norton. "The labor illusion: How operational transparency increases perceived value." In: *Management Science* 57.9 (2011), pp. 1564–1579.

[33] Diwas Singh Kc. "Heuristic thinking in patient care." In: *Management Science* 66.6 (2020), pp. 2545–2563.

[34] Tom R Tyler. "The psychology of procedural justice: A test of the group-value model." In: *Journal of Personality and Social psychology* 57.5 (1989), p. 830.

[35] Jason A Colquitt, Donald E Conlon, Michael J Wesson, Christopher OLH Porter, and K Yee Ng. "Justice at the millennium: a meta-analytic review of 25 years of organizational justice research." In: *Journal of Applied Psychology* 86.3 (2001), p. 425.

[36] Armin Falk, Ernst Fehr, and Urs Fischbacher. "Testing theories of fairness—Intentions matter." In: *Games and Economic Behavior* 62.1 (2008), pp. 287–303.

[37] Heather M Gray, Kurt Gray, and Daniel M Wegner. "Dimensions of mind perception." In: *Science* 315.5812 (2007), pp. 619–619.

[38] Kurt Gray and Daniel M Wegner. "Feeling robots and human zombies: Mind perception and the uncanny valley." In: *Cognition* 125.1 (2012), pp. 125–130.

[39] Yochi Cohen-Charash and Paul E Spector. "The role of justice in organizations: A meta-analysis." In: *Organizational Behavior and Human Decision Processes* 86.2 (2001), pp. 278–321.

[40] Ernst Fehr, Lorenz Goette, and Christian Zehnder. "A behavioral account of the labor market: The role of fairness concerns." In: *Annual Reviews of Economics* 1.1 (2009), pp. 355–384.

[41] Ryan W Buell, Ethan Porter, and Michael I Norton. "Surfacing the submerged state: Operational transparency increases trust in and engagement with government." In: *Manufacturing & Service Operations Management* (2020).

[42] Ryan W Buell, Tami Kim, and Chia-Jung Tsay. "Creating reciprocal value through operational transparency." In: *Management Science* 63.6 (2017), pp. 1673–1695.

[43] Ipek Demirdag and Suzanne Shu. "Insights Into the Black Box: Input Explainability of Algorithmic Decisions Drives Consumer Satisfaction in the Digital World." In: *ACR North American Advances* (2020).

[44] Peter M Aronow. "A general method for detecting interference between units in randomized experiments." In: *Sociological Methods & Research* 41.1 (2012), pp. 3–16.

[45] Tanjim Hossain and John A List. "The behavioralist visits the factory: Increasing productivity using simple framing manipulations." In: *Management Science* 58.12 (2012), pp. 2151–2167.

[46] Nicholas Bloom, James Liang, John Roberts, and Zhichun Jenny Ying. "Does working from home work? Evidence from a Chinese experiment." In: *The Quarterly Journal of Economics* 130.1 (2015), pp. 165–218.

[47] Joe E Isaac. "Performance related pay: The importance of fairness." In: *Journal of Industrial Relations* 43.2 (2001), pp. 111–123.

[48] Delroy L Paulhus. "Measurement and control of response bias." In: *Measures of personality and social psychological attitudes* (1991), pp. 17–59.

[49] Margie E Lachman and Suzanne L Weaver. "The sense of control as a moderator of social class differences in health and well-being." In: *Journal of Personality and Social Psychology* 74.3 (1998), p. 763.

[50] Bradley R Staats, Diwas S Kc, and Francesca Gino. "Maintaining beliefs in the face of negative news: The moderating role of experience." In: *Management Science* 64.2 (2018), pp. 804–824.

[51] Maria R Ibanez and Bradley R Staats. "Behavioral empirics and field experiments." In: *The Handbook of Behavioral Operations* (2018), pp. 121–147.

[52] Donald E Conlon, Christopher OLH Porter, and Judi McLean Parks. "The fairness of decision rules." In: *Journal of Management* 30.3 (2004), pp. 329–349.

[53] Marshall L Fisher, Santiago Gallino, and Joseph Jiaqi Xu. "The value of rapid delivery in omnichannel retailing." In: *Journal of Marketing Research* 56.5 (2019), pp. 732–748.

[54] Stanley E Griffis, Shashank Rao, Thomas J Goldsby, Clay M Voorhees, and Deepak Iyengar. "Linking order fulfillment performance to referrals in online retailing: an empirical analysis." In: *Journal of Business logistics* 33.4 (2012), pp. 279–294.

[55] Wenzheng Mao, Liu Ming, Ying Rong, Christopher S Tang, and Huan Zheng. "Faster deliveries and smarter order assignments for an on-demand meal delivery platform." In: *Available at SSRN 3469015* (2019).

[56] Ruomeng Cui, Zhikun Lu, Tianshu Sun, and Joseph Golden. "Sooner or later? Promising delivery speed in online retail." In: *Working Paper* (2020).

[57] Vinayak Deshpande and Pradeep K Pendem. "Logistics performance, ratings, and its impact on customer purchasing behavior and sales in e-commerce platforms." In: *Manufacturing & Service Operations Management* (2022).

[58] Xenophon Koufteros, Cornelia Droge, Gregory Heim, Nelson Massad, and Shawnee K Vickery. "Encounter satisfaction in e-tailing: are the relationships of order fulfillment service quality with its antecedents and consequences moderated by historical satisfaction?" In: *Decision Sciences* 45.1 (2014), pp. 5–48.

[59] Nooshin Salari, Sheng Liu, and Zuo-Jun Max Shen. "Real-time delivery time forecasting and promising in online retailing: when will your package arrive?" In: *Manufacturing & Service Operations Management* (2022).

[60] Hanzhang Qin, David Simchi-Levi, Ryan Ferer, Jonathan Mays, Ken Merriam, Megan Forrester, and Alex Hamrick. "Trading Safety Stock for Service Response Time in Inventory Positioning." In: *Available at SSRN 4066119* (2022).

[61] Sanjith Gopalakrishnan, Moksh Matta, Mona Imanpoor Yourdshahy, and Vivek Choudhary. "Go Wide or Go Deep? Assortment Strategy and Order Fulfillment in Online Retail." In: *Manufacturing & Service Operations Management* (2022).

[62] Wenchang Zhang, Christopher S Tang, Liu Ming, and Yue Cheng. "Reducing Traffic Incidents in Meal Delivery: Penalize the Platform or its Independent Drivers?" In: *Available at SSRN 4231746* (2022).

[63] Jifeng Luo, Ying Rong, and Huan Zheng. "Impacts of logistics information on sales: Evidence from Alibaba." In: *Naval Research Logistics (NRL)* 67.8 (2020), pp. 646–669.

[64] Ruomeng Cui, Meng Li, and Qiang Li. "Value of high-quality logistics: Evidence from a clash between SF Express and Alibaba." In: *Management Science* 66.9 (2020), pp. 3879–3902.

[65] Robert L Bray. "Operational transparency: Showing when work gets done." In: *Manufacturing & Service Operations Management* (2020).

[66] Ken Moon, Kostas Bimpikis, and Haim Mendelson. "Randomized markdowns and online monitoring." In: *Management Science* 64.3 (2018), pp. 1271–1290.

[67] Ruomeng Cui, Dennis J Zhang, and Achal Bassamboo. "Learning from inventory availability information: Evidence from field experiments on Amazon." In: *Management Science* 65.3 (2019), pp. 1216–1235.

[68] Xiaoyang Long, Jiankun Sun, Hengchen Dai, Dennis Zhang, Jianfeng Zhang, Yujie Chen, Haoyuan Hu, and Binqiang Zhao. "Choice Overload with Search Cost and Anticipated Regret: Theoretical Framework and Field Evidence." In: *Available at SSRN 3890056* (2021).

[69] Santiago Gallino, Nil Karacaoglu, and Antonio Moreno. "Need for Speed: The Impact of In-Process Delays on Customer Behavior in Online Retail." In: *Operations Research* (2022).

[70] Dennis J Zhang, Hengchen Dai, Lingxiu Dong, Qian Wu, Lifan Guo, and Xiaofei Liu. "The value of pop-up stores on retailing platforms: Evidence from a field experiment with Alibaba." In: *Management Science* 65.11 (2019), pp. 5142–5151.

[71] Brian Rongqing Han, Leon Yang Chu, Tianshu Sun, and Lixia Wu. "Commercializing the package flow: Cross-sampling physical products through e-commerce warehouses." In: *Available at SSRN 3566756* (2020).

[72] Ruchi Mishra, Rajesh Kumar Singh, and Bernadett Koles. "Consumer decision-making in Omnichannel retailing: Literature review and future research agenda." In: *International Journal of Consumer Studies* 45.2 (2021), pp. 147–174.

[73] Santiago Gallino and Antonio Moreno. "Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information." In: *Management Science* 60.6 (2014), pp. 1434–1451.

[74] Santiago Gallino, Antonio Moreno, and Ioannis Stamatopoulos. "Channel integration, sales dispersion, and inventory management." In: *Management Science* 63.9 (2017), pp. 2813–2831.

[75] Fei Gao, Vishal V Agrawal, and Shiliang Cui. "The effect of multichannel and omnichannel retailing on physical stores." In: *Management Science* 68.2 (2022), pp. 809–826.

[76] Fei Gao and Xuanming Su. "Omnichannel retail operations with buy-online-and-pick-up-in-store." In: *Management Science* 63.8 (2017), pp. 2478–2492.

[77] Ming Hu, Xiaolin Xu, Weili Xue, and Yi Yang. "Demand pooling in omnichannel operations." In: *Management science* 68.2 (2022), pp. 883–894.

[78] Timothy F Bresnahan and Peter C Reiss. "Entry and competition in concentrated markets." In: *Journal of political economy* 99.5 (1991), pp. 977–1009.

[79] Steven T Berry. "Estimation of a Model of Entry in the Airline Industry." In: *Econometrica: Journal of the Econometric Society* (1992), pp. 889–917.

[80] Thomas J Holmes. "The diffusion of Wal-Mart and economies of density." In: *Econometrica* 79.1 (2011), pp. 253–302.

[81] Mitsukuni Nishida. "Estimating a model of strategic network choice: The convenience-store industry in Okinawa." In: *Marketing Science* 34.1 (2015), pp. 20–38.

[82] Ali Umut Guler. "Inferring the economics of store density from closures: The Starbucks case." In: *Marketing Science* 37.4 (2018), pp. 611–630.

[83] Chloe Kim Glaeser, Marshall Fisher, and Xuanming Su. "Optimal retail location: Empirical methodology and application to practice." In: *Manufacturing & Service Operations Management* 21.1 (2019), pp. 86–102.

[84] Jingting Fan, Lixin Tang, Weiming Zhu, and Ben Zou. "The Alibaba effect: Spatial consumption inequality and the welfare gains from e-commerce." In: *Journal of International Economics* 114 (2018), pp. 203–220.

[85] Seyed M Iravani, Mark P Van Oyen, and Katharine T Sims. "Structural flexibility: A new perspective on the design of manufacturing and service operations." In: *Management Science* 51.2 (2005), pp. 151–166.

[86] Susan Feng Lu and Lauren Xiaoyuan Lu. "Do mandatory overtime laws improve quality? Staffing decisions and operational flexibility of nursing homes." In: *Management Science* 63.11 (2017), pp. 3566–3585.

[87] William C Jordan and Stephen C Graves. "Principles on the benefits of manufacturing process flexibility." In: *Management science* 41.4 (1995), pp. 577–594.

[88] Stephen C Graves and Brian T Tomlin. "Process flexibility in supply chains." In: *Management Science* 49.7 (2003), pp. 907–919.

[89] Jan A Van Mieghem. "Investment strategies for flexible resources." In: *Management Science* 44.8 (1998), pp. 1071–1078.

[90] Elena Katok, William Tarantino, and Terry P Harrison. "Investment in production resource flexibility: An empirical investigation of methods for planning under uncertainty." In: *Naval Research Logistics (NRL)* 50.2 (2003), pp. 105–129.

[91] Arash Asadpour, Xuan Wang, and Jiawei Zhang. "Online resource allocation with limited flexibility." In: *Management Science* 66.2 (2020), pp. 642–666.

[92] Zhen Xu, Hailun Zhang, Jiheng Zhang, and Rachel Q Zhang. "Online demand fulfillment under limited flexibility." In: *Management Science* 66.10 (2020), pp. 4667–4685.

[93] Levi DeValve, Yehua Wei, Di Wu, and Rong Yuan. "Understanding the value of fulfillment flexibility in an online retailing environment." In: *Manufacturing & Service Operations Management* (2021).

[94] Bicheng Yang, Tat Chan, and Raphael Thomadsen. "A Salesforce-driven model of consumer choice." In: *Marketing Science* 38.5 (2019), pp. 871–887.

[95] Daniel McFadden. "Economic choices." In: *American economic review* 91.3 (2001), pp. 351–378.

[96] Zhenling Jiang, Tat Chan, Hai Che, and Youwei Wang. "Consumer search and purchase: An empirical investigation of retargeting based on consumer online behaviors." In: *Marketing Science* 40.2 (2021), pp. 219–240.

[97] Raluca M Ursu. "The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions." In: *Marketing Science* 37.4 (2018), pp. 530–552.

[98] Paul D Sweeney and Dean B McFarlin. "Process and outcome: Gender differences in the assessment of justice." In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 18.1 (1997), pp. 83–98.

[99] Shalom H Schwartz and Tammy Rubel. "Sex differences in value priorities: cross-cultural and multimethod studies." In: *Journal of Personality and Social Psychology* 89.6 (2005), p. 1010.

[100] James Andreoni and Lise Vesterlund. "Which is the fair sex? Gender differences in altruism." In: *The Quarterly Journal of Economics* 116.1 (2001), pp. 293–312.

[101] Jean M Twenge and W Keith Campbell. "Self-esteem and socioeconomic status: A meta-analytic review." In: *Personality and Social Psychology Review* 6.1 (2002), pp. 59–71.

[102] Batia M Wiesenfeld, William B Swann Jr, Joel Brockner, and Caroline A Bartel. "Is more fairness always preferred? Self-esteem moderates reactions to procedural justice." In: *Academy of Management Journal* 50.5 (2007), pp. 1235–1253.

Appendix A

Appendix for Chapter 1

Appendix A.1: Pick List Example

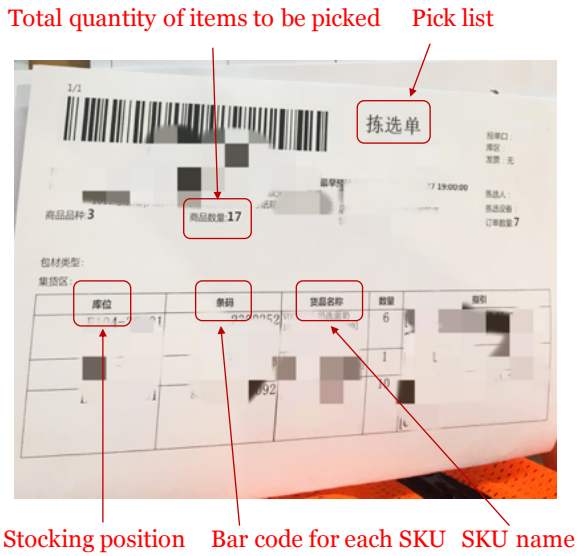


Figure A.1: Pick List Example

Appendix A.2: Second Field Experiment as Replication

We conducted another experiment from December 27th, 2019 to January 5th, 2020 to replicate the main results that we report in the paper. The second experiment involved 20 temporary picking workers. We randomly assigned them into either the algorithm group or the human group. The experimental design was the same as what we described in 1.4.1. These workers completed 3,181 pick lists in total. The sample size of workers was smaller in the second experiment than the first experiment because (1) we could only run the second experiment for 10 days before a large sales period started on January 6th, 2020 and (2) the warehouse reduced labor floating, meaning that workers came to work for more days during the second experiment and consequently leaving us with fewer unique workers. Due to the small sample size of workers, we focus on replicating main effects, rather than heterogeneous treatment effects across different types of workers.

We distributed surveys at the end of each day to workers who worked at the warehouse that day. To assess their fairness perceptions, we asked workers two questions that contrasted algorithmic vs. human-based assignment processes. The first question asked workers, "Which assignment process do you think is fairer, algorithmic assignment or human-based assignment?" This question was measured on a five-point scale, with the anchors ranging from 1 ("Definitively algorithmic assignment") to 5 ("Definitively human-based assignment") for all workers. The second question was the same as the second question in the first experiment (see Table 1.1). For workers in the algorithm group, we reverse coded their answers to *both* questions; for workers in the human group, we made no adjustment to their original answers. Therefore, for workers in both groups, a higher (vs. lower) value on a question indicates that the worker more strongly viewed their current assignment process as fairer than the alternative process. The correlation between workers' responses to these two fairness questions (after reverse

Table A.1: The Effects of Algorithmic (vs. Human-based) Assignment on Perceived Fairness and Productivity and IV-Estimated Effect of Perceived Fairness on Productivity (Replication)

Dependent variable	*Standardized perceived fairness*			*Picking efficiency*					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Algorithm	1.09***	1.13***	1.10***	1.21***	1.10***	0.98***			
	(0.35)	(0.34)	(0.34)	(0.38)	(0.35)	(0.32)			
Standardized perceived fairness							1.20***	1.04***	0.94***
							(0.38)	(0.33)	(0.31)
Day fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Hour fixed effects	No	No	No	No	Yes	Yes	No	Yes	Yes
Demographics controls	No	No	Yes	No	No	Yes	No	No	Yes
Pick list controls	No	No	No	No	No	Yes	No	No	Yes
Observations	87	87	87	3,181	3,181	3,181	3,181	3,181	3,181

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 5.03. Average picking efficiency across algorithm and human groups was 5.57.

coding) was high ($r = 0.84$; $p < 0.0001$). Following the same procedure as described in 1.4.2, we created a score of *Standardized Perceived Fairness* for each worker each day. The survey also asked the same set of demographics as the survey described in the paper.

We analyze the effect of algorithmic (vs. human-based) assignment on fairness perceptions using specification (1). As shown in Columns 1-3 of Table A.1, receiving pick lists from an algorithm significantly increases workers' perceived fairness about their assignment process by 1.09-1.13 standard deviations (depending on what we control for), compared to receiving pick lists from a human distributor (all p-values < 0.003).

We then analyze the effect of algorithmic (vs. human-based) assignment on productivity using specification (2). As shown in Columns 4-6 of Table A.1, algorithmic assignment treatment significantly improves workers' productivity by 19.48-24.06% (depending on control variables), relative to the average picking efficiency of 5.03 in the human-based assignment group (all p-values ≤ 0.002).

Columns 7-9 in Table A.1 shows the average treatment effect of perceived fairness on productivity using IV estimation based on specifications (3)-(4). As perceived fairness increases by one standard deviation, worker productivity is estimated to significantly increase

by 16.88-21.54%, relative to the average picking efficiency of 5.57 across the algorithm and human groups (all p-values < 0.003).

Appendix A.3: Discussion about Interference Between Workers

A.3.1. Discussion about Potential Alternative Experiment Designs

We consider our experimental design (similar to [45, 46]), whereby workers in the same work location were assigned to one of two experimental conditions, the cleanest among all feasible approaches. Other experimental designs that may avoid potential interference between workers are unfortunately not feasible in our case. For example, running the experiments across multiple warehouses and assigning workers to the human or algorithm condition at the warehouse level could reduce communications across conditions. However, picking tasks usually differ greatly across warehouses in terms of the number of items per pick list, the number of stocking positions covered, and walking distance, which could all affect productivity. Thus, we would have to run our experiments in a very large number of warehouses to have a comparable set of control versus treatment warehouses, which would not have been logistically infeasible based on Cainiao's priorities at the time of our experiments.

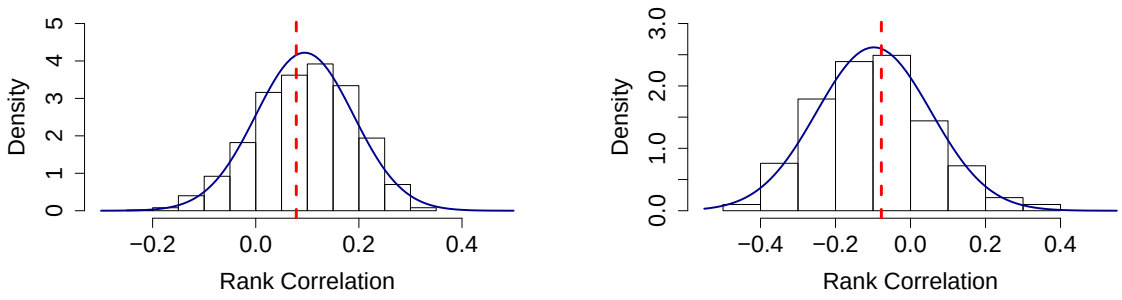
Another alternative design is to run the experiments in one warehouse and randomly assign different days (rather than workers) into one of the two conditions. Though this approach avoids co-location of the algorithm and human groups, it introduces two significant limitations in our setting. First, workers who work in a warehouse for multiple days would likely experience both the algorithm and human conditions in this alternative design, in which case interference

between conditions could arise because workers' experiences with one condition may affect their perceptions and behavior in the other condition. Second, in our setting, picking tasks differ substantially across work days (*e.g.*, promotion days versus non-promotion days). Considering the frequency of promotions for various product categories at Alibaba, we need to run our experiments for at least a few months to get a comparable set of work days assigned to the algorithm group versus the human group, which is infeasible in our setting.

A.3.2: Check Interference Between Workers via a Statistical Test

To perform the test recommended by [44] for our main field experiment, we first randomly select 12 workers from the human condition as the fixed subset. The remaining 38 workers belong to our variant subset. Then we draw 1,000 simulations on the experimental condition of the variant subset. In each simulation, we randomly select 25 workers from the variant subset to be in the algorithm condition and 13 workers to be in the human condition. For each simulation and for each day during our field experiment, we calculate the *simulated daily algorithmic treatment rate*, which equals the proportion of workers who were assigned to the algorithm condition *in the simulation* among all workers coming to work that day. Then for each simulation, we compute the Spearman's rank correlation coefficient ρ between the picking efficiency associated with pick lists assigned to workers in the fixed subset and the simulated daily algorithmic treatment rate.

Across 1,000 simulations, we obtain 1,000 values of ρ. We plot the distribution of ρ in Figure A.2(a). Since workers in the variant subset were purely randomly assigned to the algorithm vs. human condition in each simulation and 1,000 simulations were independent, Figure A.2(a) presents the approximate distribution of ρ associated with the null hypothesis that interference on productivity between workers did not occur for workers in the human group.



(a) Algorithmic Treatment Rate and Productivity (Actually Observed $\rho = 0.08$)

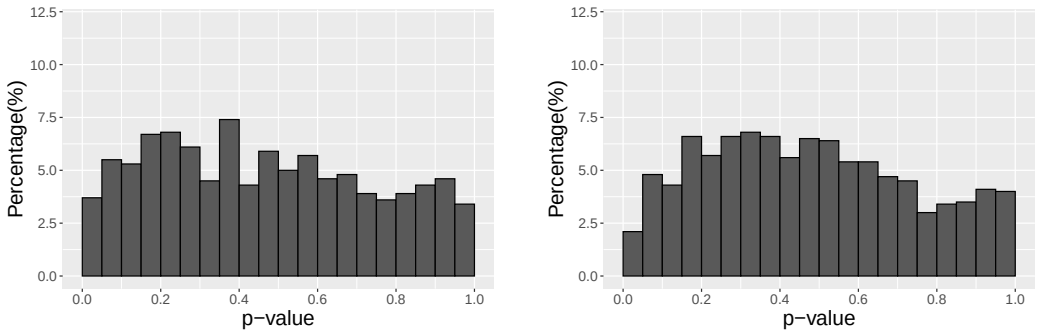
(b) Algorithmic Treatment Rate and Perceived Fairness (Actually Observed $\rho = -0.08$)

Figure A.2: Distributions of Rank Correlation Coefficients Across 1,000 Simulations and the Observed Coefficients

The dashed line in Figure A.2(a) represents the observed correlation coefficient ρ in our first field experiment. The observed ρ is around the center of null distribution, yielding $p = 0.82$ in a two-tailed test (since we do not know a priori whether the observed correlation would be negative or positive). Therefore, we cannot reject the null hypothesis at the 5% level that there is no interference on productivity for workers in the human group.

Using these 1,000 simulations, we perform a similar test on perceived fairness. For each simulation, we compute the Spearman's rank correlation coefficient ρ between the perceived fairness of each worker in the fixed subset on a day and the simulated daily algorithmic treatment rate that day. We plot the distribution of ρ in Figure A.2(b). The observed ρ, as indicated by the dashed line, is again close to the center of sharp null distribution, yielding $p = 0.86$ in a two-tailed test. Thus, we cannot reject the null hypothesis at the 5% level that there is no interference on fairness perceptions for workers in the human group.

So far, we have done 1,000 simulations by taking the same set of 12 workers in the human group as the fixed subset. To confirm the robustness of our test, we randomly draw 12 workers from 25 workers in the human group as the fixed subset for 1,000 times. Each time



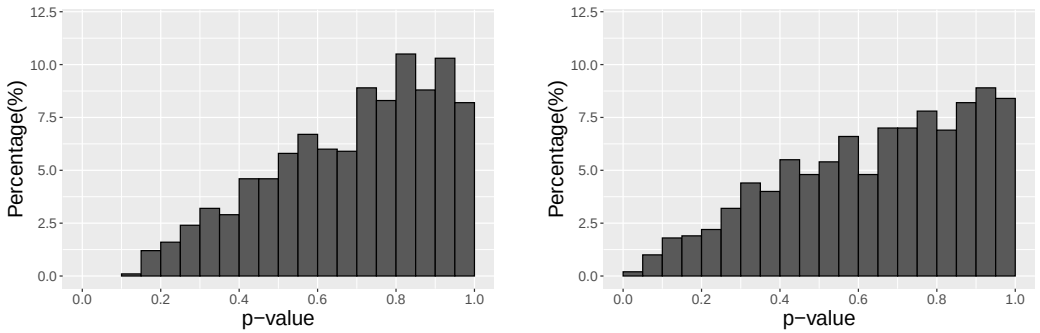
(a) Interference Test on Productivity (b) Interference Test on Fairness

Figure A.3: Distributions of p-values for the Interference Test (Human Group)

we randomly select a fixed subset, we repeat the process described above involving 1,000 simulations and obtain two p-values (one for productivity and one for perceived fairness). Figure A.3(a) shows the distribution of p-values for the interference test on productivity across 1,000 draws of fixed subsets. The p-values from our 1,000 draws of fixed subsets are smaller than 0.05 only 3.70% of the time, lower than 5% (the chance level for p-values to fall below 0.05 under uniform distribution). This suggests that the productivity of workers in the human group is unlikely to have been affected by the interference between the algorithm group and the human group.

Figure A.3(b) shows the distribution of p-values for the interference test on fairness across 1,000 draws of fixed subsets. The p-values from our 1,000 draws are smaller than 0.05 only 2.10% of the time, lower than the chance level of 5% under uniform distribution. This suggests that the perceived fairness of workers in the human group is unlikely to have been affected by the interference between the algorithm and human groups.

We next check the existence of interference for workers in the algorithm group. We follow the same steps as described above, except that we randomly select 12 workers from the algorithm condition as the fixed subset in each of the 1,000 draws. Figure A.4(a) shows the distribution of p-values for the interference test on productivity across 1,000 draws of fixed



(a) Interference Test on Productivity (b) Interference Test on Fairness

Figure A.4: Distributions of p-values for the Interference Test (Algorithm Group)

subsets. None of the p-values from the 1,000 draws is smaller than 0.05. Figure A.4(b) shows the distribution of p-values for the interference test on perceived fairness across 1,000 draws of fixed subsets. Only 0.2% of p-values from the 1,000 draws are smaller than 0.05. Thus, we further confirm that the productivity and fairness perceptions of workers in the algorithm group are unlikely to have been affected by the interference between the algorithm group and the human group.

Appendix A.4. Separately Examining the Two Questions Measuring Fairness Perceptions

In this section, we check the results about fairness perceptions separately using the first versus second question that measures perceived fairness (see Table 1.1), since one may be concerned that the second question does not directly measure workers' fairness perceptions. We show that our results about fairness perception are largely robust to using only the first question.

In the first field experiment, the first question asked workers whether they thought it would be fairer to assign pick lists using the alternative process than using their current process.

Specifically, workers in the algorithm group were asked, "Do you think it would be fairer if pick lists were assigned by a human distributor?"; and workers in the human group were asked, "Do you think it would be fairer if pick lists were assigned by an algorithm?" Workers in both groups responded using a five-point Likert scale from 1 ("Definitively would") to 5 ("Definitively would not"). We constructed *Standardized perceived fairness Q1* (or *Standardized perceived fairness Q2*), which equaled each worker's response to the first question (or the second question) divided by the standard deviation of the responses in the whole sample. As shown in Column 1 of Table A.2, the effect of algorithmic assignment on perceived fairness about worker's assignment process is positive but not statistically significant if we only use the first question to measure fairness perceptions ($p = 0.38$). If we use the second question to measure fairness perceptions, algorithmic assignment significantly increases perceived fairness by 1.36 standard deviations ($p < 0.0001$ in Column 4 of Table A.2). We speculate that the first question alone does not yield a statistically significant effect on fairness perceptions because half (50.93%) of the responses were 3 (the middle option of the five-point scale), indicating a neutral preference between the current and alternative assignment processes.

Table A.2: The Effects of Algorithmic (vs. Human-based) Assignment on Perceived Fairness (Based on the First vs. Second Question)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	*Standardized perceived fairness Q1*			*Standardized perceived fairness Q2*		
Data:	First experiment	Second experiment	Combined experiment	First experiment	Second experiment	Combined experiment
Algorithm	0.23	1.02***	0.65***	1.36****	1.08****	1.26****
	(0.26)	(0.38)	(0.24)	(0.20)	(0.30)	(0.19)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	106	87	193	106	87	193

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001.

We actually noticed that many people chose the middle option in the first field experiment after we collected data. We thought it may be because the anchors on our Likert scale did not make the current and alternative assignment processes clear. Thus, to reduce the chance that workers simply picked the middle option, we made a small change to the anchors of the first question in the second experiment. The question itself was almost identical as before and asked workers, "Which assignment process do you think is fairer, algorithmic assignment or human-based assignment?". But this time, we explicitly labeled each anchor with an assignment process, such that 1 = "Algorithmic assignment is obviously fairer", 2 = "Algorithmic assignment is slightly fairer", 3 = "They are equally fair", 4 = "Human-based assignment is slightly fairer", and 5 = "Human-based assignment is obviously fairer". The second question was identical between the first and second field experiments. As shown in Columns 2 and 5 of Table A.2, regardless of whether we use the first or second question, the algorithmic assignment significantly increases fairness perceptions to a similar extent (around one standard deviation) in the second experiment ($p = 0.009$ in Column 2 and $p = 0.0007$ in Column 5).

Further, as shown in Column 3 of Table A.2, if we combine the two field experiments and measure fairness perceptions using the first question, we also observe that the algorithmic assignment significantly improves workers' perceived fairness about their assignment process ($p = 0.007$). Similarly, as shown in Column 6 of Table A.2, if we use the second question to measure fairness perceptions and analyze data from two field experiments combined, the result that algorithmic assignment significantly increases perceived fairness holds ($p < 0.0001$).

In addition, we test the inter-item reliability between the first and second questions. The Pearson correlation coefficient was 0.31 ($p = 0.001$), 0.84 ($p < 0.001$), and 0.53 ($p < 0.001$) in the first field experiment, the second experiment, and two experiments combined, respectively.

In summary, we find that (1) the first and second questions behaved similarly and were highly correlated in the second experiment, (2) the second question behaved similarly between the first and second field experiments, and (3) the first question worked well when two experiments are combined. These observations give us faith in our results about fairness perceptions.

Appendix A.1.5: Results of Interviews

In September 2020, we conducted structured interviews with 13 picking workers (61.54% females, $M_{age} = 30.54$) in the warehouse where our field experiments were implemented. The interviews lasted about 25 minutes on average. At the time of our interviews, hard-copy pick lists were printed and laid out on a table at the distribution station for workers to take. Note that the interviews took place one year after our first field experiment and eight months after our second field experiment. Considering that the workers in our field experiments were temporary workers, the workers in our interviews have a low chance of overlapping with workers in our field experiments. We could not verify this for sure since workers took our interviews anonymously and we could not match them with our field experiment data. In this online appendix, we summarized the key questions in order we asked workers, along with the key insights we gleaned from each question.

We first asked workers, "what factors usually influence your motivation and productivity?" The most frequently mentioned factors, brought up by 7 out of 13 workers, involve pick list characteristics including the number of items they have to collect and how many stocking positions they have to get products from. Another factor, which was brought up by 2 out of 13 workers, is the convenience of obtaining pick lists.

Next, we asked workers, "what factors could influence whether you find a pick list assignment process fair or unfair?" The most frequently mentioned factor, brought up by seven workers, is whether pick lists of varying task difficulty are assigned evenly across workers, especially in terms of how many stocking positions they have to get products from. In addition, some workers (n = 4) focused on the assignment process in the warehouse at the time of our interviews and complained that since pick lists were put on a table for workers to take, some of their colleagues tended to take pick lists according to their own preferences and leave harder pick lists to others, causing unfair task allocations.

Then we asked workers whether they thought the pick list assignment process would be fair or unfair if it was run by a human distributor as well as why they thought one way or the other. Among workers who indicated that a human-based assignment process might cause unfair outcomes (n = 7), most (n = 5) justified their judgment by mentioning that they believed human distributors are subject to personal biases. For example, human distributors could give easier pick lists to workers who they personally know or who they have a good relationship with. Or workers could get difficult pick lists if they refuse to do personal favors for human distributors. We found out later that among workers who indicated that a human-based assignment process would be fair (n = 6), two workers misunderstood our question. Specifically, they thought about human-based assignment as having workers take pick lists printed out by a human (*i.e.*, the same as what was actually going on in their warehouse at the time of our interviews), rather than having a human allocate pick lists (*i.e.*, what we were interested in knowing their thoughts about).

Furthermore, to get some sense about when workers care about fairness, we asked workers to rate how much they would care about the fairness of a pick list assignment process under three circumstances (from 1 = "Not at all" to 5 = "Very much"): their average response was 3.42 if they were paid based on the number of items they picked; 2.38 if they were paid

by hour; and 4.00 if they were paid by their performance ranking among workers in the warehouse.

Next, we asked whether they thought the pick list assignment process would be more or less fair if they could receive pick lists by scanning a bar code than if they could receive pick lists from a human distributor. Most workers (n = 10) believed the assignment process run by a machine would be fairer. When asked why they believed so, most workers (n = 8) explained that they believed an algorithmic assignment process does not follow human distributors' personal preferences, would be able to deliver equal treatments across workers, and would not selectively favor or disadvantage certain workers.

In the end, we asked workers, "besides pick list characteristics and the assignment process, what other factors may influence your productivity?". Common factors brought up by workers included special circumstances (whether certain products are out of stocks, whether picking carts are temporarily unavailable), physical work environment (warehouse temperature, weather), and workers' physical well-being.

Appendix A.6. Additional Results in Heterogeneous Treatment Effects in Our Main Field Experiment

A.6.1. Heterogeneous Treatment Effects Based on Workers' Education Level

In this section, we explore whether workers' education levels affect how they respond to algorithmic assignment. We split our sample by education level: workers whose education level is at or below middle school form a subsample (n = 28), and workers whose education

Table A.3: Heterogeneous Treatment Effect Based on Workers' Education Level

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Standardized perceived fairness			Picking efficiency		
Subsample of workers:	High school or above	Middle school or under	All sample	High school or above	Middle school or under	All sample
Algorithm	1.39***	0.87**	0.72**	1.34****	0.28	0.45
	(0.40)	(0.37)	(0.32)	(0.23)	(0.32)	(0.28)
High school or above			−0.35			0.91**
			(0.59)			(0.40)
*Algorithm*High school or above*			0.64			0.79**
			(0.46)			(0.36)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	No	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Pick list controls	No	No	No	Yes	Yes	Yes
Observations	52	54	106	2,177	2,238	4,415

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 3.84 for workers with a high school degree or above and 4.02 for workers with a middle school degree or under.

is at or above high school form another subsample (n = 22). We separate the sample based on whether a worker achieved a degree higher than middle school because China has the nine-year compulsory education policy: citizens are required to attend school for at least nine years including six years of primary education and three years of middle school, which are all funded by the government.

As shown in Columns 1-2 of Table A.3, for both subsamples of workers, algorithmic assignment process is perceived as fairer than human-based assignment. Specifically, assigning pick lists via an algorithm (vs. via a human distributor) significantly increases workers' fairness perceptions of their current assignment process by 1.39 standard deviations among workers with a high school degree or above (p = 0.001) and by 0.87 standard deviations among workers without a high school degree (p = 0.025). When we add an interaction between the algorithmic treatment and a dummy variable indicating whether workers had a high school degree or above to predict fairness perceptions, the treatment effect of algorithmic assignment on fairness perception is not significantly moderated by education (p = 0.17, Column 3 in Table A.3).

As shown in Columns 4-5 of Table A.3, for workers with a high school degree or above, algorithmic assignment boosts productivity by 34.90%, relative to the average picking efficiency of 3.84 among workers with at least a high school degree in the human group ($p < 0.0001$); however, algorithmic (vs. human-based) assignment has no significant effect on the productivity of workers without a high school degree ($p = 0.37$). Furthermore, as shown in Column 6 of Table A.3, the interaction between the indicator for algorithmic treatment and the indicator for having at least a high school degree is positive and significant ($p = 0.026$). This indicates that algorithmic assignment (vs. human assignment) improves productivity to a significantly larger extent among workers with at least a high school degree than workers with a lower level of education. This result is consistent with prior research showing that people with higher levels of education tend to have higher self-esteem, maintain a more positive evaluation of their own worth [101]; and that people with high (vs. low) self-esteem are more eager to embrace fair treatments and more likely to adjust their attitudes and effort at work based on their fairness perceptions [102]. Future research that systematically tests the role of education and the underlying reasons would be interesting.

A.6.2 Additional Information about Task Difficulty in the Main Field Experiment

To examine the relationship between picking efficiency and the number of stocking positions, we run an ordinary least squares regression to predict picking efficiency as a function of the number of stocking positions. As the number of stocking positions in a pick list increases by one, picking efficiency on average decreases by 0.17 items per minute (or 4.01% relative to the average picking efficiency in the sample; $p < 0.001$). Compared to a regression that only uses time fixed effects and worker demographics to predict picking efficiency, adding the

number of stocking positions as a predictor increases the R^2 of the regression by 0.0516. This suggests that the number of stocking positions can explain 5.16% of the variation in picking efficiency on top of day fixed effects, hour fixed effects, and worker demographics combined.

As mentioned in Section 1.6.3, we calculate the average stocking positions across all pick lists assigned to a given worker on a given day. We then categorize worker-day level observations into three groups by splitting the range between the minimum (2.41) and maximum (15.17) of average stocking positions into three equal intervals. If a worker's average stocking positions on a day were in the interval of [2.41,6.66], (6.66,10.92], or (10.92,15.17], this worker was considered as on average facing low, medium, or high task difficulty that day, respectively. We similarly categorize pick lists into three groups by splitting the range between the minimum (1.00) and maximum (49.00) of the number of stocking positions into three equal intervals. Specifically, if the number of stocking positions in a pick list was in the interval of [1.00,17.00], (17.00,33.00], or (33.00,49.00], we deem this pick list as having a low, medium, or high task difficulty, respectively. As shown in Figure 1.2(b) in the paper and Figure A.5, compared to the distribution of average stocking positions at the worker-day level, the distribution of the number of stocking positions at the pick list level is much more positively skewed. Thus, the cutoffs used to mark the levels of task difficulty are higher at the pick list level than at the worker-day level.

In Figures A.6(a)-A.6(c), the x-axis indicates the percentage of pick lists that are deemed as high difficulty across all pick lists a worker received on a day (*i.e.*, *percentage of high difficulty pick lists*), and each bar represents the number of worker-day level observations within a given subsample whose percentage of high difficulty pick lists fall into a specific range. For example, Figure A.6(a) indicates that among the worker-day subsample with low task difficulty, all of the 52 worker-day level observations have 0-2% of pick lists deemed as high difficulty. That is, no observations have more than 2% of high difficulty pick lists.

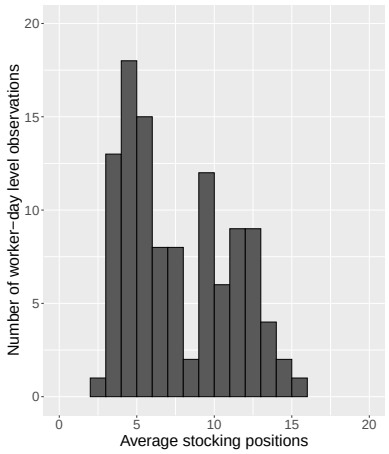
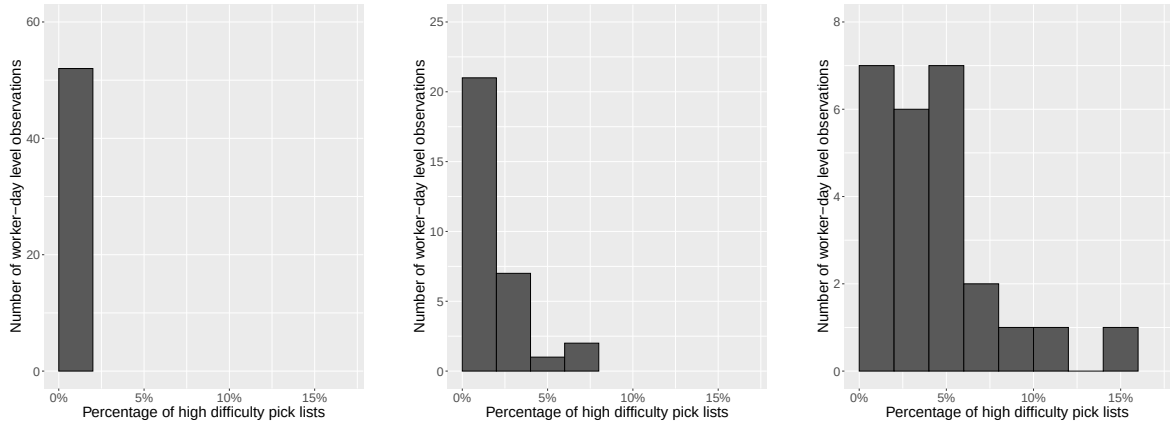


Figure A.5: Distribution of Average Stocking Positions Across Worker-day Level Observations



(a) Worker-day Subsample with Low Task Difficulty

(b) Worker-day Subsample with Medium Task Difficulty

(c) Worker-day Subsample with High Task Difficulty

Figure A.6: Distributions of Percentage of High Difficulty Pick Lists in Three Worker-day Level Subsamples

Figures A.6(a)-A.6(c) show that if a worker-day observation is categorized into the high task difficulty subsample, the worker would be more likely to receive a higher percentage of high difficulty pick lists that day, compared to if the worker-day observation is categorized into the medium or low task difficulty subsample.

A.6.3 Heterogeneous Treatment Effects Based on Task Difficulty—Median Split

In addition to splitting the observations into three subsamples to identify particularly difficult or easy tasks, we have also done a median split to separate our observations into two subsamples. Specifically, we split the worker-day level observations based on whether a given worker's average stocking positions on a given day was greater than the median value of 7.00. As shown in Table A.4, when workers experience below-median task difficulty (*i.e.*, average stocking positions below or equal to 7.00), algorithmic (vs. human-based) assignment increases fairness perceptions by 1.51 standard deviations ($p < 0.001$ in Column 1). When workers experience above-median task difficulty (*i.e.*, average stocking positions above 7.00), algorithmic (vs. human-based) assignment does not significantly increase fairness perceptions ($p = 0.14$ in Column 2). We further split pick lists based on whether the number of stocking positions in a given pick list was greater than the median value of 5.00. When workers handle tasks with below-median difficulty (*i.e.*, the number of stocking positions no more than 5.00), algorithmic assignment significantly boosts productivity by 0.75 items per minute (or 16.48% relative to the average productivity for low difficulty pick lists in the human group; $p = 0.014$ in Column 3). This impact is not statistically significant when workers handle tasks with above-median difficulty (*i.e.*, the number of stocking positions above 5.00; $p = 0.62$ in Column 4).

Table A.4: Effects of Algorithmic (vs. Human-based) Assignment Broken Down by Median Task Difficulty

	(1)	(2)	(3)	(4)
Dependent variable:	Standardized perceived fairness		Picking efficiency	
Subsample:	Below median task difficulty	Above median task difficulty	Below median task difficulty	Above median task difficulty
Algorithm	1.51****	0.83	0.75**	0.05
	(0.32)	(0.55)	(0.31)	(0.10)
Day fixed effects	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes
Pick list controls	No	No	Yes	Yes
Observations	55	51	2,250	2,165

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 4.55 for tasks below median difficulty and 3.30 for tasks above median difficulty.

Altogether, when we split our observations into two subsamples based on the median task difficulty, the effect of algorithmic (vs. human-based) assignment seems weaker in the above-median subsample than in the below-median subsample. However, it is important to note that since task difficulty is positively skewed, especially at the pick list level (as shown in Appendix A.6.2), the above-median subsample includes the majority or all of observations in the range of medium task difficulty, whereas the below-median subsample primarily consists of observations in the range of low task difficulty. This, along with our finding in Section 1.6.3 that the effect of algorithmic assignment tends to emerge when workers are faced with particularly easy or difficult tasks (*i.e.*, in the range of low or high task difficulty), helps explain why the effect of algorithmic assignment seems weaker in the above-median subsample than in the below-median subsample.

A.6.4 Heterogeneous Treatment Effect Based on Sensitivity to Task Difficulty When Task Difficulty Is Above Median

In this section, we examine heterogeneous treatment effect based on sensitivity to task difficulty using only the subset of observations where task difficulty is above the median. As shown in

Table A.5, when faced with above-median task difficulty on a day, high-sensitivity workers perceive algorithmic assignment to be significantly fairer than human-based assignment by 2.34 standard deviations ($p < 0.0001$ in Column 1); but for low-sensitivity workers, there is no significant difference in perceived fairness between the two assignment processes ($p = 0.40$ in Column 2). In fact, the treatment effect of algorithmic assignment on fairness perceptions is significantly amplified among high-sensitivity workers than among low-sensitivity workers ($p = 0.0006$ in Column 3).

Table A.5: Heterogeneous Treatment Effect Based on Sensitivity to Task Difficulty When Task Difficulty Is Above Median

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Standardized perceived fairness			Picking efficiency		
Subsample of workers:	High sensitivity	Low sensitivity	All sample	High sensitivity	Low sensitivity	All sample
Algorithm	2.34****	−0.47	−0.57	0.39**	−0.19	−0.14
	(0.35)	(0.54)	(0.54)	(0.18)	(0.13)	(0.13)
High Sensitivity			−2.65****			0.03
			(0.61)			(0.12)
*Algorithm * High Sensitivity*			2.68****			0.34*
			(0.70)			(0.19)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	No	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Pick list controls	No	No	No	Yes	Yes	Yes
Observations	24	27	51	1,070	1,095	2,165

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. This table includes worker-day or pick list level observations that are associated with above-median task difficulty (*i.e.*, if the average stocking positions for a worker on a day or the number of stocking positions in a pick list was above median). For pick lists whose number of stocking positions was above median, average picking efficiency in the human group was 3.59 among high-sensitivity workers and 3.10 among low-sensitivity workers.

Further, when workers handle pick lists with above-median task difficulty, algorithmic assignment boosts the productivity of high-sensitivity workers by 0.39 items per minute (or 10.86% relative to the average picking efficiency of high-sensitivity workers in the human group; $p = 0.026$ in Column 4), but does not significantly impact the productivity of low-sensitivity workers ($p = 0.13$ in Column 5). The treatment effect of algorithmic assignment

on productivity is marginally significantly amplified among high-sensitivity workers relative to low-sensitivity workers (p = 0.073 in Column 6).

Appendix A.7: Supplement Results of Online Experiments

A.7.1 Additional Results about the Experiment Reported in the Paper

In addition to the measure described in the paper, we also measured the relative importance of equality and uniqueness using one scale. We asked participants which of the two objectives they thought the warehouse should prioritize when it comes to assign picking tasks: treating all workers equality or taking into consideration personal characteristics. The anchors on the scale ranged from 1 ("Definitely should treat all workers equally") to 7 ("Definitely should consider workers' characteristics"). Choosing a higher (vs. lower) value indicates that the participant put less (more) weight on equality (uniqueness). Choosing the midpoint of the scale (*i.e.*, 4) means that the participant thought it equally important to ensure equality and consider workers' unique characteristics. We confirm that people on average prioritize equality over uniqueness in the warehouse task assignment setting ($M = 3.46 < 4$, $SD = 1.85$; $t(200.00) = 26.55$, $p < 0.0001$ for a one-sample t-test).

A.7.2 Results about the Replication Online Experiment

We conducted another online experiment that followed the same design as the online experiment reported in the paper, except that we did not present information about pick list size in this additional experiment. A total of 200 participants from Amazon's Mechanical Turk

comprised our study sample (38.50% female, $M_{age} = 38.925$). They were randomly assigned to either the algorithm condition (N = 99) or the human condition (N = 101).

Supporting the assumption underlying our Hypothesis 1, people view the assignment process run by a machine as more capable of preserving equality than the assignment process run by a human ($M_{algorithm} = 5.21$, $SD = 1.53$ vs. $M_{human} = 4.27$, $SD = 1.68$; $t(196.91) = 4.17$, $p < 0.0001$, Cohen's $d = 0.59$). Following the analysis in our field experiment and online experiment reported in the paper, we constructed *Standardized Perceived Fairness*, which equaled *Perceived Fairness* divided by its standard deviation in the whole sample. In support of Hypothesis 1, participants in the algorithm condition perceived their assignment process fairer than those in the human condition ($M_{algorithm} = 3.56$, $SD = 0.87$ vs. $M_{human} = 2.98$, $SD = 1.04$; $t(193.23) = 4.32$, $p < 0.0001$, Cohen's $d = 0.61$).

Appendix A.8: Heterogeneous Treatment Effects by Gender

In this appendix, we report findings about the heterogeneous treatment effects of algorithmic assignment based on gender in both our field and online experiments.

In terms of fairness perceptions, as shown in Columns 1-2 of Table A.6, algorithmic assignment (vs. human-based assignment) significantly improves fairness perceptions for both female workers and male workers during our main field experiment (all p-values ≤ 0.002). The effect size seems directionally larger among females than among males. When we predict fairness perceptions as a function of gender, the algorithmic treatment, and their interaction, gender does not significantly moderate the effect of algorithmic assignment on fairness perceptions: as shown in Column 3 of Table A.6, p-value for the interaction term is 0.11. As for the online experiment reported in the paper, Columns 1-2 of Table A.7 indicate that algorithmic (vs. human-based assignment) significantly increases fairness perceptions among females (p

= 0.0012 in Column 1) but does not significantly change fairness perceptions among males (p = 0.13 in Column 2). The impact of algorithmic assignment on fairness perceptions is directionally but not significantly larger among females than among males, as shown in Column 3 of Table A.7 (p = 0.18).

In terms of productivity, as shown in Columns 4-5 of Table A.6, algorithmic assignment has a positive but not statistically significant effect among female workers (p = 0.12 in Column 4), and has a significant, positive impact among male workers (p < 0.0001 in Column 5) in our main field experiment. Notably, the estimated effect size is similar in magnitude between female workers and male workers, and the non-significant effect among females may be related to sample size: During our main field experiment, male workers on average came to work in the warehouse for 2.59 days, while female workers only worked for 1.57 days; and female workers in total contributed only 28.64% of all pick lists. Gender does not significantly moderate the effect of algorithmic assignment on productivity (p = 0.76 in Column 6 of Table A.6).

Table A.6: Effects of Algorithmic (vs. Human-based) Assignment Broken Down by Gender (Main Field Experiment)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	Standardized perceived fairness			Picking efficiency		
Subsample of workers:	Female	Male	All sample	Female	Male	All sample
Algorithm	1.66***	0.87***	0.85****	0.82	0.88*****	0.57**
	(0.45)	(0.27)	(0.23)	(0.52)	(0.22)	(0.22)
Female			−0.52*			−0.50*
			(0.31)			(0.26)
*Algorithm*Female*			0.53			0.12
			(0.33)			(0.39)
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Hour fixed effects	No	No	No	Yes	Yes	Yes
Demographics controls	Yes	Yes	Yes	Yes	Yes	Yes
Pick list controls	No	No	No	Yes	Yes	Yes
Observations	31	75	106	1,214	3,201	4,415

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001. Average picking efficiency in the human group was 3.94 for female workers and 3.92 for male workers.

Table A.7: Effects of Algorithmic (vs. Human-based) Assignment Broken Down by Gender (Online Experiment)

	(1)	(2)	(3)
Dependent variable	*Standardized perceived fairness*		
Subsample of workers:	Female	Male	All sample
Algorithm	0.67***	0.29	0.29
	(0.20)	(0.19)	(0.18)
Female			0.003
			(0.20)
*Algorithm*Female*			0.38
			(0.28)
Observations	83	116	199

Note: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001.

Appendix B

Appendix for Chapter 2

Appendix B.1. Sample Robustness

Table B.1: Impact of Pick-up stations–Sample Robustness on City

Dependent variable:	*GMV*	*Orders*
	(1)	(2)
After	−62.181	−0.406
	(102.478)	(0.538)
Station	1,384.067***	10.161***
	(50.282)	(0.264)
*Station*After*	477.985***	1.602***
	(70.902)	(0.372)
Relative Effect Size	10.5%	4.9%
Time fixed effects	Yes	Yes
Area fixed effects	Yes	Yes
Observations	1,097,100	1,097,100

Note: *p<0.05; **p<0.01; ***p<0.001. The sample is based on another city (Hangzhou) during the same period, the average of *GMV* was 4543.372, the average of *Orders* was 32.884.

Appendix B.2: Parallel Trend in Pre-treatment Period

To examine the assumption that the parallel trends assumption is satisfied, following [74], we run the following analogue of our main specification in the pre-treatment period:

$$Outcome\ Variable_{it} = \beta_1 * Days_{it} + \beta_2 Station_i + \beta_3 Station_i * Days_{it} + \mu_i + \lambda_t + \varepsilon_{it},$$

where $Days_t \in \{-1, -2, ..., -90\}$, and the absolute value of $Days_t$ is the number of days before the treatment of location i on day t. From Columns 1-4 in Table B.2, the estimates of β_3 are not significant (p-values > 0.204). This result is consistent with the parallel trends assumption.

Table B.2: Parallel Trend Test

	Dependent variable:			
	GMV	*Items value*	*Orders*	*New comers*
	(1)	(2)	(3)	(4)
Days	22.773***	−0.168*	0.274***	0.003***
	(1.307)	(0.068)	(0.009)	(0.001)
Station	251.220***	−1.139	2.641***	0.022
	(28.354)	(1.458)	(0.199)	(0.011)
*Station*Days*	0.673	0.019	0.005	<0.0001
	(0.540)	(0.028)	(0.004)	(0.0002)
Relative Effect Size	0.03%	0.02%	0.02%	<0.01%
Time fixed effects	Yes	Yes	Yes	Yes
Area fixed effects	Yes	Yes	Yes	Yes
Observations	598,950	479,990	598,950	598,950

Note: *p<0.05; **p<0.01; ***p<0.001.
Note: Average of *GMV* was 2310.026, average of *Items value* was 106.171, average of *Orders* was 21.506, average of *New comers* was 0.202.

Appendix B.3: Estimation Results With Multiple Classes

Table B.3: Estimation Results—Multiple Latent Classes

Number of latent classes	1	2		3			4			
(Consumer Type)	(1)	(1)	(2)	(1)	(2)	(3)	(1)	(2)	(3)	(4)
Daytime receiving value	21.800	1.739	10.612	2.291	0.618	10.218	3.060	0.934	0.639	10.167
	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)	(0.003)	(0.009)	(0.007)	(0.015)	(0.015)
After-work receiving value	47.851	3.902	23.134	4.916	11.536	22.122	3.935	12.123	12.204	21.828
	(0.013)	(0.002)	(0.004)	(0.003)	(0.002)	(0.003)	(0.019)	(0.020)	(0.029)	(0.020)
Scale parameter (uncertainty)	37.711	4.473	17.718	5.361	14.802	16.642	5.514	14.424	15.094	16.596
	(<0.001)	(0.001)	(<0.001)	(0.001)	(<0.001)	(<0.001)	(0.008)	(0.004)	(0.003)	(0.003)
Package waiting sensitivity	−2.303	−0.248	−1.088	−0.308	−0.939	−1.032	-0.304	-0.917	-0.978	-0.999
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Station distance sensitivity	−0.301	−2.577	−0.092	−2.596	−0.991	−0.106	-2.073	-3.749	-0.192	-0.042
	(0.001)	(0.006)	(0.002)	(0.009)	(0.002)	(0.001)	(0.011)	(0.013)	(0.001)	(0.002)
Home constant	-18.938	-3.268	-11.050	−2.422	−1.968	−10.254	-2.681	-2.251	-2.797	-9.912
	(0.003)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.008)	(0.004)	(0.009)	(0.007)
Station constant	-115.153	-17.370	-54.763	−18.996	−40.914	−51.081	-19.856	-40.659	-39.479	-51.202
	(0.006)	(0.004)	(0.002)	(0.005)	(0.001)	(0.002)	(0.026)	(0.014)	(0.008)	(0.009)
Type probability	1.000	0.813	0.187	0.184	0.711	0.105	0.163	0.635	0.146	0.055
	–	–	(<0.001)	(<0.001)	–	(<0.001)	(<0.001)	(<0.001)	(<0.001)	–
BIC	29281518.755	28032264.547		26795602.199			26614183.994			
Observations	893,126	893,126		893,126			893,126			

Note: Standard errors are presented in parentheses. Daytime receiving is between 10:00 a.m.-3:59 p.m., and after-work receiving is between 4:00 p.m.-7:59 p.m..

Appendix B.4: Gaussian Mixture Model on the Distribution of Delivery Time

Figure B.1: Distribution of Gaussian Mixture Model