Learning on the Go: Understanding How Gig Economy Workers Learn with Recommendation Algorithms

SHUNAN JIANG, Department of Industrial Engineering and Operations Research, UC Berkeley, USA
WICHINPONG PARK SINCHAISRI, Haas School of Business, UC Berkeley, USA

As gig economy platforms increasingly rely on algorithms to manage workers, understanding how algorithmic recommendations influence worker behavior is critical for optimizing platform design and improving worker welfare. In this paper, we investigate the dynamic interactions between gig workers and platform algorithms, with a particular focus on how workers learn to improve their strategy and performance over time. Using multiple quantitative methods, including two-way fixed-effects regression and multinomial logit modeling, we analyze over one million orders completed by gig workers on a retail delivery platform. Our findings reveal a clear learning curve: workers progressively improving their efficiency and on-time delivery performance with increased experience. We also find that while newcomers heavily rely on algorithmic recommendations for task selection, more experienced workers tend to deviate from these recommendations, developing and employing personalized strategies. This shift suggests that experienced workers may perceive algorithmic recommendations as less beneficial or misaligned with their evolved preferences, highlighting the necessity for adaptive recommendation systems. Our research underscores the importance of designing human-centric recommendation algorithms that accommodate workers' learning trajectories, incorporate their feedback, and offer flexibility to support personalized strategies, ultimately enhancing collaborative dynamics and outcomes for both workers and platforms.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: gig economy, worker performance, worker learning, task bundling, recommendation algorithms, empirical analysis

ACM Reference Format:
Shunan Jiang and Wichinpong Park Sinchaisri. 2025. Learning on the Go: Understanding How Gig Economy Workers Learn with Recommendation Algorithms. In *Proceedings of ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '25).* ACM, New York, NY, USA, 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As technology-mediated work continues to reshape the labor landscape, gig economy platforms, such as grocery delivery, ride-hailing, and freelancing, have become crucial sources of flexible, task-based employment. These platforms offer workers independence in task selection but also present them with complex decision-making challenges, especially as the volume and diversity of tasks grow. Without the traditional support networks of coworkers, supervisors, or mentors, gig workers must independently navigate these challenges, often learning through trial and error [40].

Authors' Contact Information: Shunan Jiang, Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, California, USA, shunan_jiang@berkeley.edu; Wichinpong Park Sinchaisri, Haas School of Business, UC Berkeley, Berkeley, California, USA, parksinchaisri@berkeley.edu.

CSCW '25, Oct 18–22, 2025, Bergen, Norway

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

To enhance operational efficiency, many platforms now rely on algorithmic recommendation systems to support workers' decision-making. In grocery delivery, for example, platforms frequently suggest bundling multiple orders into a single trip to streamline routes, reduce idle time, and boost earnings. However, these algorithmic recommendations introduce new layers of complexity, requiring workers to integrate automated suggestions with their own personal strategies. Navigating this balance can be difficult, leading to misaligned decisions that may reduce performance or result in suboptimal outcomes. For example, workers on food delivery platforms often rejected bundled orders due to difficulties associated with batch tasks, such as increased complexity and effort required [19].

When workers have the autonomy to create their own task bundles, they may overestimate their capacity or fail to account for the logistical challenges of complex deliveries, resulting in delayed orders, missed service windows, or customer dissatisfaction. The cognitive load imposed by the vast number of available tasks further complicates decision-making, forcing workers to juggle multiple, often competing, priorities. These priorities include maximizing earnings, minimizing effort, and meeting performance benchmarks imposed by the platform.

In this paper, we investigate how gig economy workers learn and adapt through interactions with platform recommendation algorithms in a real-world retail delivery setting. Specifically, we address three research questions: (i) How do gig workers learn to enhance their performance over time? (ii) How do gig workers respond to the platforms' bundling and task recommendation algorithms? (iii) How does workers' decision-making evolve as they accumulate experience with the platform? Our findings offer insights into designing recommendation systems that better support worker autonomy and performance, ultimately facilitating more effective collaboration between platform algorithms and worker strategies.

We adopt multiple quantitative methods to investigate how gig workers engage in learning and decision-making on an on-demand retail delivery platform. Using a dataset of 1.2 million orders completed by 5,292 gig workers over 364 days in New York City, we apply a two-way fixed-effects regression model, controlling for external variables such as weather conditions and traffic patterns, to assess how workers learn to perform better through experience. We then perform a descriptive analysis to uncover how workers learn to bundle tasks with platform recommendation and how the interaction influences their performances. Additionally, to analyze workers' task selection behaviors, we employ a multinomial logit (MNL) model that captures how workers respond to platform recommendations and explore new stores as they accumulate experience. This methodological framework offers a comprehensive view of worker strategies, shedding light on how workers co-adapt with platform algorithms to optimize performance over time.

Our findings reveal several key insights into worker adaptation to algorithmic systems. First, workers demonstrate a clear learning curve, with significant improvements in efficiency and on-time delivery as they gain experience. The regression analysis shows that store-specific experience plays a crucial role in enhancing performance, while skills acquired from other stores also contribute to improvement. These transferable skills, such as navigating store layouts and managing customer expectations, enable workers to adapt more efficiently across different contexts, highlighting the importance of cross-context learning.

When choosing which orders to accept, workers can decide whether to follow platform recommendations or select from a pool of non-recommended orders. Our findings indicate that newer workers are more inclined to rely on platform suggestions, while more experienced workers develop their own strategies, increasingly deviating from algorithmic recommendations. This progression highlights how workers gradually build confidence and optimize their task selection strategies, ultimately improving both performance and earnings. Our results suggest that as workers gain proficiency, platforms should adapt their algorithms to offer greater flexibility, enabling experienced

workers to align task selection with their evolving strategies. Additionally, incorporating worker feedback mechanisms can further personalize recommendations, ensuring suggestions remain relevant and responsive to workers' changing needs.

While prior research on gig worker learning often uses behavioral proxies such as the rate of visiting new areas to highlight phases of exploration, particularly among newer workers [9], the strategic balance between exploration and exploitation at the moment of decision remains less well understood. Our paper offers a new perspective by applying a multinomial logit (MNL) choice model that explicitly accounts for the set of alternatives available to workers when making task selections. Viewed through this lens, we find a surprisingly strong and consistent preference for familiar options (i.e., exploitation), evident across experience levels and especially pronounced among top-performing workers. Understanding this pattern within discrete choice contexts is critical for designing platform mechanisms that better support real-world decision-making.

These findings highlight the value of human-centric recommendation systems that evolve in tandem with workers' learning trajectories and preferences. By aligning algorithmic recommendations with workers' strategies and experience levels, platforms can improve collaborative interactions, enhance performance outcomes, and foster long-term engagement within the gig economy. This adaptive approach can empower workers while ensuring platform systems remain both efficient and supportive of worker autonomy and development.

Our paper is organized as follows: Section 2 reviews related work and outlines our contributions. Section 3 introduces the context of our study and describes the dataset. In Section 4, we provide empirical evidence on how workers learn to improve performance through experience. Section 5 explores how workers adapt to the platform's recommendations for bundling tasks. Section 6 examines how workers learn to select tasks with the platform's recommendation algorithms and discusses implications for recommendation algorithm design. Finally, Section 8 presents our concluding remarks.

2 Related Works and Contributions

Our work relates to two major streams of literature: the interactions between humans and algorithms or computer-supported platforms, and worker learning and performance improvement in operations management.

2.1 Human-Algorithm Interactions at Work

This paper contributes to a growing body of work examining how digital platforms shape worker behavior and performance through algorithmic management and recommendation systems. A central theme in this literature concerns the tradeoffs between worker autonomy, engagement, and the structure imposed by algorithmic management.

Several studies have highlighted how platform features and feedback channels shape performance and satisfaction. Higher-quality platforms have been associated with greater worker autonomy and job satisfaction [26]. The introduction of dedicated communication spaces between gig workers and restaurants has been shown to facilitate cooperation on food delivery platforms [38], while structured feedback systems can improve outcomes among crowd workers by guiding their attention and effort [17]. Customizable and evolving avatars have also been explored as tools for boosting worker engagement [13], and recent work emphasizes how design features that prioritize well-being can strengthen worker-platform relationships [37].

A growing stream of research investigates how algorithmic and AI systems influence worker autonomy and learning. Algorithmic management has been shown to reshape power dynamics and compel workers to develop new interpretive skills for navigating data-driven systems [23, 24]. Perceptions of fairness, trust, and emotional response play a critical role in determining how

workers engage with algorithmic decisions [28]. More broadly, platform data-driven systems have been shown to influence worker autonomy and job satisfaction [29].

Several recent studies propose new directions for worker-facing AI tools. Stakeholder-centered design approaches have been used to co-create tools that align algorithmic management with worker needs [41, 42]. AI-guided systems can improve service quality, particularly for novice gig workers, though they may also increase task completion times due to added overhead from AI consultations [27]. Other research focuses on resistance when algorithmic tools are seen as intrusive. Passive sensing systems, despite potential benefits, may be rejected due to concerns over surveillance and control [11, 12]. In contrast, tools that give workers autonomy to track their own performance have been proposed to increase transparency and accountability [16]. Participatory design and collective action strategies have also been proposed as ways to create more empowering and worker-centric platform environments [32, 35].

At the same time, algorithmic recommendations are not always embraced. Users may exhibit algorithm aversion, forming biased perceptions against algorithmic advice [14, 15], or fail to incorporate such recommendations effectively into their workflows, even when open to them [5].

Our work also intersects with core challenges within the broader field of recommender systems (RecSys), particularly the Exploration-Exploitation (E&E) dilemma and the Cold-Start problem [4, 43]. The E&E dilemma involves the fundamental trade-off between exploiting known user preferences to maximize immediate satisfaction and exploring new or uncertain items to gather information and improve future recommendations [4, 43]. This balance is often modeled using Multi-Armed Bandit (MAB) or Reinforcement Learning (RL) frameworks [25, 30, 31] and is crucial for long-term engagement and discovery [4, 43].

Our study builds on this literature by examining how gig workers engage with task and bundling recommendations over time. While prior work has focused largely on the design and short-term impact of algorithmic tools, less is known about how workers develop long-term strategies and adapt as they accumulate experience. We explore how workers initially respond to recommendations, how their behavior evolves over time, and how this adaptation affects their long-term performance. In doing so, our findings contribute to a deeper understanding of human-algorithm collaboration in gig economy settings.

2.2 Worker Learning and Performance Improvement

Worker learning has long been a foundational topic in operations and organizational research. Comprehensive reviews have described how individuals and teams improve over time, with experience-based learning often cited as the primary mechanism [3, 10]. Empirical studies document learning curves in settings ranging from software development [18] and assembly lines [36] to item-picking [20] and emergency response services [6]. Learning can also occur through peer interactions [1] and customer feedback [8]. Reinforcement learning models, such as the experience-weighted attraction model, offer theoretical foundations for understanding how workers update strategies over time in response to performance feedback [7].

As the gig economy has grown, scholars have turned their attention to learning dynamics in these more flexible, data-driven work environments. Gig workers are often influenced by internal targets such as income and time goals in addition to pay rates [2], and their day-to-day experiences help shape productivity and service quality [21]. Research on early-stage gig work behavior shows that workers tend to explore unfamiliar regions at first, which may reduce short-term performance but enables longer-term gains as they learn to batch more effectively and improve delivery quality [9]. Workers also engage in self-tracking practices to maintain personal accountability and reflect on past outcomes [22].

To better support gig workers' task selection strategies, recent studies have analyzed the heuristics used to accept or reject batched orders. These efforts have led to order batching algorithms that better accommodate courier needs [19]. In parallel, emotional labor and perceptions of control have been linked to job satisfaction, further highlighting the complex motivational landscape of gig work [33].

Our research extends this body of work by focusing on how gig workers learn and refine decision-making strategies through repeated interactions with platform recommendation systems. In particular, we study how heterogeneous strategies emerge and result in divergent learning paths. Using a large-scale dataset and a multinomial logit (MNL) framework, we analyze how workers respond to platform recommendations, how their task selection behavior evolves, and how that evolution impacts long-term performance.

3 Empirical Context: Retail Delivery Platform in the U.S.

We collaborate with an on-demand retail delivery company (hereafter referred to as "the company" or "the platform") to analyze a comprehensive dataset consisting of online retail orders completed in New York City over a 364-day period, spanning from November 2022 to October 2023. This dataset captures a wide range of information, including completed orders by workers, order characteristics, and productivity metrics such as time spent shopping, checkout time, and driving time. Additionally, the dataset provides detailed evaluations of each completed order, such as whether the delivery was on time.

One of the key advantages of this dataset is its granularity, which allows us to observe: (1) the full list of orders algorithmically highlighted as recommendations, alongside a separate list of other accessible but non-highlighted orders, meaning those available to the worker but not explicitly promoted by the platform, during the one-hour window immediately preceding each accepted order; and (2) detailed information about orders that were bundled together by the platform for simultaneous delivery.

In the following sections, we first provide an overview of the platform's operations and describe the interface through which workers interact with the system. We then present descriptive statistics on worker activity and the order recommendations they received. Finally, we outline the supplementary datasets used in our analysis.

3.1 Platform Overview

The company operates as an online retail delivery platform that provides on-demand delivery of retail and essential goods across multiple metropolitan areas in the United States. Customers place orders through the platform's mobile application or website and can schedule deliveries within flexible time windows. The platform facilitates timely delivery by matching customers with gig workers who are responsible for visiting physical retail stores, hand-picking the ordered items, and maintaining real-time communication with customers via integrated chat. After shopping, gig workers then deliver the items directly to the customers' addresses.

Gig workers are compensated on a per-order basis, with pay depending on factors such as order size, complexity, and delivery distance. In addition to their base earnings, workers can receive customer tips and platform-issued bonuses for meeting specific performance thresholds, such as delivering during high-demand hours.

3.2 Worker Process and Platform Algorithmic Features

To participate on the platform, workers must first undergo a screening process that includes verification of eligibility criteria such as being over 18 years of age, possession of a valid driver's license, and access to a vehicle. Upon approval, workers are officially granted access to the gig

platform and can define their working hours and preferred delivery regions each day, typically within the platform's operating window of 7:00 AM to 12:00 AM.

During their self-declared working hours, gig workers are shown a comprehensive list of all available delivery orders in their selected region. Within this list, a subset of orders is algorithmically marked as recommended based on factors such as the worker's current location, historical performance metrics, availability, and prior customer ratings. These recommendations are visually tagged to indicate that they may offer better alignment with the worker's profile or potential efficiency gains. However, workers retain full autonomy to choose from the entire pool of available orders, not just those marked as recommended. Each order, whether recommended or not, includes key information such as estimated pay, item composition, store and customer locations, and delivery time windows.

A key distinction between this platform and traditional ride-hailing services (e.g., Uber or Lyft) lies in the task allocation process. While ride-hailing drivers are typically auto-assigned rides and cannot browse or select among alternatives, workers on this platform are given the full flexibility to evaluate a menu of available tasks and make informed decisions based on personal preferences, operational constraints, and expected earnings. This design gives rise to strategic behavior in task selection and opens up opportunities for personalized optimization.

The platform also supports order bundling, which allows workers to fulfill multiple orders in a single shopping and delivery trip. Bundling takes two primary forms. First, the platform algorithm occasionally generates pre-bundled orders by pairing two deliveries that share similar attributes, such as store origin, item composition, and destination proximity. If a worker accepts a platform-generated bundle, they are required to fulfill both orders together. These system-generated bundles always consist of exactly two orders.

Second, workers may choose to *self-bundle* by manually selecting and sequencing multiple orders, whether they are recommended or not, for concurrent fulfillment. There is no platform-imposed limit on the number of orders that can be self-bundled, and workers exercise full discretion in determining whether and how to do so.

This paper focuses on how workers navigate this hybrid environment, where algorithmic guidance is combined with high levels of worker autonomy. We analyze how workers respond to platform recommendations and how they develop bundling strategies over time, either by accepting platform-generated bundles or by constructing self-bundles. These decisions reveal how workers learn and adapt in algorithmically mediated labor settings and offer insights into the evolving dynamics of human-algorithm collaboration.

3.3 Descriptive Statistics

Having described the platform's operational model and algorithmic features, we now present descriptive statistics to contextualize the scale and heterogeneity of worker behavior in our dataset.

The dataset comprises detailed operational records from 5,292 gig workers who collectively completed approximately 1.2 million orders across 800 retail stores. There is substantial heterogeneity in worker engagement: some individuals completed only a single order before dropping out from the platform, while others fulfilled over 6,000 orders within a one-year period. On average, each worker completed 230 orders annually, with the 25th, 50th (median), and 75th percentiles at 5, 28, and 136 orders, respectively.

Order volume at the store level also exhibits considerable dispersion. While some stores processed only one order during the observation window, others handled more than 100,000. On average, each store processed 1,600 orders annually, with the 25th, 50th, and 75th percentiles at 5, 13, and 39 orders, respectively.

Bundling represents a key operational feature of the platform: approximately 60% of all orders are delivered as part of a bundle. The platform employs a proprietary algorithm to construct these bundles by identifying orders with similar attributes, including store origin, item composition, and delivery destination. Each bundle consists of two orders assigned for simultaneous fulfillment by a single gig worker.

3.4 Supplementary Data: TLC Trip Records and Weather Records

To account for the potential influence of traffic and weather conditions on workers' behaviors, we incorporate two additional datasets into our analysis.

The first dataset is the New York City Taxi and Limousine Commission (TLC) dataset, which provides detailed trip-level records for taxi and ride-hailing services in New York City (NYC). This dataset includes information such as pickup and drop-off locations, timestamps, trip distances, fares, and payment methods, encompassing millions of rides over multiple years. From the TLC dataset, we derive two key traffic-related proxies: the traffic volume for each hour and the average hourly speed of taxis, both serving as indicators of overall traffic conditions in NYC.

The second dataset is sourced from the OpenWeather platform, which offers global meteorological data across a broad range of parameters, including temperature, humidity, wind speed, and precipitation, as well as specialized metrics like air pollution and UV index. We initially extracted over 50 weather variables from this platform. After performing variance inflation factor (VIF) testing to address multicollinearity, we selected three weather parameters, apparent temperature, rainfall, and wind speed, for inclusion in our subsequent regression analyses.

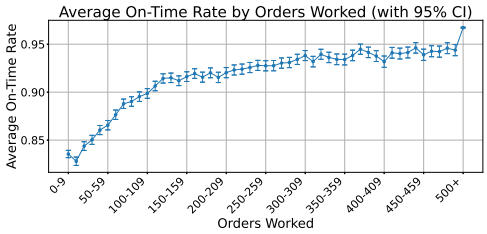
4 Learning to Improve: *How Do Gig workers Learn to Improve Performance?*

In this section, we analyze how gig workers learn to improve their performance over time as they gain experience on the platform. We focus on two key performance outcomes. First, we consider the *on-time probability (OTP)*, defined as the proportion of orders delivered no later than the time specified by the platform. This measure serves as the platform's primary indicator of service quality and reflects a binary classification of each order as on-time or not. Second, we examine the *number of items picked per hour*, a proxy for worker productivity that captures task efficiency. Together, these two metrics provide a comprehensive view of service reliability and operational efficiency of gig workers. We begin with model-free descriptive evidence to document performance trends, followed by an empirical strategy using two-way fixed effects (2FE) regression models that control for time-invariant differences across workers and stores. These analyses form the empirical foundation for the next section, where we examine how workers learn to adapt strategically to the platform's recommendation algorithms.

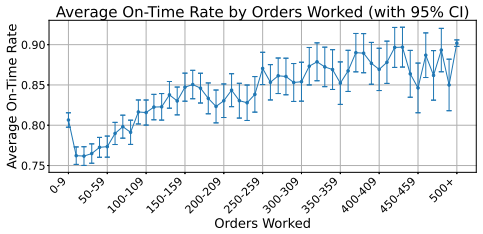
4.1 Model-free Evidence of Performance Improvement

To visualize how performance evolves with experience, we plot the relationship between on-time delivery probability and the cumulative number of orders completed by a worker during the one-year observation period. Figure 1 presents this relationship for two worker populations. Panel (a) includes all workers active during the study period ($N = 5,292$), while Pnael (b) focuses on a cohort of newcomers ($N = 1,131$) who joined the platform after November 1, 2022.

Comparing these two populations reveals important differences. The trend for all workers in Panel (a) shows a relatively high initial on-time rate and a gradual upward trajectory, whereas newcomers in Panel (b) begin with lower average performance and display a steeper early learning curve. This difference is expected, as the full-sample trend averages over workers at various tenure levels, including many who were already experienced and performing well at the start of the study. In contrast, the newcomer cohort more clearly isolates the early-stage learning process from



(a) All Workers



(a) New Workers

Fig. 1. Average On-Time Delivery Probability by Cumulative Orders Completed

Notes: This figure compares average on-time delivery probability as a function of cumulative experience, measured by the number of orders completed. Panel (a) shows trends for all workers active during the study period ($N = 5,292$), while Panel (b) focuses on newcomers who joined after November 1, 2022 ($N = 1,131$). The x-axis represents cumulative orders completed, grouped into bins of 10 (e.g., 0–9, 10–19), up to 500 or more. The y-axis displays the average on-time delivery probability for orders completed within each experience bin. Error bars indicate 95% confidence intervals. The comparison illustrates the aggregate smoothing effect in the full sample and the steeper learning trajectory visible among newcomers.

platform entry. Moreover, the full-sample trend may also be shaped by survivorship bias, since low-performing workers may have exited the platform earlier and are thus underrepresented at higher experience levels.

To mitigate these confounds and more directly examine the learning process, we focus our regression analysis in Section 4 on the newcomer cohort. Because these workers joined during the observation window, we can observe their full experience trajectory from initiation, enabling more reliable identification of learning patterns.

Within the newcomer cohort (Figure 1b), we observe lower average performance relative to the full sample, particularly in the early stages. Notably, there is a small dip in on-time delivery probability between the first (0–9 orders) and second (10–19 orders) experience bins. This initial fluctuation may reflect several dynamics: for instance, some early departures in the 0–9 bin may be experienced workers briefly testing the platform, artificially raising the average performance in that bin. Alternatively, workers may face adaptation challenges or increased task complexity, such as exposure to order bundling, shortly after joining. Beyond this point, performance steadily improves through approximately the first 350 orders. Afterward, the upward trend slows and becomes noisier, with no clear sustained gains. Overall, this descriptive evidence suggests a positive association between cumulative experience and service quality, particularly during early tenure.

4.2 Two-way Fixed-Effects Regression Analysis of Worker Performance

To investigate the relationship between gig worker experience and performance, we perform a two-way fixed effects (2FE) regression analysis for each of the two performance metrics: on-time delivery status ($OnTime$) and the number of items picked per hour ($ItemsPerHour$) [39].

$$\begin{aligned} PerformanceMetric_{ist} = \beta_0 &+ \beta_1 OTS_{ist} + \beta_2 OTS_{ist}^2 \\ &+ \beta_3 OOS_{ist} + \beta_4 OOS_{ist}^2 \\ &+ \mathbf{X}'_{ist}\beta + \mu_i + \delta_s + \gamma_t + \epsilon_{ist} \end{aligned} \tag{1}$$

where

- $PerformanceMetric_{ist}$ is either the on-time delivery outcome or the number of items picked by gig worker i when shopping at store s at time t.
- OTS_{ist} denotes the number of prior orders completed by worker i at store s by time t (within-store experience), while OOS_{ist} denotes the number of prior orders completed by worker i at all other stores (cross-store experience).
- OTS^2_{ist} and OOS^2_{ist} are the squared terms included to capture potential nonlinearities in experience effects.
- $\mathbf{X}_{ist}$ is a vector of control variables, including external factors such as weather conditions (e.g., temperature, rain, wind speed), order-specific variables (e.g., total payment, bonuses, requested item quantities, delivery distance), and urban traffic metrics (e.g., taxi volume, average traffic speed).
- γ_t denotes time fixed effects, capturing any temporal patterns such as day-of-week or seasonal variations that might influence performance.
- ϵ_{ist} is the idiosyncratic error term, assumed to be independently and identically distributed across i, s, and t.

4.2.1 Description of Key Variables.

Dependent variables. The first dependent variable, *OnTime*, is a binary indicator equal to 1 if the delivery was completed on or before the platform-specified deadline, and 0 otherwise. This variable aligns with how the platform evaluates service quality and is not subject to the noise or skew commonly found in continuous delivery time metrics. The second dependent variable, *ItemsPerHour*, is computed as the number of items picked during the shopping process divided by the time spent in-store. This metric serves as a proxy for worker productivity.

Independent variables. To examine the relationship between a gig worker's accumulated experience with specific stores and their delivery performance, we introduce two key independent variables: *OTS* (*Orders This Store*) and *OOS* (*Orders Other Stores*).

The variable *OTS* measures the number of prior deliveries a worker has completed at a particular store. It serves as a proxy for the worker's familiarity with that store's unique operational environment, including its layout, inventory system, and staff routines. We hypothesize that repeated exposure to the same store enables gig workers to improve their performance over time. For instance, workers may become more efficient at locating items, make fewer picking errors, and develop better rapport with store employees. Familiarity may also lead to improved route planning, both within the store during item collection and externally during the delivery phase. To capture potential nonlinearities in this relationship, we include the squared term OTS^2. This allows us to test whether the marginal benefits of store-specific experience diminish after a certain threshold or continue to accrue.

In contrast, *OOS* measures the number of deliveries completed by a worker at all other stores, excluding the focal one. This variable enables us to investigate whether broader experience across diverse store environments translates to improved performance in a given store. Such cross-store learning may reflect general improvements in workflow, adaptability, or task management that are not store-specific. As with *OTS*, we include a squared term OOS^2 to allow for nonlinear effects in the relationship between broader experience and performance.

Control variables. We include a comprehensive set of controls to account for factors that may confound the relationship between experience and performance. These include:

- *Order characteristics:* total payout, bonuses, item quantities, store-to-customer distance, and delivery time window length, which collectively reflect order complexity and incentives.
- *Traffic conditions*: hourly taxi volume and average taxi speed in New York City, capturing time-varying urban congestion that could delay deliveries.
- *Weather conditions*: apparent temperature, precipitation, and wind speed, which may influence travel time and worker comfort.
- *Time fixed effects* are included at the day-of-week and calendar month levels to control for temporal fluctuations in demand, traffic, and worker availability.
- *Worker–store fixed effects* capture persistent heterogeneity in individual performance that is specific to a worker-store pair, such as skill, local knowledge, or route familiarity.
- We also include total order value as part of the control variables to account for the possibility that lower-priced orders may include a higher number of small, low-value items, which could inflate the $ItemsPerHour$ metric.

4.3 Results: *Diminishing Positive Return on Experience*

Table 1 presents the estimated associations between gig worker experience and two performance outcomes: on-time delivery probability ($OnTime$) and the number of items picked per hour ($ItemsPerHour$). Each specification includes worker and store fixed effects, as well as controls for weather, traffic, and order characteristics.

Table 1. The impact of experience on performance among new gig workers

	$OnTime$	$ItemsPerHour$
OTS	6.0552e-05***	9.4281e-03***
	(0.003)	((0.008))
OTS^2	-8.9995e-09***	-5.4681e-06**
	(0.000)	(0.003)
OOS	5.9109e-05***	6.3414e-03*
	(0.000)	(0.022))
OOS^2	-8.8744e-09***	-7.6253e-07
	(0.015)	(0.235)
Fixed Effects controls	✓	✓
Weather controls	✓	✓
Traffic controls	✓	✓
R^2	0.029	0.013
Observations	105543	105543

We find consistent evidence that both store-specific and cross-store experience are positively associated with performance improvements. In particular, store-specific experience (OTS) is strongly and positively related to both outcomes. The estimated coefficients on OTS^2 are negative and statistically significant, suggesting diminishing marginal returns to store-specific experience. These results support the hypothesis that familiarity with a store's layout, inventory systems, and routines leads to greater efficiency, but that the incremental benefit of additional experience declines over time.

Cross-store experience (OOS) is also positively associated with both on-time delivery and picking efficiency, although the estimated effects are smaller in magnitude. This finding suggests that workers acquire generalizable skills from exposure to diverse store environments, such as navigating

item lists, managing time pressure, or interacting with platform logistics. The weaker curvature in OOS^2 compared to OTS^2 implies that cross-store learning may continue to yield value over a longer horizon, although the evidence for diminishing returns is less robust in this case.

Taken together, the results indicate that performance improves with accumulated experience, particularly in the early stages of store-specific learning. These improvements appear to taper off as workers become more familiar with store processes and stabilize their operational routines. Although the analysis is observational, the use of worker and store fixed effects, along with a comprehensive set of controls, allows us to account for time-invariant differences and to isolate performance dynamics related to accumulated experience. We note that the observed R^2 values are modest, which is expected in models of individual-level performance in operational settings. A substantial portion of outcome variability is likely driven by task-specific factors, such as in-store congestion, product availability, or customer-specific constraints, which are not directly captured in our dataset but are common sources of variation in real-world gig work environments.

5 Responding to Recommendations: *Orders to Bundle*

Our earlier analyses suggest that gig workers improve their performance as they gain experience, but the extent and pattern of this improvement vary considerably across individuals. In this section, we focus on this heterogeneity by examining how workers differ in their learning trajectories and in how they respond to platform-generated task recommendations, particularly those involving order bundling.

We begin by segmenting workers based on their overall tenure on the platform and show that distinct worker groups follow different performance trajectories. We then analyze how workers respond to the platform's bundling recommendations and whether such responses are associated with improved outcomes. While workers may continue to adapt long after joining the platform, we concentrate on the first 350 orders. This range captures the period in which platform-related performance metrics such as service quality and productivity show the most observable change. Although workers may also learn to optimize for personal objectives, such as minimizing stress or maximizing earnings per unit of effort, those dimensions fall outside the scope of this study. We focus here on outcomes that are directly relevant to platform design and performance management.

5.1 Model-free Evidence: Worker Heterogeneity

To better understand variation in learning patterns, we segment the newcomer cohort into five groups based on the quantiles of total orders completed during the study period. We refer to these as *worker tenure groups*, which reflect different levels of platform engagement. The five groups are defined as follows: 0 to 1 orders (261 workers), 2 to 12 orders (197 workers), 13 to 53 orders (228 workers), 54 to 128 orders (221 workers), and 129 or more orders (224 workers).

Figure 2 shows the average on-time delivery probability by cumulative order bin for each tenure group, focusing on the first 350 orders. Several patterns are apparent. Workers in the highest tenure group start with relatively high on-time rates and improve steadily. In contrast, workers in lower tenure groups begin with lower performance and display greater volatility, particularly in the first 50 to 100 orders. This suggests that workers who ultimately stay longer on the platform may follow more consistent learning paths or adopt more effective strategies early on.

These differences point to substantial heterogeneity in how workers engage with the platform. Differences in motivation, operational skill, or familiarity with similar systems may all play a role. To avoid masking this heterogeneity, our subsequent analyses of bundling behavior (Section 5.2) and response to recommendations (Section 6) are conducted separately by tenure group.

One concern in defining worker groups by total orders is the potential confounding effect of join date. Workers who join earlier in the year naturally have more time to accumulate orders.

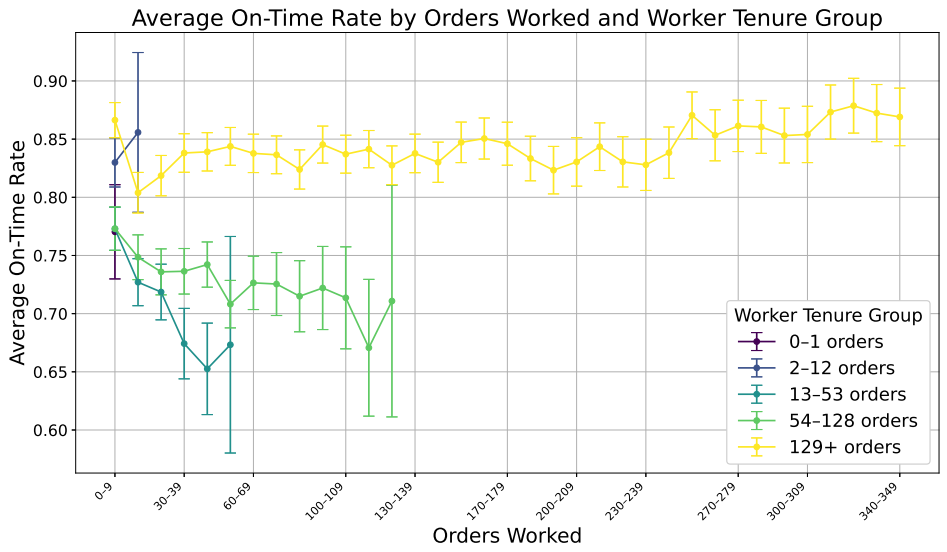


Fig. 2. Average on-time delivery rate by cumulative orders and final tenure group

Notes: This figure plots the average on-time delivery probability (y-axis) as a function of cumulative orders completed (x-axis, grouped in increments of 10) for workers in the newcomer cohort ($N = 1,131$). Workers are divided into five groups based on the total number of orders completed during the one-year study period: 0–1, 2–12, 13–53, 54–128, and 129 or more. Each line represents one tenure group, and error bars indicate 95% confidence intervals for the mean within each bin.

To address this, we conduct robustness checks reported in the Appendix, which control for join date and restrict comparisons to workers who entered the platform during the same time windows. These analyses confirm that the performance differences we observe are not driven solely by timing.

5.2 Model-free Evidence: Learning to Bundle

A key feature of the platform is its use of algorithmic bundling. Approximately 60 percent of all orders are created as bundles by the platform. After completing a few initial orders, workers begin to see these bundled tasks, which consist of two individual orders grouped by an algorithm based on store origin, item similarity, or proximity of delivery destinations. These bundles appear in the same interface as all other available tasks and are labeled as bundled orders. However, they are not explicitly highlighted as recommended. Once a bundled order is accepted, the two constituent deliveries must be completed together and cannot be separated..

In addition to accepting these platform-created bundles, workers have the autonomy to create their own bundles by selecting multiple individual orders to complete concurrently. We define this behavior as *self-bundling*. Using platform timestamps, we identify self-bundling as cases in which a worker has overlapping shopping intervals across multiple orders, indicating that they independently chose to fulfill them at the same time.

While bundling can increase efficiency, it may also introduce complexity. As shown in Figures 1 and 2, average on-time delivery probability tends to decline during the early stage when bundling becomes available. This dip may reflect the added complexity of learning to manage multiple tasks, though it could also result from survivorship effects, as lower-performing workers may exit the

platform before performance recovers. Understanding how bundling behavior evolves and differs across worker groups helps illuminate the broader dynamics of learning and retention.

5.2.1 Overall Bundle Behaviors. Figure 3 shows the average *bundle volume*, defined as the number of orders fulfilled concurrently, across tenure groups during their first 350 orders. All groups exhibit a sharp increase in bundle volume shortly after the first few orders, which aligns with the point at which platform-generated bundles begin appearing in the interface.

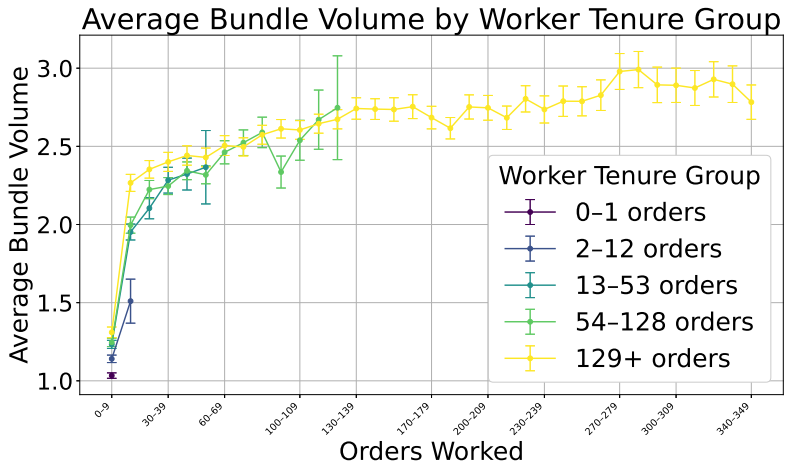


Fig. 3. Average bundle volume across worker groups and over time

Notes: This figure plots the average number of orders fulfilled concurrently (y-axis) against the cumulative number of orders completed (x-axis) for the newcomer cohort ($N = 1,131$). Lines represent different tenure groups, defined by total orders completed over the one-year study period. Error bars indicate 95 percent confidence intervals.

Initially, all worker groups exhibit a sharp increase in bundling behavior within their first few orders, coinciding with the point at which platform-generated bundles begin appearing in the interface. These bundles are created algorithmically and presented to workers as grouped tasks, which cannot be accepted separately. Workers in the highest tenure group (those completing more than 129 orders during the study period, shown in yellow) consistently display the highest average bundle volume across the experience range. This group reaches a peak of approximately three bundled orders per 100 orders and maintains a relatively stable level of bundling throughout their tenure. This pattern suggests that these workers integrate bundling more fully into their routines, potentially using it to increase efficiency or earnings.

In contrast, early exiters (workers who completed between 0 and 12 total orders, shown in purple and blue) engage only minimally with bundled tasks before leaving the platform. The mid-tier groups (13–53 and 54–128 orders) show moderate uptake of bundling, maintaining levels that are slightly lower than the highest tenure group, particularly in the first 30 to 50 orders.

Overall, the results suggest that frequent engagement with bundling is is more common among workers who remain active on the platform for longer periods. Whether bundling behavior contributes to retention or is simply correlated with other factors such as familiarity, motivation, or efficiency is unclear from these descriptive trends. Nonetheless, the consistent pattern of higher bundling among long-tenure workers suggests that it is a feature more fully adopted by those

who are most engaged with the platform. Further analysis of these patterns, including statistical comparisons and controls for worker characteristics, can be found in the appendix.

5.2.2 *Platform-Bundled vs. Self-Bundled Orders.*

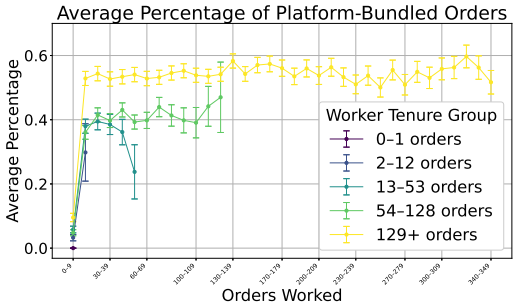


Fig. 4. Platform-bundled orders as a share of all orders

Notes: This figure shows the percentage of all completed orders that were part of a platform-generated bundle (y-axis), plotted by cumulative order bins (x-axis). Workers are grouped by total order volume during the study period. Each percentage is calculated within each group and bin. Error bars represent 95 percent confidence intervals.

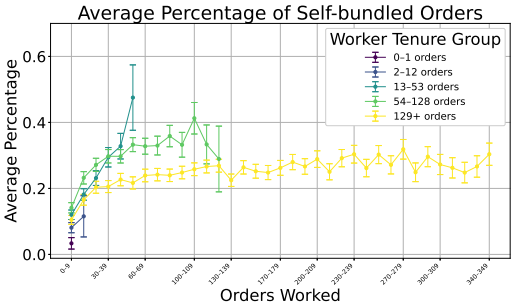


Fig. 5. Self-bundled orders as a share of all orders

Notes: This figure shows the percentage of all completed orders that were part of a self-initiated bundle (y-axis), where bundling is inferred from overlapping shopping intervals. The x-axis represents cumulative order bins, and lines reflect different tenure groups. Error bars represent 95 percent confidence intervals.

Figure 4 presents the share of completed orders that were part of platform-generated bundles. Each value reflects the proportion of bundled tasks among all completed orders in a given experience bin, including single, platform-bundled, and self-bundled orders. Across all tenure groups, there is a sharp initial increase in the adoption of platform bundles within the first 20 orders, coinciding with the point at which bundled tasks begin appearing in the interface. After this early phase, the highest tenure group (129 or more orders completed) stabilizes at a platform bundling rate above 50 percent and maintains relatively low within-group variation. By contrast, lower tenure groups converge to lower and more variable platform bundle adoption rates, typically below 50 percent.

Figure 5 displays the corresponding share of self-bundled orders. Here, mid-tenure groups (13–53 and 54–128 orders completed) exhibit the highest average rates of self-bundling, peaking at approximately 30 to 40 percent. These workers appear to explore self-initiated bundling strategies more frequently during their time on the platform. In contrast, the highest tenure group engages in self-bundling at consistently lower rates, stabilizing below 30 percent after the early stages of experience. This group may be more likely to integrate platform-generated bundles into their routines, potentially reducing the need for manual coordination of tasks.

Taken together, these descriptive trends point to meaningful differences in bundling strategies across worker tenure groups. While all workers encounter bundling opportunities early in their experience, those with longer tenure are more likely to maintain consistent use of platform-generated bundles. In contrast, mid-tenure workers exhibit higher levels of self-bundling, potentially reflecting a greater degree of experimentation with independent strategies. These patterns suggest that sustained engagement with platform-generated bundles may be associated with higher retention and performance, whereas self-bundling appears more common among workers who exit the platform earlier in their tenure.

6 Responding to Recommendations: *Orders to Select*

While gig workers can freely choose to work on any available order on the platform, the platform typically recommends a subset of orders through its recommendation algorithm. These *algorithmically recommended orders* are determined based on platform-side considerations such as demand patterns and the worker's past performance. Recommended and non-recommended orders are presented in separate tabs within the worker interface, allowing workers to browse both categories at the time of selection.

To examine how workers interact with these recommendations, we model each task selection as a discrete choice among a set of available alternatives. Specifically, we apply a multinomial logit (MNL) model [34] to estimate the probability that a worker selects a given order, conditional on the attributes of the order and the worker's experience level. This framework allows us to quantify how order features such as pay, distance, and recommendation status influence task selection, and how these preferences evolve as workers accumulate experience.

The MNL model is well-suited to this setting for both behavioral and structural reasons. It assumes that each worker, when presented with a choice set, selects the order that maximizes their utility based on observable order attributes and their own evolving preferences. Importantly, it allows us to estimate trade-offs that workers make between recommended and non-recommended tasks, while conditioning on the full set of options available to them at the time of decision.

To capture how decision-making changes with experience, we focus on each worker's first 200 completed orders. We divide this period into five consecutive 40-order segments. While alternative specifications using continuous experience variables are possible, we use equal-width bins to facilitate comparison with the tenure-based grouping introduced in Section 5 and to allow for nonparametric heterogeneity in choice behavior over time. This discretization enables a clearer examination of how alignment with platform recommendations shifts as workers gain familiarity with the system.

In the following subsection, we describe the choice set construction, the dependent variable, and the key features included in the MNL specification.

6.1 Multinomial Logit Model of Workers' Selected Orders

To investigate how gig workers respond to platform recommendations and how this behavior evolves with experience, we estimate a multinomial logit (MNL) model across sequential phases of a worker's early experience. We segment each worker's first 200 completed orders into five experience-based periods and estimate separate MNL models for each segment. The MNL framework allows us to analyze how workers choose one order from a set of simultaneously available alternatives, modeling the probability of choosing a particular order as a function of its attributes and the worker's characteristics.

The MNL model is grounded in utility maximization theory and is widely used to model decision-making in discrete choice contexts. In our setting, each worker is presented with a list of orders—some algorithmically recommended, others not—and selects one to fulfill. The model estimates how features such as recommendation status, store familiarity, pay, and bundling influence the likelihood of selection, and how these relationships evolve with increasing platform experience.

6.1.1 Description of Key Variables. We define each choice occasion as a one-hour window prior to the time a worker accepts an order. The choice set includes all tasks available to the worker during that hour, including both algorithmically recommended and non-recommended orders. Each alternative within a choice set represents a single task unit: either an individual delivery order or a platform-generated bundled task (i.e., two orders combined and offered as one indivisible unit).

Workers typically observe dozens of alternatives within a given one-hour window. To normalize scales and improve interpretability, all variables are normalized before estimation.

Dependent variable. The dependent variable *CHOSEN* in the model is a binary indicator representing whether a specific alternative was chosen that takes the value 1 if the alternative was chosen by the gig worker, and 0 otherwise.

Independent variables. We document the important independent variables included in the MNL model here. We defer the complete list of variables to Appendix D. These variables represent characteristics of the alternatives and other relevant attributes influencing the decision. Each of these variables contributes to the deterministic component of the utility function, V_{ij}, for each alternative j faced by gig worker i.

- *LIST*: Indicator variable equal to 1 if the order appeared in the platform's recommendation list (i.e., in the recommended tab), and 0 otherwise.
- *PastFrequency*: The proportion of all previous orders completed by the worker that were fulfilled at the same store. This serves as a proxy for store familiarity. If the worker has not previously completed an order at the store, this value is set to 0.

To capture how the influence of key factors evolves with experience, we incorporate two complementary approaches. First, we introduce interaction terms between the main explanatory variables and the cumulative number of orders completed by the worker at the time of decision. These interactions allow us to model how the marginal effects of recommendations and store familiarity change as workers gain experience. Second, we classify workers into five tenure groups based on the total number of orders completed during the one-year period (0–1, 2–12, 13–53, 54–128, and 129+ orders) and introduce group-specific effects using one-hot encoded dummy variables, with the 129+ group as the reference. We then interact these group indicators with key independent variables to estimate how different types of workers respond to the same order-level features. This modeling strategy allows us to analyze both within-worker learning dynamics and cross-worker heterogeneity in decision-making. Only interaction terms that provide non-redundant information are retained in the model to ensure efficient and interpretable specifications.

Although external factors such as traffic and weather (discussed in Section 3) were included in the performance models in Section 4, they are omitted from this MNL specification. This is because all alternatives within a given choice set are evaluated at the same point in time, meaning that variables such as temperature or traffic congestion do not vary across options in the same set. Since the MNL model estimates utility differences across alternatives within each choice occasion, only variables that vary across those alternatives can be identified. As such, we focus on order-specific features and experience-based variables that are observable and differentiable across the set of available tasks at the time of decision.

6.1.2 *Model specification.*

Utility function. The utility function U_{ij} for alternative j and gig worker i is composed of a deterministic component V_{ij} and a stochastic component ϵ_{ij}:

$$U_{ij} = V_{ij} + \epsilon_{ij} \tag{2}$$

The deterministic component V_{ij} is modeled as a linear function of the explanatory variables:

$$V_{ij} = \beta_0 + \beta_1 \cdot \text{Bundled}_{ij} + \beta_2 \cdot \text{List}_{ij} + \dots \tag{3}$$

Here, β_k represents the coefficient associated with each explanatory variable, and ϵ_{ij} is the error term, assumed to follow a Gumbel distribution.

Choice probabilities. The probability that gig worker i chooses alternative j is given by the following multinomial logit probability function:

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{l=1}^{J} \exp(V_{il})} \tag{4}$$

where J is the number of available alternatives in the choice set.

6.1.3 Estimation method. Model parameters are estimated using Maximum Likelihood Estimation (MLE). The log-likelihood function is defined as:

$$\ln L(\beta) = \sum_{i=1}^{N} \left(V_{iy_i} - \ln \left(\sum_{l=1}^{J} \exp(V_{il}) \right) \right) \tag{5}$$

where y_i denotes the alternative chosen by worker i within their observed choice set. Coefficient estimates are computed separately for each of the five cumulative order intervals: 0–1 (excluded due to sparsity), 2–12, 13–53, 54–128, and 129+ orders. Standard errors are clustered at the worker level. This modeling structure enables us to observe how worker decision-making evolves both over time and across different types of workers.

6.2 Results: *Workers Follow Recommendations Less with More Experience*

We estimate separate multinomial logit (MNL) models across four cumulative order ranges (2–12, 13–53, 54–128, and 129+) and segment workers into four corresponding tenure groups based on their total number of completed orders. We exclude the 0–1 group and order range due to limited data and noise.

We focus on two main predictors: *PastFrequency*, which captures store familiarity and reflects exploitation behavior, and *LIST*, which indicates whether an order was shown as a platform recommendation. Each MNL model includes interaction terms between these predictors and both cumulative experience and tenure group indicators to capture heterogeneity in decision-making behavior over time.

6.2.1 Store Familiarity. Figure 6 presents the estimated coefficients for $PastFrequency$, which captures a worker's tendency to return to stores they have visited before. For the reference group of workers who completed 129 or more orders, we observe consistently strong and positive coefficients across all order ranges. This pattern indicates that high-tenure workers are significantly more likely to choose tasks at familiar stores, suggesting a strong preference for exploitation strategies.

In contrast, the interaction terms for the lower-tenure groups (2–12, 13–53, and 54–128) are predominantly negative, especially in the earlier order ranges. This means that less experienced workers place substantially less weight on store familiarity relative to the 129+ group. For instance, in the 2–12 range, both the 2–12 and 13–53 groups show a significantly weaker association between past frequency and choice probability.

The purple bars in Figure 6 display the interaction between $PastFrequency$ and cumulative orders completed within each experience bin. These interactions reveal a more nuanced trend. In the earliest ranges, there is no clear evidence that workers change their reliance on store familiarity as they complete more orders. However, in the 54–128 range, the interaction becomes significantly positive, suggesting that workers begin to favor familiar stores more as they gain experience within that phase. Interestingly, this trend reverses in the final bin (129+), where the coefficient turns negative. This reversal implies that workers may shift away from familiarity-based heuristics after gaining extensive platform experience.

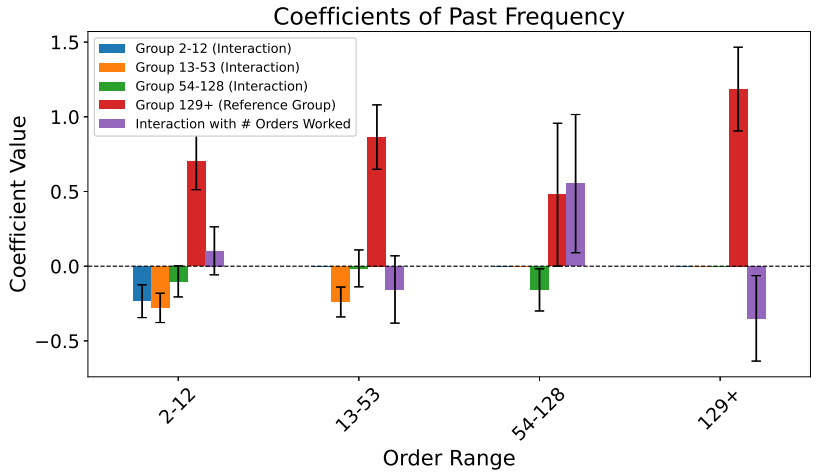


Fig. 6. Estimated MNL Coefficients of Past Frequency by Order Range and Tenure Group.

Notes: This figure reports coefficients from separate multinomial logit models estimated across four cumulative order bins (x-axis: 2–12, 13–53, 54–128, and 129+ orders). The red bar shows the baseline coefficient for *PastFrequency* in the reference group (workers with 129+ total orders). Blue, orange, and green bars represent interaction terms between *PastFrequency* and dummy variables for other tenure groups (2–12, 13–53, and 54–128, respectively), capturing group-specific deviations from the reference. The purple bar indicates the interaction between *PastFrequency* and the worker's cumulative experience (number of orders completed at the time of decision). Error bars denote 95% confidence intervals. For non-reference groups, the total marginal effect of *PastFrequency* equals the sum of the red bar and the corresponding group-specific interaction coefficient.

Overall, these findings indicate that store familiarity plays an important role in decision-making for the most experienced workers, but its influence is not static. Less experienced workers rely less on this heuristic, and even the most experienced users appear to reduce their dependence on store familiarity as their order volume increases beyond a certain threshold.

These insights provide a different perspective from prior work such as Dai et al. (2022) [9], which used proxies like visiting new areas or stores to infer exploration behavior. While that study suggested that early-stage workers are more exploratory, our choice-based model shows that exploitation based on familiarity is prominent, especially among high-tenure workers, when choices are analyzed at the moment of decision.

6.2.2 Algorithmic Recommendations. Figure 7 shows the coefficients related to the variable *LIST*, indicating whether an order was included in the platform's recommendation list. For the 129+ reference group, we observe a strong and positive effect of recommendation status in the early stages, especially in the 2–12 and 13–53 ranges. This implies that workers who eventually reach high tenure tend to follow recommendations early in their experience. However, this effect drops sharply in the 54–128 and 129+ ranges, where the coefficients become statistically indistinguishable from zero. This suggests that the role of recommendations weakens as experienced workers become more independent in their task selection.

When comparing other tenure groups to the reference group, we find additional differences. In the 2–12 range, both the 13–53 and 54–128 groups show significantly negative interaction terms.

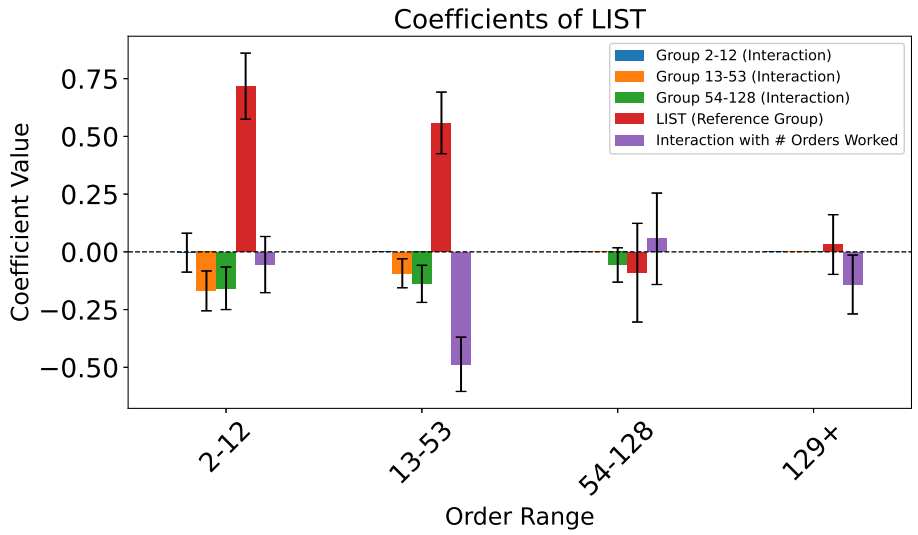


Fig. 7. Estimated MNL Coefficients Related to Recommendation Status *LIST* by Order Range and Tenure Group.

Notes: This figure reports coefficients from separate multinomial logit models estimated across four cumulative order bins (x-axis: 2–12, 13–53, 54–128, and 129+ orders). The red bar shows the baseline coefficient for *LIST* in the reference group (workers with 129+ total orders). Blue, orange, and green bars represent interaction terms between *LIST* and dummy variables for other tenure groups (2–12, 13–53, and 54–128, respectively), capturing group-specific deviations from the reference. The purple bar indicates the interaction between *LIST* and the worker's cumulative experience (number of orders completed at the time of decision). Error bars denote 95% confidence intervals. For non-reference groups, the total marginal effect of *LIST* equals the sum of the red bar and the corresponding group-specific interaction term.

This indicates that these workers are less likely to follow platform recommendations compared to the highest-tenure workers during the same early phase. These group-level differences fade in later stages.

The purple bars, which show the interaction between *LIST* and cumulative orders worked, consistently take on negative values across the order ranges. The coefficients are significantly negative in both the 13–53 and 129+ ranges, providing robust evidence that the influence of recommendations declines with experience. As workers complete more orders, they are less likely to rely on the recommendation tab and more likely to make independent selections.

Summary of Insights. Together, these results show that worker decision-making evolves in systematic and measurable ways. High-tenure workers rely heavily on familiarity and follow recommendations early in their experience, but both tendencies diminish as they complete more orders. Workers with lower tenure tend to show weaker reliance on both recommendations and store familiarity, and their behaviors exhibit more heterogeneity across time.

These findings suggest that recommendation systems might be more effective during the early stages of a worker's tenure but may lose their influence as workers gain confidence and experience. Designing adaptive recommendation mechanisms that evolve with the user's level of familiarity and strategic intent could improve long-term engagement and efficiency on the platform.

7 Discussion

This study advances understanding of how gig workers interact with algorithmic systems by analyzing how task selection and bundling strategies evolve with experience. We contribute to the literature on human-algorithm interaction and gig worker learning by modeling worker responses to platform-provided alternatives at the moment of decision. Our findings show how workers differentiate among algorithmic tools and adjust their reliance over time.

Contribution to literature. We build on prior work documenting gig worker adaptation through exploration and heuristic learning [9, 11, 19]. Unlike studies that infer learning from broad behavior patterns (e.g., entering new areas), we model decisions using multinomial logit analysis with full choice sets, enabling us to estimate how specific task attributes—such as store familiarity and recommendation status—influence selection. On bundling, our findings extend earlier work noting that workers often reject platform-generated batches [19]. We show that workers' bundling behavior is dynamic: high-tenure workers increasingly adopt platform-generated bundles, while mid-tenure workers experiment more with self-bundling. These trends suggest that bundling strategies are not static but evolve with experience and engagement. To our knowledge, this is one of the first studies to empirically track bundling behavior over time using large-scale data.

Understanding reliance on Recommendations. Our choice model results reveal a declining influence of platform recommendations. The coefficient on recommendation status *LIST* is positive for early-stage workers but becomes statistically negligible as tenure increases. While our analysis compares coefficients across models and interaction terms, the consistency of this pattern suggests a robust shift in behavior. This decline coincides with increased acceptance of platform-generated bundles, as shown in Section 5. These patterns are not contradictory. Platform-generated bundles are indivisible units optimized for efficiency and may be evaluated using observable features. In contrast, recommendations rely more on subjective fit and may be compared against personal heuristics. Experienced workers may retain trust in optimization-based tools while becoming less reliant on recommendation algorithms that no longer align with their strategies. This distinction contributes to the literature on algorithmic management and human-AI collaboration. Rather than treating algorithmic advice uniformly, workers differentiate based on task type and perceived value. Our findings align with research on selective algorithm use [5, 14] and the need for human-centered AI systems that adjust to user experience [16, 41].

Implications for adaptive platform design. Our findings suggest that recommendation systems should adapt as workers gain experience. While static suggestions may assist with onboarding, they become less effective once workers develop their own task selection strategies. The decline in responsiveness to recommendations, especially during the intermediate tenure phase, indicates an opportunity for platforms to adjust how they prioritize or present recommendations. In contrast, the continued adoption of platform-generated bundles by experienced workers shows that algorithmic tools remain valuable when their benefits are clearly observable. Although this study focuses on a platform that grants workers significant autonomy, the behavioral patterns we observe may extend to other gig platforms where workers can exercise some degree of choice. These include early reliance on platform guidance, increasing preference for familiar tasks, and selective use of algorithmic inputs as workers gain experience. In settings where worker discretion is more limited, adaptation may take the form of modifying the timing, presentation, or framing of tasks rather than allowing direct task selection. Overall, these findings point to the importance of aligning algorithmic support with the evolving preferences and capabilities of workers. Systems that adjust to worker experience levels and behavioral signals may be more effective in promoting both worker satisfaction and platform performance. Future work should examine how these strategies perform across a broader range of platform models and labor contexts.

Limitations and future work. Several limitations warrant caution. First, our choice set approximations are based on hourly snapshots and do not capture visibility, ranking, or UI design. Second, tenure-based grouping may reflect duration or cohort effects, though robustness checks help mitigate this. Third, our findings stem from a single platform in one urban context and may not generalize to platforms with different dispatch models or lower autonomy. We also do not observe algorithm or interface changes, which could influence user behavior. Nor can we capture rejected or unseen tasks. Finally, our focus is on individual decision-making and does not incorporate social learning or coordination, which may be influential in gig work environments. Future research could integrate interface logs or worker interviews to contextualize decisions, study multiple platform models, or explore how worker and algorithm behavior co-evolve. Extending this work to other labor platforms will help assess generalizability and refine our understanding of human-algorithm adaptation.

8 Concluding Remarks

This study examines how gig workers learn and adapt to algorithmic systems on a U.S.-based retail delivery platform. We address three core research questions: (1) how workers improve performance over time, (2) how they respond to platform-generated bundling and recommendation systems, and (3) how decision-making evolves with experience. These questions speak to broader concerns in human-algorithm interaction and the design of decision-support systems in labor platforms.

Our analysis proceeds in three parts. First, we examine learning curves using descriptive trends and two-way fixed effects regressions to measure how worker performance improves with experience. Second, we analyze bundling behavior by comparing engagement with platform-generated and self-initiated bundles across tenure groups. Finally, we apply a multinomial logit (MNL) model to study task selection behavior at the moment of choice, estimating how workers trade off algorithmic recommendations and store familiarity over time.

Across all methods, we find that workers improve both service quality and productivity with experience, though learning is heterogeneous. High-tenure workers adopt platform-generated bundles more frequently, suggesting continued reliance on optimization tools. At the same time, these workers rely less on platform-generated order recommendations, preferring familiar stores and self-developed heuristics as they gain confidence. The divergence in responses to different types of algorithmic support illustrates how workers learn not only to perform tasks more efficiently, but also to evaluate and selectively engage with the tools offered by the platform.

By linking learning, task selection, and algorithmic response within a unified empirical framework, this paper contributes to research on gig economy labor, adaptive algorithm design, and behavioral operations. It highlights the importance of tailoring platform support based on workers' experience and revealed preferences. As platforms increasingly shape how work is allocated, understanding how workers develop expertise and autonomy in response to algorithmic systems will be critical for improving both platform outcomes and worker well-being.

References

[1] Zeynep Akşin, Sarang Deo, Jónas Oddur Jónasson, and Kamalini Ramdas. 2021. Learning from many: Partner exposure and team familiarity in fluid teams. *Management Science* 67, 2 (Feb. 2021), 854–874.

[2] Gad Allon, Maxime C Cohen, and Wichinpong Park Sinchaisri. 2023. The impact of behavioral and economic drivers on gig economy workers. *Manufacturing & Service Operations Management* 25, 4 (2023), 1376–1393.

[3] Linda Argote. 2012. *Organizational learning: Creating, retaining and transferring knowledge* (2 ed.). Springer, New York, NY.

[4] Andrea Barraza-Urbina. 2017. The Exploration-Exploitation Trade-off in Interactive Recommender Systems. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys)*. 157–161. https://doi.org/10.1145/3109859.3109866

[5] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2024. Improving Human Sequential Decision-Making with Reinforcement Learning. https://doi.org/10.48550/ARXIV.2108.08454

[6] Hessam Bavafa and Jónas Oddur Jónasson. 2021. The Variance Learning Curve. *Management Science* 67, 5 (May 2021), 3104–3116.

[7] Colin Camerer and Teck Hua Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* 67, 4 (1999), 827–874.

[8] Jonathan R Clark, Robert S Huckman, and Bradley R Staats. 2013. Learning from customers: Individual and organizational effects in outsourced radiological services. *Organization Science* 24, 5 (Oct. 2013), 1539–1557.

[9] Hongyan Dai, Jayashankar M Swaminathan, and Yuqian Xu. 2022. Leveraging the Experience: Exploration and Exploitation in Gig Worker Learning Process. (May 2022).

[10] Ezey M Dar-El. 2013. *HUMAN LEARNING: From learning curves to learning organizations* (2000 ed.). Springer, New York, NY.

[11] Vedant Das Swain, Lan Gao, Abhirup Mondal, Gregory D Abowd, and Munmun De Choudhury. 2024. Sensible and sensitive AI for worker wellbeing: Factors that inform adoption and resistance for information workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 105. ACM, New York, NY, USA, 1–30.

[12] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D Abowd, and Munmun De Choudhury. 2023. Algorithmic power or punishment: Information worker perspectives on passive sensing enabled AI phenotyping of performance and wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 105. ACM, New York, NY, USA, 1–17.

[13] Esra Cemre Su de Groot and Ujwal Gadiraju. 2024. "Are we all in the same boat?" Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 640, 26 pages. https://doi.org/10.1145/3613904.3642429

[14] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126.

[15] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (March 2018), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[16] Kimberly Do, Maya De Los Santos, Michael Muller, and Saiph Savage. 2024. Designing gig worker sousveillance tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 1. ACM, New York, NY, USA, 1–19.

[17] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1013–1022.

[18] Wai Fong Boh, Sandra A Slaughter, and J Alberto Espinosa. 2007. Learning from experience in software development: A multilevel analysis. *Management Science* 53, 8 (Aug. 2007), 1315–1331.

[19] Shreepriya Gonzalez-Jimenez, Cecile Boulard, Clara Tuco, and Romane Calleau. 2022. Designing Food Delivery Gig-platforms for Courier Needs: the Case of Batched Orders. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, Taiwan) *(CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 163–167. https://doi.org/10.1145/3500868.3559452

[20] Eric H Grosse and Christoph H Glock. 2015. The effect of worker learning on manual order picking processes. *Int. J. Prod. Econ.* 170 (Dec. 2015), 882–890.

[21] Reeju Guha and Daniel Corsten. 2023. The Role of Within-Day Learning on Gig Workers' Performance and Task Allocation: Evidence from an On-demand Platform. *Available at SSRN* (2023).

[22] Rie Helene (lindy) Hernandez, Qiurong Song, Yubo Kou, and Xinning Gui. 2024. "At the end of the day, I am accountable": Gig Workers' Self-Tracking for Multi-Dimensional Accountability Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 54. ACM, New York, NY, USA, 1–20.

[23] Mohammad Hossein Jarrahi, Gemma Newlands, Min Kyung Lee, Christine T Wolf, Eliscia Kinder, and Will Sutherland. 2021. Algorithmic management in a work context. *Big Data & Society* 8, 2 (2021), 20539517211020332.

[24] Mohammad Hossein Jarrahi and Will Sutherland. 2019. Algorithmic management and algorithmic competencies: Understanding and appropriating algorithms in gig work. In *International Conference on Information*. Springer, 578–589.

[25] Gi-Soo Kim and Myunghee Cho Paik. 2019. Contextual Multi-armed Bandit Algorithm for Semiparametric Reward Model. *arXiv preprint arXiv:1901.11221* (2019). https://arxiv.org/abs/1901.11221

[26] Sangmi Kim, Elizabeth Marquis, Rasha Alahmad, Casey S. Pierce, and Lionel P. Robert Jr. 2018. The Impacts of Platform Quality on Gig Workers' Autonomy and Job Satisfaction. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 181–184. https://doi.org/10.1145/3272973.3274050

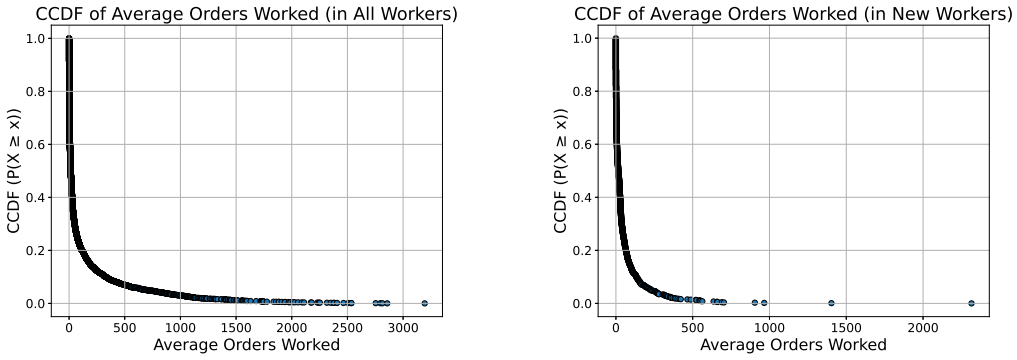
[27] Benjamin Knight, Dmitry Mitrofanov, and Serguei Netessine. 2022. The impact of ai technology on the productivity of gig economy workers. *Available at SSRN 4372368* (2022).

[28] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
[29] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1603–1612.
[30] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*. 661–670. https://doi.org/10.1145/1772690.1772758
[31] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*. 297–306. https://doi.org/10.1145/1935826.1935875
[32] Shuhao Ma. 2024. Advancing HCI and design methods to empower gig workers. In *Designing Interactive Systems Conference*. ACM, New York, NY, USA.
[33] Elizabeth B. Marquis, Sangmi Kim, Rasha Alahmad, Casey S. Pierce, and Lionel P. Robert Jr. 2018. Impacts of Perceived Behavior Control and Emotional Labor on Gig Workers. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 241–244. https://doi.org/10.1145/3272973.3274065
[34] Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior. (1972).
[35] Saiph Savage. 2024. Unveiling AI-driven collective action for a worker-centric future. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA.
[36] Scott M Shafer, David A Nembhard, and Mustafa V Uzumeri. 2001. The Effects of Worker Learning, Forgetting, and Heterogeneity on Assembly Line Productivity. *Management Science* 47, 12 (Dec. 2001), 1639–1653.
[37] Riyaj Shaikh, Anubha Singh, Barry Brown, and Airi Lampinen. 2024. Not Just A Dot on The Map: Food Delivery Workers as Infrastructure. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 385, 15 pages. https://doi.org/10.1145/3613904.3641918
[38] Clara Tuco, Cécile Boulard, Romane Calleau, and Shreepriya Shreepriya. 2021. Food Delivery Eco-System: When Platforms Get Enterprises and Gig-Workers to Implicitly Cooperate. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 183–186. https://doi.org/10.1145/3462204.3481757
[39] Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data*. MIT press.
[40] Zheng Yao, Silas Weden, Lea Emerlyn, Haiyi Zhu, and Robert E Kraut. 2021. Together but alone: Atomization and peer support among gig workers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.
[41] Angie Zhang, Alexander Boltz, Jonathan Lynn, Chun-Wei Wang, and Min Kyung Lee. 2023. Stakeholder-centered AI design: Co-designing worker tools with gig workers through data probes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 63. ACM, New York, NY, USA, 1–19.
[42] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.
[43] Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang, Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, and Philip S. Yu. 2025. Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap. *arXiv preprint arXiv:2501.01945* (2025). https://arxiv.org/abs/2501.01945

A Additional Statistics of Our Dataset for Section 3

To understand the distribution of worker activity on the platform, Figure 8 presents the Complementary Cumulative Distribution Function (CCDF) of the total orders completed per worker during the one-year study period. Panel 8a displays this distribution for all approximately 5000 active workers, while panel 8b focuses on the subset of 1131 newcomers who joined during the study period and are the primary focus of our subsequent learning analyses. Both plots reveal a highly skewed distribution characteristic of many online platforms: a large fraction of workers completes a very small number of orders (indicated by the steep initial drop in the CCDF), while a long tail consists of a smaller number of highly active workers responsible for a large volume of orders. This significant heterogeneity in engagement is evident even within the newcomer cohort, motivating

our later analyses which segment workers based on their activity levels (tenure groups) to explore variations in learning and strategy development.



(a) All Workers (N ≈ 5000)

(b) Newcomer Cohort (N = 1131)

Fig. 8. Complementary Cumulative Distribution Function (CCDF) of Total Orders Completed per Worker.

Notes: These figures show the empirical CCDF of the total number of orders completed per worker over the one-year study period. The Y-axis represents the probability P(X ≥ x), i.e., the fraction of workers completing X or more orders. (a) Shows the distribution for all active workers in the dataset (N ≈ 5000). (b) Shows the distribution for the cohort of newcomers (N = 1131) who joined during the study period and form the basis for analyses in Sections 4-6. Both plots illustrate the highly skewed distribution of worker activity, highlighting heterogeneity in engagement.

In addition to worker activity, we also examined the distribution of order volume across the approximately 800 stores included in the dataset. Figure 9 displays the CCDF for the total number of orders processed per store over the study year. Similar to worker activity, store order volume exhibits a highly skewed distribution. A substantial fraction of stores handles a relatively low volume of orders, indicated by the steep initial decline in the CCDF curve. Conversely, a long tail emerges, representing a smaller number of high-volume stores that process a disproportionately large share of the total orders. This heterogeneity in store activity levels is important context for understanding potential variations in worker experience, the relevance of store-specific learning (Section 4), and the operational landscape of the platform.

Complementing the distribution of order volume per store, Figure 10 illustrates the heterogeneity in worker traffic across stores by plotting the CCDF of the number of unique workers completing at least one order per store. Echoing the pattern seen for order volume and worker activity, this distribution is also highly skewed. A large proportion of the approximately 800 stores in the dataset received orders from only a small number of distinct workers during the year. Conversely, a long tail indicates that certain high-traffic stores were visited by a much larger and more diverse set of workers. This variation in the number of unique workers interacting with each store provides further context for understanding the environment, potentially influencing factors like store-specific congestion or the diversity of worker experience associated with particular locations.

To provide context for the recommendation environment faced by workers, Figure 11 illustrates the distribution of the number of algorithmically recommended orders available in workers' choice sets at the time they made a selection. The CCDF plot reveals a highly skewed distribution. In the vast majority of decision instances, workers were presented with only a small number of

Fig. 9. Complementary Cumulative Distribution Function (CCDF) of Total Orders Processed per Store.

Notes: This figure shows the CCDF of the total number of orders processed per store over the one-year study period, including all participating stores (N ≈ 800). The Y-axis represents the probability that a randomly selected store processed X or more orders, where X is the value on the X-axis. The plot illustrates the highly skewed distribution of store activity, with a large proportion of stores handling relatively few orders and a long tail of stores processing a very high volume of orders. This highlights the heterogeneity in store importance and workload within the platform's operations.

recommended orders (often fewer than 10). However, the long tail indicates that occasionally, workers encountered choice sets containing a very large number of recommended options (up to 200).

B Robustness Check of Join Date and Worker Tenure Groups for Section 5

Regarding the potential confounding effect between worker tenure (defined by total orders completed) and worker join date, we conducted an additional analysis. Workers achieving higher tenure might have simply joined the platform earlier in the study period, giving them more time to accumulate orders.

We first explicitly analyzed the distribution of join dates across our newcomer tenure quantile groups (defined in Section 5.1). The analysis revealed a statistically significant difference in mean join dates across the groups (ANOVA $p < 0.001$), confirming that workers achieving higher tenure generally joined earlier in the study year. We acknowledge this statistical finding and its implication that tenure and start date within the year are correlated in this cohort.

However, we also examined the practical distribution visually. As shown in Figure 12, while the median join dates differ (noticeably earlier for the 130+ group, corresponding to the 129+ group label used elsewhere), there is considerable overlap in the interquartile ranges and overall

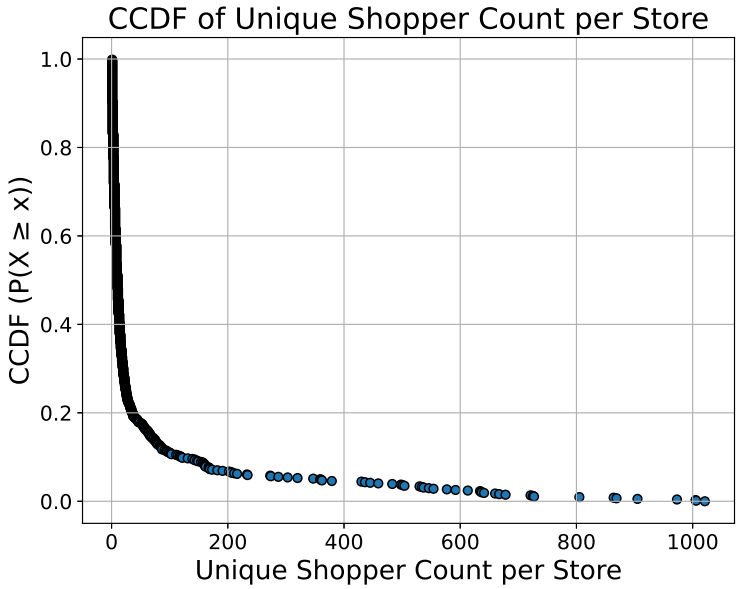


Fig. 10. Complementary Cumulative Distribution Function (CCDF) of Unique Worker Count per Store.

Notes: This figure shows the CCDF of the number of unique workers (shoppers) who completed at least one order from each store over the one-year study period, including all participating stores ($N \approx 800$). The Y-axis represents the probability $P(X \geq x)$, i.e., the fraction of stores visited by X or more unique workers. The plot illustrates a highly skewed distribution: many stores are served by only a few unique workers, while a small number of stores attract orders from a large, diverse set of workers. This highlights heterogeneity in worker exposure across different store locations.

distributions across the five quantile groups. This suggests that workers who joined at various points throughout the year are still represented across most tenure levels.

To further assess the impact of this correlation, we performed a robustness check focusing only on the cohort of workers who joined in the first half of the study year. Figure 13 presents the performance trajectories (Average On-Time Rate vs. Orders Worked) for this restricted cohort. By analyzing only workers who started within a similar timeframe, we significantly mitigate the confounding effect of join date.

Crucially, Figure 13 demonstrates that even within this cohort, significant differences in performance trajectories exist between the eventual tenure groups. The group that eventually completed the most orders (129+) still exhibits a distinctly higher and more stable On-Time Rate from early on compared to groups that completed fewer orders.

This finding provides strong evidence that the observed association between higher eventual tenure and better performance is not solely an artifact of earlier join dates. It suggests there are underlying differences in learning, strategy, or initial characteristics associated with workers who persist and achieve higher engagement, even when comparing workers who started around the same time.

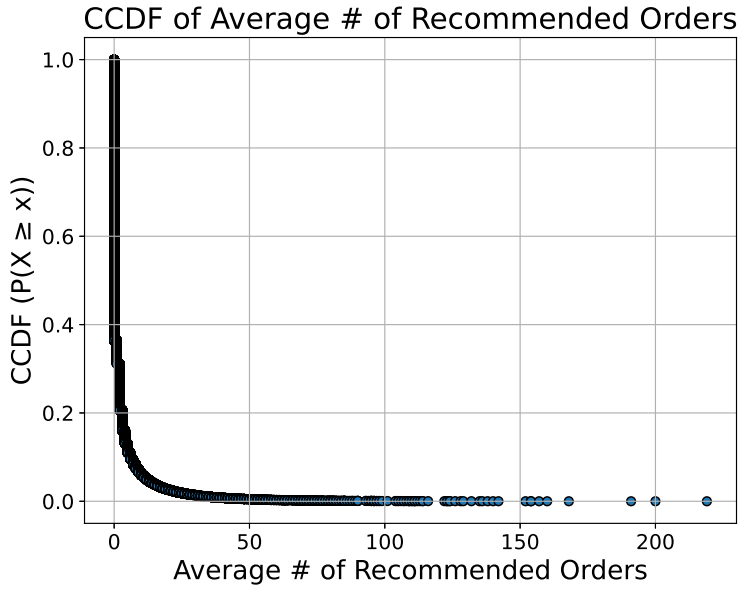


Fig. 11. Complementary Cumulative Distribution Function (CCDF) of the Number of Recommended Orders per Choice Set.

Notes: This figure shows the empirical CCDF of the number of algorithmically recommended orders present in the choice sets faced by workers each time they selected an order during the study period (mean = 2.47, Quantiles (25%, 50%, 75%) = [0. 0. 2.]. The Y-axis represents the probability P(X ≥ x), i.e., the fraction of observed choice sets containing X or more recommended orders. The X-axis represents the number of recommended orders in the choice set (despite the axis label mentioning "Average"). The plot shows a highly skewed distribution: most choice sets presented to workers contained very few recommended orders, while a small fraction of choice sets contained a large number of recommendations. This provides context for the recommendation environment workers navigated.

Therefore, while the statistically significant correlation between join date and final tenure indeed is a limitation, the results of this robustness check support the value of using the quantile-based tenure groupings for our descriptive analyses in Section 5 and Section 6.

C Statistical Testing for Section 5

C.1 Statistical Testing for On-time Rate Difference

Table 2 presents the detailed results of the Tukey HSD (Honestly Significant Difference) post-hoc test (significan level = 0.05), conducted following a significant ANOVA result ($F = 8.114$, $p < 0.001$) comparing the mean On-Time rate across the five worker tenure quantile groups. This specific analysis focuses on the performance achieved during the workers' very earliest experience on the platform, specifically within the 0–1 orders completed bin. The test performs pairwise comparisons between all tenure groups to identify which specific groups had significantly different mean OTPs during this initial period.

The columns in the table represent:

- **Group 1 / Group 2**: The pair of worker tenure quantile groups being compared.

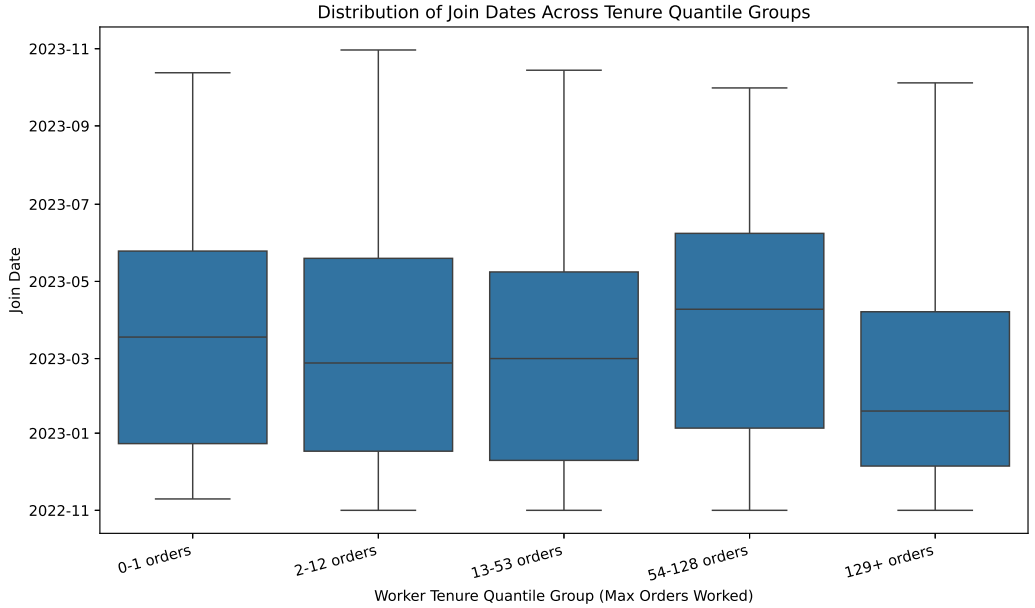


Fig. 12. Distribution of Worker Join Dates Across Tenure Quantile Groups. This figure shows the distribution of worker join dates within each of the five tenure quantile groups, where tenure is defined by the maximum number of orders completed during the study period. The boxplots illustrate the median, 25% and 75% quantile range, and overall range of join dates for workers in each group. Tenure quantile groups are labeled as: '0-2 orders', '3-13 orders', '14-54 orders', '55-129 orders', and '130+ orders'.

- **Mean Diff**: The difference between the mean On-time rate of Group 2 and the mean On-time rate of Group 1, calculated using only data from the 0–1 orders completed bin.
  A positive value indicates Group 2 had a higher average OTP in this initial period.
- **p-adj**: The p-value for the pairwise comparison, adjusted for multiple comparisons using the Tukey HSD method.
- **Lower / Upper**: The lower and upper bounds of the 95% confidence interval for the difference in means. If this interval does not contain zero, the difference is statistically significant at the $\alpha = 0.05$ level.
- **Reject**: Indicates whether the null hypothesis (that the two group means are equal) should be rejected. (True = significant difference; False = no significant difference).

Key findings from this table indicate that, even within the first two orders completed: workers who eventually achieved the highest tenure (129+ orders) had a significantly higher mean on-time rate compared to those in the lowest tenure group (0–1 orders), the 13–53 orders group, and the 54–128 orders group. Interestingly, the 2–12 orders group also showed a significantly higher mean on-time rate during this initial 0-1 order period compared to the 13–53 orders group. Other pairwise comparisons did not show statistically significant differences in on-time rate during this very early stage. These results suggest that differences in performance trajectories between workers who ultimately achieve different levels of tenure manifest very early in their engagement with the platform.

Table 3 provides the corresponding pairwise comparisons for the subsequent 2–12 orders period. During this phase, the 129+ orders group continued to significantly outperform the 13–53 orders

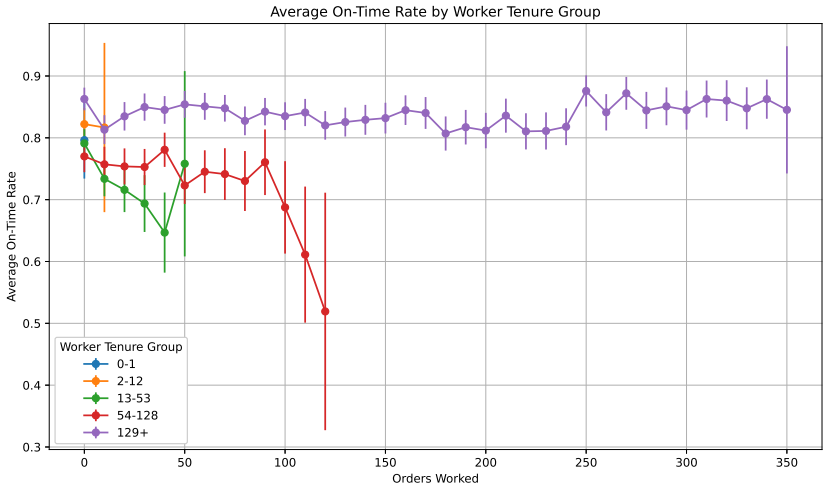


Fig. 13. Average On-Time Rate by Worker Tenure Group (Restricted Cohort). This figure plots the average on-time delivery rate (Y-axis) against the cumulative number of orders worked (X-axis, binned) for workers who joined the platform during the first half of the study period (N=690 workers total in this cohort). Separate lines represent workers belonging to different eventual tenure groups, defined by quantiles of the maximum total orders completed over the full one-year study period (0-1 orders: N=117, 2-12 orders: N=151, 13-53 orders: N=141, 54-128 orders: N=124, 129+ orders: N=157). Error bars indicate 95% confidence intervals for the mean on-time rate within each X-axis bin for each group. This analysis is restricted to a specific join cohort to mitigate potential confounding effects of join date when comparing performance trajectories across eventual tenure groups.

Table 2. Tukey HSD Test Results for Mean On-Time Rate by Worker Tenure Group (Period: 0–1 Orders). ANOVA F-statistic = 8.114, p-value = 1.745×10^{-6}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 orders	129+ orders	0.1123	0.0010	0.0334	0.1912	True
0–1 orders	13–53 orders	-0.0255	0.9010	-0.1037	0.0528	False
0–1 orders	2–12 orders	0.0661	0.1798	-0.0159	0.1481	False
0–1 orders	54–128 orders	-0.0057	0.9997	-0.0845	0.0732	False
129+ orders	13–53 orders	-0.1378	<0.001	-0.2149	-0.0606	True
129+ orders	2–12 orders	-0.0462	0.5255	-0.1272	0.0348	False
129+ orders	54–128 orders	-0.1180	0.0003	-0.1958	-0.0402	True
13–53 orders	2–12 orders	0.0916	0.0162	0.0112	0.1719	True
13–53 orders	54–128 orders	0.0198	0.9563	-0.0573	0.0969	False
2–12 orders	54–128 orders	-0.0718	0.1102	-0.1527	0.0092	False

and 54–128 orders groups. The 2–12 orders group also maintained significantly higher performance compared to both the 13–53 orders and 54–128 orders groups.

Table 4 shows the pairwise comparisons for during the 13–53 orders completed period. During this stage, significant performance differences were observed between all three groups compared:

Table 3. Tukey HSD Test Results for Mean On-time Rate by Worker Tenure Group (Period: 2–12 Orders). ANOVA F-statistic = 13.715, p-value = 3.811×10^{-11}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.0751	<0.001	-0.1080	-0.0421	True
129+ orders	2–12 orders	-0.0126	0.9233	-0.0543	0.0291	False
129+ orders	54–128 orders	-0.0689	<0.001	-0.1019	-0.0358	True
13–53 orders	2–12 orders	0.0625	0.0004	0.0207	0.1042	True
13–53 orders	54–128 orders	0.0062	0.9863	-0.0269	0.0393	False
2–12 orders	54–128 orders	-0.0563	0.0022	-0.0981	-0.0145	True

the highest tenure group (129+ orders) significantly outperformed both the 13–53 orders and 54–128 orders groups, and the 54–128 orders group significantly outperformed the 13–53 orders group.

Table 5 presents the comparison for the 54–128 orders completed period. Within this experience range, the only possible comparison between the defined tenure groups confirms that the highest tenure group (129+ orders) continued to significantly outperform the 54–128 orders group. Note that no Tukey HSD test is performed here since we only have two groups.

Table 4. Tukey HSD Test Results for Mean On-time Rate by Worker Tenure Group (Period: 13–53 Orders). ANOVA F-statistic = 106.590, p-value = 1.833×10^{-68}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.1288	<0.001	-0.1494	-0.1081	True
129+ orders	54–128 orders	-0.0928	<0.001	-0.1100	-0.0756	True
13–53 orders	54–128 orders	0.0359	<0.001	0.0153	0.0566	True

Table 5. Mean Diff of On-time Rate by Worker Tenure Group (Period: 54–128 Orders). ANOVA F-statistic = 196.957, p-value = 1.858×10^{-85}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	54–128 orders	-0.1206				True

C.2 Statistical Testing for Bundle Volume Difference

The subsequent tables detail pairwise comparisons of mean bundle volume across tenure groups for different early order periods, following significant ANOVAs.

Table 6 presents comparisons for the initial 0–1 orders completed period (ANOVA F = 7.828, $p < 0.001$). During this very early stage, significant differences ($\alpha = 0.05$) in mean bundle volume were observed where: the highest tenure group (129+ orders) bundled significantly more than the 0–1, 2–12, and 54–128 orders groups; additionally, the 13–53 orders group bundled significantly more than the 0–1 orders group.

Table 7 shows comparisons for the 2–12 orders completed period (ANOVA F = 37.250, $p < 0.001$). Significant differences ($\alpha = 0.05$) indicate: the '129+ orders' group bundled more than all other groups compared; the 13–53 orders group bundled more than the 2–12 orders group; and the 54–128 orders group bundled more than the 2–12 orders group.

Table 6. Tukey HSD Test Results for Mean Bundle Volume by Worker Tenure Group (Period: 0–1 Orders). ANOVA F-statistic = 7.828, p-value = 2.961×10^{-6}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 orders	129+ orders	0.1607	<0.001	0.0792	0.2422	True
0–1 orders	13–53 orders	0.0967	0.0096	0.0160	0.1775	True
0–1 orders	2–12 orders	0.0515	0.4594	-0.0332	0.1362	False
0–1 orders	54–128 orders	0.0779	0.0683	-0.0035	0.1594	False
129+ orders	13–53 orders	-0.0639	0.1834	-0.1436	0.0157	False
129+ orders	2–12 orders	-0.1092	0.0034	-0.1929	-0.0256	True
129+ orders	54–128 orders	-0.0828	0.0397	-0.1631	-0.0024	True
13–53 orders	2–12 orders	-0.0453	0.5692	-0.1282	0.0377	False
13–53 orders	54–128 orders	-0.0188	0.9676	-0.0984	0.0608	False
2–12 orders	54–128 orders	0.0265	0.9099	-0.0571	0.1100	False

Table 7. Tukey HSD Test Results for Mean Bundle Volume by Worker Tenure Group (Period: 2–12 Orders). ANOVA F-statistic = 37.250, p-value = 6.784×10^{-31}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.1580	<0.001	-0.2263	-0.0897	True
129+ orders	2–12 orders	-0.3671	<0.001	-0.4535	-0.2807	True
129+ orders	54–128 orders	-0.1369	<0.001	-0.2054	-0.0684	True
13–53 orders	2–12 orders	-0.2091	<0.001	-0.2956	-0.1226	True
13–53 orders	54–128 orders	0.0211	0.9186	-0.0475	0.0897	False
2–12 orders	54–128 orders	0.2302	<0.001	0.1435	0.3169	True

Table 8. Tukey HSD Test Results for Mean Bundle Volume by Worker Tenure Group (Period: 13–53 Orders). ANOVA F-statistic = 35.257, p-value = 1.022×10^{-22}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.2317	<0.001	-0.2943	-0.1690	True
129+ orders	54–128 orders	-0.1465	<0.001	-0.1987	-0.0943	True
13–53 orders	54–128 orders	0.0851	0.0026	0.0226	0.1477	True

Table 9. Mean Difference of Bundle Volume by Worker Tenure Group (Period: 54–128 Orders). ANOVA F-statistic = 13.840, p-value = 9.85×10^{-7}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	54–128 orders	-0.1069				True

Table 8 details comparisons for the 13–53 orders completed period (ANOVA F = 35.257, p < 0.001). Significant differences ($\alpha = 0.05$) were found where: the 129+ orders group bundled more than both the 13–53 and 54–128 orders groups; and the 54–128 orders group bundled more than the 13–53 orders group.

Finally, Table 9 presents the comparison for the 54–128 orders completed period (ANOVA F = 13.840, p < 0.001). The only possible comparison within this range confirms that the 129+ orders group bundled significantly more than the 54–128 orders group ($\alpha = 0.05$).

C.3 Statistical Testing for Platform-Bundled Percentage

Table 10. Tukey HSD Test Results for Platform-Bundled Percentage by Worker Tenure Group (Period: 0–1 Orders). ANOVA F-statistic = 4.429, p-value = 1.45×10^{-3}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 orders	129+ orders	0.0387	<0.001	0.0123	0.0650	True
0–1 orders	13–53 orders	0.0199	0.2305	-0.0063	0.0461	False
0–1 orders	2–12 orders	0.0210	0.2236	-0.0064	0.0484	False
0–1 orders	54–128 orders	0.0103	0.8246	-0.0161	0.0367	False
129+ orders	13–53 orders	-0.0188	0.2738	-0.0446	0.0070	False
129+ orders	2–12 orders	-0.0176	0.3868	-0.0447	0.0095	False
129+ orders	54–128 orders	-0.0284	0.0245	-0.0544	-0.0024	True
13–53 orders	2–12 orders	0.0011	1.0000	-0.0257	0.0280	False
13–53 orders	54–128 orders	-0.0096	0.8471	-0.0354	0.0162	False
2–12 orders	54–128 orders	-0.0107	0.8154	-0.0378	0.0163	False

Table 11. Tukey HSD Test Results for Platform-Bundled Percentage by Worker Tenure Group (Period: 2–12 Orders). ANOVA F-statistic = 36.265, p-value = 4.566×10^{-30}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.0744	<0.001	-0.1033	-0.0454	True
129+ orders	2–12 orders	-0.1483	<0.001	-0.1850	-0.1117	True
129+ orders	54–128 orders	-0.0794	<0.001	-0.1084	-0.0503	True
13–53 orders	2–12 orders	-0.0739	<0.001	-0.1106	-0.0373	True
13–53 orders	54–128 orders	-0.0050	0.9904	-0.0340	0.0241	False
2–12 orders	54–128 orders	0.0690	<0.001	0.0322	0.1057	True

Table 12. Tukey HSD Test Results for Platform-Bundled Percentage by Worker Tenure Group (Period: 13–53 Orders). ANOVA F-statistic = 132.465, p-value = 5.608×10^{-85}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	-0.1511	<0.001	-0.1752	-0.1269	True
129+ orders	54–128 orders	-0.1364	<0.001	-0.1565	-0.1162	True
13–53 orders	54–128 orders	0.0147	0.4001	-0.0095	0.0388	False

The tables report post-hoc Tukey HSD comparisons of the average platform-bundled percentage across tenure groups within distinct order-completion intervals. All tests follow a significant one-way ANOVA conducted at each interval and control for multiple comparisons using the family-wise error rate ($\alpha = 0.05$).

Table 13. Mean Difference of Platform-Bundled Percentage by Worker Tenure Group (Period: 54–128 Orders). ANOVA F-statistic = 140.542, p-value = 2.372×10^{-61}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	54–128 orders	-0.1277				True

Table 10 shows comparisons for the initial 0–1 orders completed (ANOVA F = 4.429, p < 0.01). During this earliest stage, the only statistically significant pairwise differences indicate that workers who ultimately reached 129+ orders had a higher average platform-bundled percentage than both the 0–1 and 54–128 groups. No other group differences were significant.

In the 2–12 orders period (Table 11, ANOVA F = 36.265, p < 0.001), a more pronounced pattern of significant differences emerges. The 129+ orders group had significantly higher platform-bundled percentages than all other tenure groups. The 13–53 group also significantly outperformed the 2–12 group, while the 54–128 group bundled more than the 2–12 group.

Table 12 summarizes results for the 13–53 orders completed period (ANOVA F = 132.465, p < 0.001). At this stage, the 129+ group remained significantly above both the 13–53 and 54–128 groups in terms of platform-bundled allocation. However, no significant differences emerged between the 13–53 and 54–128 groups.

Finally, Table 13 displays the comparison for the 54–128 orders completed period (ANOVA F = 140.542, p < 0.001). The 129+ group continued to receive significantly more platform-bundled tasks compared to the 54–128 group.

C.4 Statistical Testing for Percentage of Self-Bundled Orders

The following tables present Tukey HSD post-hoc test results comparing the percentage of self-bundled orders across worker tenure groups, within specific early experience bins. Each table follows a one-way ANOVA that identified significant overall differences between groups ($\alpha = 0.05$).

Table 14 shows results for the first 0–1 orders completed period (ANOVA F = 5.183, p < 0.001). At this very early stage, workers in the 129+, 13–53, and 54–128 order groups exhibited significantly higher self-bundled percentages than those in the 0–1 group. No other comparisons reached statistical significance.

In the 2–12 orders bin (Table 15, ANOVA F = 10.110, p < 0.001), significant differences were observed across several groupings. Specifically, the 129+ group has a lower share of self-bundled orders than the 54–128 group. Additionally, the 54–128 group has more self bundle percentage than the 2–12 group, and the 13–53 group has more self bundle percentage than the 2–12 group significantly.

Table 16 presents results for the 13–53 orders completed segment (ANOVA F = 46.612, p < 0.001). At this stage, workers in the 129+ group has significantly less self-bundled orders than both the 13–53 and 54–128 groups. The 13–53 group also showed a significantly lower self bundling rate than the 54–128 group.

Finally, Table 17 presents results for the 54–128 orders segment (ANOVA F = 84.727, p < 0.001). The 129+ group still has a lower self bundle percentage than the 54–128 group.

D Full Independent Variables in the MNL Model in Section 6

- **LIST**: Indicates whether the order is in the recommendation list by the platform's algorithm (1) or it is not (0).
- **BUNDLED**: Indicates whether the order is part of a bundle (1) or not (0).

Table 14. Tukey HSD Test Results for Self-Bundled Percentage by Worker Tenure Group (Period: 0–1 Orders). ANOVA F-statistic = 5.183, p-value = 3.747×10^{-4}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 orders	129+ orders	0.0576	0.0198	0.0060	0.1092	True
0–1 orders	13–53 orders	0.0619	0.0086	0.0108	0.1131	True
0–1 orders	2–12 orders	0.0124	0.9697	-0.0412	0.0661	False
0–1 orders	54–128 orders	0.0651	0.0053	0.0135	0.1166	True
129+ orders	13–53 orders	0.0043	0.9993	-0.0461	0.0548	False
129+ orders	2–12 orders	-0.0452	0.1365	-0.0981	0.0078	False
129+ orders	54–128 orders	0.0075	0.9945	-0.0434	0.0584	False
13–53 orders	2–12 orders	-0.0495	0.0761	-0.1020	0.0031	False
13–53 orders	54–128 orders	0.0032	0.9998	-0.0473	0.0536	False
2–12 orders	54–128 orders	0.0526	0.0522	-0.0003	0.1056	False

Table 15. Tukey HSD Test Results for Self-Bundled Percentage by Worker Tenure Group (Period: 2–12 Orders). ANOVA F-statistic = 10.110, p-value = 3.684×10^{-8}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	54–128 orders	0.0476	<0.001	0.0193	0.0759	True
13–53 orders	2–12 orders	-0.0445	0.0062	-0.0803	-0.0088	True
2–12 orders	54–128 orders	0.0705	<0.001	0.0347	0.1063	True
129+ orders	13–53 orders	0.0216	0.2237	-0.0066	0.0499	False
129+ orders	2–12 orders	-0.0229	0.4044	-0.0586	0.0128	False
13–53 orders	54–128 orders	0.0260	0.0908	-0.0024	0.0543	False

Table 16. Tukey HSD Test Results for Self-Bundled Percentage by Worker Tenure Group (Period: 13–53 Orders). ANOVA F-statistic = 46.612, p-value = 5.21×10^{-30}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	13–53 orders	0.0438	<0.001	0.0228	0.0649	True
129+ orders	54–128 orders	0.0807	<0.001	0.0631	0.0982	True
13–53 orders	54–128 orders	0.0368	<0.001	0.0158	0.0579	True

Table 17. Mean Difference of Self-Bundled Percentage by Worker Tenure Group (Period: 54–128 Orders). ANOVA F-statistic = 84.727, p-value = 2.259×10^{-37}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
129+ orders	54–128 orders	0.0894				True

- **Past Frequency**: This variable represents the proportion of orders that a worker has completed in the store relative to the total number of orders completed across all stores, measured prior to selecting the current order. It is set to 0 if no orders have been completed in the current store.

- **ORDER_TYPE_ID**: The type of order associated with the alternative (e.g., delivery or pickup).
- **MILES_DISTANCE_STORE_CUST**: The distance (in miles) between the store and the customer's location.
- **REQUESTED_ITEMS**: The number of items requested in the order.
- **DOLLARS_BONUS**: The dollar amount of any bonuses offered for completing the order.
- **DOLLARS_PAY**: The total payment offered for completing the order, excluding bonuses.
- **LOCAL_DELIVERY_WINDOW**: The delivery time window of the order.
- **PCT_DAILY_NONFOOD_ITEMS**: The percentage of daily non-food items in the alternative (rounded to 0, 0.5, 1 due to data privacy) .
- **PCT_EXPANDED_FOOD_ITEMS**: The percentage of expanded food items in the alternative (rounded to 0, 0.5, 1 due to data privacy).
- **PCT_GENERAL_MERCH_ITEMS**: The percentage of general merchandise items in the alternative (rounded to 0, 0.5, 1 due to data privacy).
- **MAX_CONTRIBUTIONS_CAT_PCT**: The percentage of the main item category (continuous from 0 to 1, this serves as a proxy of how distributed items are in the order).