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

SHUNAN JIANG, University of California, Berkeley, USA
WICHINPONG PARK SINCHAISRI, Haas School of Business, University of California, Berkeley, USA

As gig economy platforms increasingly rely on algorithms to manage on-demand workers, understanding how algorithmic recommendations influence worker behavior is critical for optimizing platform design and improving worker experience. This paper examines the dynamic interactions between gig workers and platform algorithms, focusing on how workers learn to refine their strategies and performance over time. Using multiple quantitative methods, including two-way fixed effects regression and multinomial logit modeling, we analyze more than a million orders completed by gig workers on a retail delivery platform. Our findings reveal a clear learning curve: workers progressively improve their efficiency and on-time delivery performance with experience. Newcomers rely heavily on algorithmic recommendations for task selection, but 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 need for adaptive, human-centric systems that evolve with workers' learning trajectories, incorporate their feedback, and offer flexibility to support personalized strategies to enhance collaboration and outcomes for both workers and platforms.

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

Additional Key Words and Phrases: gig economy, on-demand work, worker learning, recommendation algorithm, human–AI collaboration, 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. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW440 (November 2025), 35 pages. https://doi.org/10.1145/3757621

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 complex decision-making challenges, especially as the volume and diversity of tasks grow. Without traditional support networks of colleagues, supervisors, or mentors, gig workers must independently navigate these challenges, often learning through trial and error [40].

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 combining multiple orders into a single trip to streamline routes, reduce idle time, and increase earnings. However, these platform-provided recommendations introduce new layers of complexity, requiring workers to integrate them with their own strategies. Navigating this balance

Authors' Contact Information: Shunan Jiang, University of California, Berkeley, Berkeley, California, USA, shunan_jiang@berkeley.edu; Wichinpong Park Sinchaisri, Haas School of Business, University of California, Berkeley, Berkeley, California, USA, parksinchaisri@berkeley.edu.

ACM 2573-0142/2025/11-ARTCSCW440
https://doi.org/10.1145/3757621

can be difficult, leading to misaligned decisions that reduce performance or result in suboptimal results. For example, food delivery platform workers often reject bundled orders due to the increased complexity and effort required [19].

When workers have the autonomy to create their own task bundles, they may overestimate their capacity or overlook the logistical challenges of complex deliveries. This can lead to 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 platforms' bundling and task recommendation algorithms? (iii) How does workers' decision-making evolve as they accumulate experience with the platform? Our findings offer insights for designing recommendation systems that better support worker autonomy and performance, ultimately fostering more effective collaboration between platform algorithms and worker strategies.

We adopt multiple quantitative methods to examine 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 improve performance through experience. We then conduct a descriptive analysis to examine how workers learn to bundle tasks with platform recommendations and how this interaction influences their performance. Additionally, to analyze workers' task selection behaviors, we employ a multinomial logit (MNL) model to capture how workers respond to platform recommendations and explore new stores as they gain experience. This methodological framework provides a comprehensive view of worker strategies, illuminating how workers co-adapt with platform algorithms to optimize performance over time.

Our findings reveal several key insights into how workers adapt to algorithmic systems. First, workers show a clear learning curve, with significant gains in efficiency and on-time delivery as they gain experience. Regression analysis indicates that store-specific experience plays a crucial role in improving performance, while skills acquired from other stores also contribute to gains. These transferable skills, such as navigating store layouts and managing customer expectations, enable workers to adapt more effectively across contexts, underscoring the value of cross-context learning.

When choosing which orders to accept, workers can either follow platform recommendations or select from a pool of available orders. Our findings show that newer workers tend to rely more on platform suggestions, whereas experienced workers develop their own strategies and increasingly deviate from algorithmic recommendations. This progression illustrates how workers gradually build confidence and refine their task selection strategies, leading to improvements in both performance and earnings. Our results suggest that as workers gain proficiency, platforms should adapt their algorithms to provide greater flexibility, enabling experienced workers to align task selection with their evolving strategies. Incorporating worker feedback mechanisms could further personalize recommendations, ensuring suggestions remain relevant and responsive to changing needs.

While prior research on gig worker learning often uses behavioral proxies, such as the rate of visiting new areas, to identify 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. Through this lens, we find a 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 underscore the value of human-centric recommendation systems that evolve alongside 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. Such an 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. Section 4 presents empirical evidence on how workers improve performance through experience. Section 5 examines how workers adapt to the platform's recommendations for bundling tasks. Section 6 analyzes how workers select tasks with the platform's recommendation algorithms and discusses implications for recommendation algorithm design. Section 7 discusses the broader implications of our findings. Finally, Section 8 offers concluding remarks.

2 Related Works and Contributions

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

2.1 Human-Algorithm Interactions at Work

In this subsection, we focus on the first stream, examining how digital platforms shape worker behavior and performance through algorithmic management and recommendation systems. A central theme in this literature is the trade-off between worker autonomy, engagement, and the structure imposed by algorithmic management.

Several studies have shown how platform features, feedback channels, and algorithmic systems shape performance, satisfaction, and autonomy. 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 facilitated 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 to increase worker engagement [13], and recent work emphasizes how design features that prioritize well-being can strengthen worker–platform relationships [37]. Algorithmic management can reshape power dynamics and compel workers to develop new interpretive skills for navigating data-driven systems [23, 24], while perceptions of fairness, trust, and emotional response play a critical role in determining how workers engage with algorithmic decisions [28]. More broadly, data-driven systems have been found to influence worker autonomy and job satisfaction [29]. Taken together, these findings suggest that engagement with algorithmic systems depends not only on task structures and interface features but also on trust, perceived fairness, and opportunities for self-directed learning.

Building on these themes, 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]. 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 recommendations effectively into their workflows, even when they are open to them [5]. Concerns about surveillance and control can also drive resistance to passive systems [11, 12]. In contrast, tools that give workers autonomy to track their own performance have been proposed to increase transparency and accountability [16], and participatory design or collective action strategies have been suggested to create more empowering, worker-centric platform environments [32, 35].

These adoption challenges are particularly relevant in recommender systems, where user acceptance interacts with algorithmic trade-offs such as exploration and exploitation. The exploration–exploitation (E&E) dilemma involves the fundamental choice between exploiting known user preferences to maximize immediate satisfaction and exploring new or uncertain options 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]. In gig work, this can involve decisions such as whether to accept a familiar delivery route or try a new store to expand future opportunities.

Our study builds on this literature by examining how gig workers engage with task and bundle recommendations over time. Although 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, particularly in settings where algorithmic recommendations are central to daily decision making. Using longitudinal behavioral data from thousands of workers, we analyze how workers initially respond to recommendations, how their behavior evolves over time, and how this adaptation affects long-term performance. In doing so, we contribute to a deeper understanding of human–algorithm collaboration in gig economy settings and demonstrate how discrete choice modeling can capture the dynamics of this adaptation.

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, provide theoretical foundations for understanding how workers update strategies over time in response to performance feedback [7]. However, the dynamics of learning in gig work differ from these traditional contexts: platform-mediated environments often lack stable teams, consistent workflows, and face-to-face feedback, requiring workers to self-direct their learning.

As the gig economy has grown, scholars have examined 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 shape both productivity and service quality [21]. Research on early-stage gig work behavior shows that workers tend to explore unfamiliar regions at first, an approach that 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 informed the development of 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 examine how heterogeneous strategies emerge and lead to divergent learning paths. Although prior studies have examined either short-term performance effects or the design of order batching algorithms, little is known about how workers' decision-making strategies evolve with experience and how these strategies interact with platform recommendations over time. Using a large-scale dataset and a multinomial logit framework, we analyze how workers respond to platform recommendations, how their task selection behavior evolves, and how that evolution affects long-term performance. Our approach enables us to capture fine-grained choice patterns over time while quantifying how platform guidance shapes learning trajectories.

3 Empirical Context: Online Retail Delivery Platform

We collaborate with an on-demand retail delivery company (hereafter referred to as "the company" or "the platform") to analyze a comprehensive dataset of online retail orders completed in New York City over a 364-day period, from November 2022 to October 2023. The dataset contains detailed information on completed orders, order characteristics, and productivity metrics such as time spent shopping, checkout time, and driving time. It also includes evaluations for each completed order, such as whether the delivery was on time.

A key strength 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, that is, orders 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 on orders that the platform bundled together for simultaneous delivery.

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

3.1 Platform Overview

The company operates 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 service by matching customers with gig workers, who visit physical retail stores, hand-pick the ordered items, and maintain real-time communication with customers via integrated chat. After shopping, gig workers deliver the items directly to customers' addresses.

Gig workers are compensated on a per-order basis, with pay determined by factors such as order size, complexity, and delivery distance. In addition to base earnings, workers may receive customer tips and platform-issued bonuses for meeting 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 complete a screening process that verifies eligibility criteria such as being at least 18 years old, possessing a valid driver's license, and having access to a vehicle. Upon approval, workers are granted access to the gig platform and can set 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 see 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 potential alignment with the worker's profile or opportunities for 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, displays 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 have full flexibility to evaluate a menu of available tasks and make informed decisions based on personal preferences, operational constraints, and expected earnings. This design encourages strategic behavior in task selection and enables 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 attributes such as store origin, item composition, and destination proximity. If a worker accepts a platform-generated bundle, they must fulfill both orders together. These system-generated bundles always contain exactly two orders. Second, workers may *self-bundle* by manually selecting and sequencing multiple orders, recommended or not, for concurrent fulfillment. There is no platform-imposed limit on the number of orders that can be self-bundled, and workers retain full discretion in deciding whether and how to do so.

This paper examines 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, whether by accepting platform-generated bundles or creating self-bundles. These decisions provide insight into how workers learn and adapt in algorithmically mediated labor settings and illuminate 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. Worker engagement is highly heterogeneous: some individuals completed only a single order before leaving the platform, while others fulfilled more than 6,000 orders during the 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.

The volume of orders at the store level also exhibits substantial variation. Some stores processed only a single order, while 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, and approximately 60% of all orders were delivered as part of a bundle.

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) trip records, which provide detailed trip-level data for taxi and ride-hailing services in New York City (NYC). These records include pickup and drop-off locations, timestamps, trip distances, fares, and payment methods, encompassing millions of rides over multiple years. From this dataset, we derive two key traffic-related proxies: (i) hourly traffic volume and (ii) the average hourly taxi speed, 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 wide range of parameters, including temperature, humidity, wind speed, and precipitation, as well as specialized metrics such as 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 retained three parameters: apparent temperature, rainfall, and wind speed, for inclusion in our regression analyses.

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

In this section, we analyze how gig workers 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.

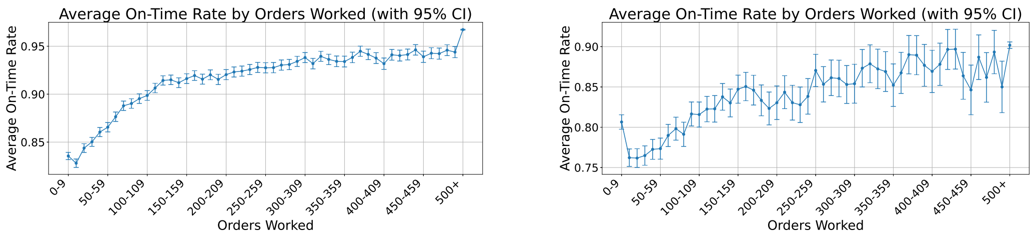
We begin with model-free descriptive evidence to document performance trends, followed by an empirical strategy using two-way fixed effects (2FE) regression models that control for time-invariant differences across workers and stores. These analyses form the empirical foundation for the next section, where we examine how workers 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). Panel (a) shows trends for all workers active during the study period ($N = 5,292$), while Panel (b) focuses on newcomers ($N = 1,131$) who joined the platform after November 1, 2022.

Comparing the two populations reveals distinct patterns. In Panel (a), all workers exhibit a relatively high initial on-time rate and a gradual upward trend. In contrast, newcomers in Panel (b) start with lower average performance but display a steeper early learning curve. This difference is expected: the full-sample trend averages across workers with varied tenure, including many already experienced at the start of the study. It is also influenced by survivorship bias, as lower-performing workers are more likely to exit early and are thus underrepresented at higher experience levels.

To address these confounds, our regression analysis in Section 4 focuses on the newcomer cohort. Because these workers entered during the observation window, we observe their complete experience trajectory from platform entry, enabling cleaner identification of learning patterns. Our



(a) All Workers

(b) New Workers

Fig. 1. Average on-time delivery probability by cumulative orders completed. Panel (a) includes all workers active during the study period (N = 5,292). Panel (b) focuses on newcomers who joined after November 1, 2022 (N = 1,131). Cumulative orders are grouped into bins of 10 (e.g., 0–9, 10–19) up to 500+, and error bars indicate 95% confidence intervals.

two-way fixed effects model with worker-specific intercepts further controls for time-invariant heterogeneity and helps mitigate survivorship bias.

Within the newcomer cohort (Panel b), performance is lowest in the earliest stages. Notably, there is a dip between the first (0–9 orders) and second (10–19 orders) experience bins. This could reflect: (i) experienced workers briefly testing the platform in their first few orders, artificially boosting the first bin; or (ii) adaptation challenges and increased task complexity, such as the introduction of bundled orders, in early tenure. Beyond this point, performance rises steadily through roughly the first 350 orders before plateauing and becoming noisier, suggesting diminishing returns to experience.

4.2 Two-Way Fixed Effects Regression Analysis of Worker Performance

To quantify the relationship between gig worker experience and performance, we estimate two-way fixed effects (2FE) regressions for each of our two outcome variables: the on-time delivery indicator ($OnTime$) and the number of items picked per hour ($ItemsPerHour$) [39]. Restricting the sample to newcomers enables us to observe their complete learning trajectory from platform entry, mitigating survivorship and left-censoring biases.

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

where:

- $PerformanceMetric_{ist}$ is either the on-time delivery outcome or the number of items picked by worker i when shopping at store s at time t.
- OTS_{ist} is the number of prior orders completed by worker i at store s by time t (within-store experience), and OOS_{ist} is the number of prior orders completed at all other stores (cross-store experience).
- OTS_{ist}^2 and OOS_{ist}^2 capture potential nonlinearities in the returns to experience.
- $\mathbf{X}_{ist}$ is a vector of time-varying controls, including:
  - *Weather conditions*: temperature, rainfall, wind speed
  - *Order characteristics*: total payment, bonuses, requested item quantities, delivery distance

  - *Traffic conditions*: hourly taxi volume, average taxi speed
- μ_i: worker fixed effects, capturing all time-invariant worker-specific characteristics (e.g., innate skill, motivation).
- δ_s: store fixed effects, capturing persistent store characteristics (e.g., layout, size, location).
- γ_t: time fixed effects, capturing day-of-week, seasonal, or macro shocks affecting performance.
- ϵ_{ist}: idiosyncratic error term.

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 measure aligns with how the platform evaluates service quality and avoids the noise and skew that often affect continuous delivery-time metrics. The second dependent variable, $ItemsPerHour$, is computed as the total number of items picked during the shopping process divided by the time spent in-store, serving as a proxy for worker productivity.

Independent variables. We focus on two primary independent variables that capture different dimensions of a gig worker's accumulated experience: *OTS* (*Orders This Store*) and *OOS* (*Orders Other Stores*).

OTS measures the number of prior deliveries a worker has completed at a particular store, reflecting store-specific familiarity with its layout, inventory systems, and staff routines. We hypothesize that repeated exposure to the same store improves performance; for example, by enabling faster item location, reducing picking errors, fostering rapport with store employees, and supporting more efficient route planning both inside and outside the store. To capture potential nonlinearities in this relationship and allow for diminishing or accelerating returns to store-specific experience, we include the squared term OTS^2.

OOS measures the number of prior deliveries a worker has completed at all other stores, excluding the focal one. This variable captures broader cross-store learning that may improve performance through generalizable skills such as workflow optimization, adaptability, or task management. As with OTS, we include OOS^2 to capture nonlinear effects of broader experience.

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

- *Order characteristics*: total payout, bonuses, item quantities, store-to-customer distance, delivery time window length, and total order value. These variables capture order complexity, incentives, and the possibility that lower-priced orders may involve more small, low-value items, potentially inflating the $ItemsPerHour$ metric.
- *Traffic conditions*: hourly taxi volume and average taxi speed in New York City, serving as proxies for time-varying urban congestion.
- *Weather conditions*: apparent temperature, precipitation, and wind speed, which may affect travel time and worker comfort.
- *Time fixed effects*: day-of-week and calendar-month dummies to control for temporal fluctuations in demand, congestion, and worker availability.
- *Worker–store fixed effects*: capture persistent heterogeneity in performance specific to a worker–store pair, such as skill, local knowledge, or route familiarity.

4.3 Results: *Diminishing Positive Return on Experience*

Table 1 reports 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	$6.06 \times 10^{-5***}$	$9.43 \times 10^{-3***}$
	(2.00×10^{-5})	(3.53×10^{-3})
OTS^2	$-1.75 \times 10^{-8***}$	$-5.47 \times 10^{-6**}$
	(1.05×10^{-8})	(1.85×10^{-6})
OOS	$5.91 \times 10^{-5***}$	$6.34 \times 10^{-3*}$
	(1.58×10^{-5})	(2.78×10^{-3})
OOS^2	$-8.87 \times 10^{-9***}$	-7.63×10^{-7}
	(3.64×10^{-9})	(6.42×10^{-7})
Fixed effects	✓	✓
Weather controls	✓	✓
Traffic controls	✓	✓
R^2	0.029	0.013
Observations	105,543	105,543

Notes: Standard errors in parentheses. Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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 show that gig workers improve their performance with experience, but the pace and pattern of this improvement vary considerably across individuals. In this section, we examine this heterogeneity by analyzing how workers differ in their learning trajectories and in their responses to platform-generated task recommendations, specifically, recommendations to bundle multiple orders.

We begin by segmenting workers according to their overall tenure on the platform, revealing distinct performance trajectories across groups. We then investigate how workers respond to bundling recommendations and whether these responses are associated with improved operational outcomes. While workers may continue to adapt well beyond their initial period on the platform, we focus on the first 350 orders. This window captures the phase in which platform-relevant performance measures, such as service quality and in-store productivity, exhibit the most pronounced changes.

Although workers may also learn to optimize for personal objectives (e.g., minimizing stress or maximizing earnings per unit of effort), those dimensions fall outside the scope of our analysis. Here, we concentrate on outcomes directly tied 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 *worker tenure groups* based on the quantiles of total orders completed during the study period. These groups reflect different levels of platform engagement: 0–1 orders (261 workers), 2–12 orders (197 workers), 13–53 orders (228 workers), 54–128 orders (221 workers), and 129 or more orders (224 workers). These cutoffs correspond approximately to the 0–20%, 20–40%, 40–60%, 60–80%, and 80–100% percentiles of tenure.

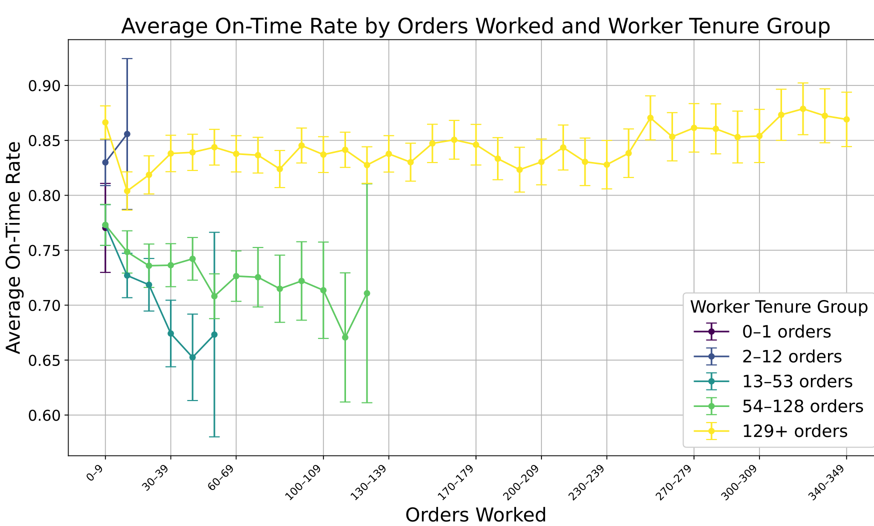


Fig. 2. Average on-time delivery rate by cumulative orders and tenure group

Notes: The 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 the five tenure groups described in the main text. Error bars indicate 95% confidence intervals for the mean within each bin.

Figure 2 shows the average on-time delivery probability for each tenure group, focusing on the first 350 orders. Several patterns emerge. Workers in the highest tenure group begin with relatively high on-time rates and improve steadily over time. In contrast, workers in lower tenure groups start with lower performance and display greater volatility, particularly during their first 50-100 orders. These differences suggest that workers who ultimately remain longer on the platform may follow more consistent learning trajectories or adopt more effective strategies early in their tenure.

The observed heterogeneity may stem from differences in motivation, operational skills, or prior experience with similar systems. To avoid masking these differences, our subsequent analyses of bundling behavior (Section 5.2) and responses to platform recommendations (Section 6) are conducted separately for each tenure group.

One potential concern in defining groups by total orders is the confounding effect of join date, since earlier joiners have more time to accumulate orders. To address this, we perform robustness checks in the Appendix that control for join date and compare only workers who entered the platform during the same time windows. The results confirm that the performance differences are not driven solely by timing.

5.2 Model-free Evidence: Learning to Bundle

A central operational feature of the platform is its use of algorithmic bundling: approximately 60% of all orders are generated as bundles by the system. After completing a few initial tasks, workers begin to encounter these bundled orders, which combine two individual deliveries selected by an algorithm based on store origin, item similarity, or proximity of drop-off locations. Bundles appear in the same task selection interface as other orders and are labeled as bundled, but they are not explicitly flagged as platform recommendations. Once accepted, the two component deliveries must be completed together.

In addition to these system-generated bundles, workers can also create their own by selecting multiple individual orders to fulfill concurrently, a practice we refer to as *self-bundling*. Using platform timestamps, we classify self-bundling as cases where a worker's shopping intervals for different orders overlap, indicating that they independently chose to execute them in parallel.

While bundling offers opportunities for efficiency gains, it also introduces additional coordination complexity. As illustrated in Figures 1 and 2, average on-time delivery rates tend to dip during the early stage when bundling first becomes available. This decline may reflect the learning curve associated with managing multiple simultaneous tasks, though it could also be partially explained by survivorship effects if lower-performing workers exit the platform before their performance recovers. Examining how bundling behavior evolves across worker groups therefore provides insight into both learning dynamics and retention patterns.

5.2.1 Overall Bundle Behaviors. Figure 3 plots the average *bundle volume*, defined as the number of orders fulfilled concurrently, for each tenure group during their first 350 orders. Across all groups, bundle volume rises sharply after the first few orders, coinciding with the introduction of platform-generated bundles into the task interface. These algorithmically created bundles are presented as grouped tasks that cannot be accepted separately.

Workers in the highest tenure group (more than 129 total orders, shown in yellow) consistently record the highest bundle volumes across the experience range, peaking at roughly three bundled orders per 100 total orders and maintaining a stable level thereafter. In contrast, early exiters (0–1 and 2–12 total orders, shown in purple and blue) engage only minimally with bundled tasks before leaving the platform. Mid-tier groups (13–53 and 54–128 orders) show moderate uptake, maintaining bundling rates slightly below those of the highest tenure group, especially in the first 30–50 orders.

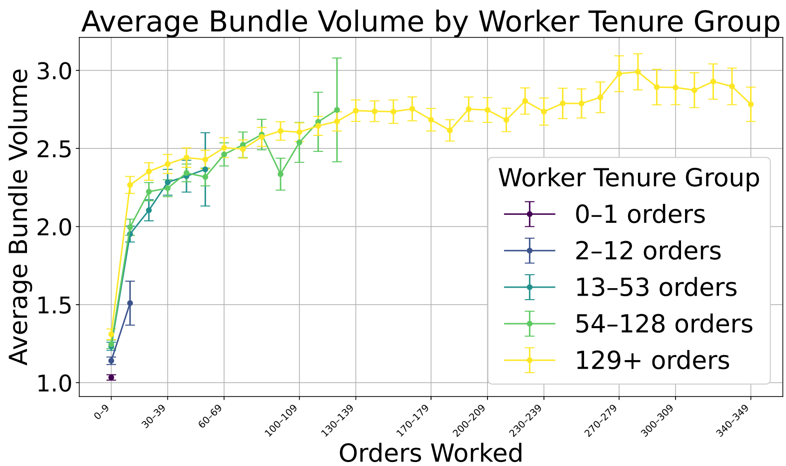


Fig. 3. Average bundle volume across worker groups and over time

Notes: The y-axis shows the average number of orders fulfilled concurrently; the x-axis shows cumulative orders completed (in bins of 10) for the newcomer cohort ($N = 1,131$). Lines represent tenure groups defined by total orders completed during the one-year study period. Error bars denote 95% confidence intervals.

These descriptive patterns indicate that frequent engagement with bundling is more common among workers who remain active on the platform for longer periods. Whether bundling contributes to retention or simply correlates with factors such as familiarity, motivation, or efficiency cannot be determined from this model-free evidence alone. Additional analyses, including statistical comparisons controlling for worker characteristics, are reported in the Appendix.

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

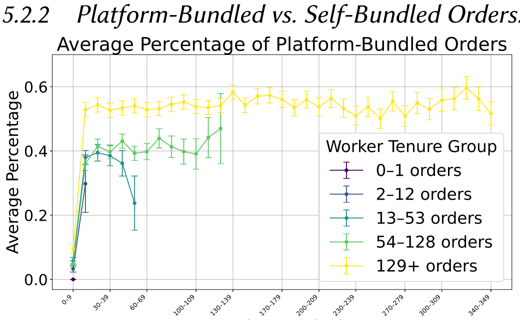


Fig. 4. Platform-bundled orders as a share of all orders

Notes: The y-axis shows the proportion of all completed orders that were part of a platform-generated bundle, plotted by cumulative order bins (x-axis) for different tenure groups. Each proportion is calculated within group–bin pairs. Error bars denote 95% confidence intervals.

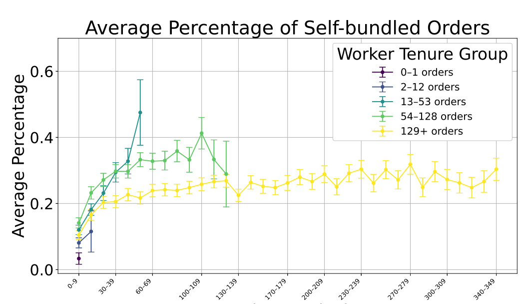


Fig. 5. Self-bundled orders as a share of all orders

Notes: The y-axis shows the proportion of all completed orders that were part of a self-initiated bundle, inferred from overlapping shopping intervals. The x-axis represents cumulative order bins; lines reflect different tenure groups. Error bars denote 95% confidence intervals.

Figure 4 shows the share of completed orders that were part of platform-generated bundles. Adoption rises sharply within the first 20 orders, coinciding with the introduction of bundled tasks into the interface. The highest tenure group (129+ orders) quickly stabilizes above a 50% bundling rate and maintains low within-group variation. Lower tenure groups converge to lower and more volatile adoption rates, typically below 50%.

Figure 5 presents the corresponding share of self-bundled orders. Mid-tenure groups (13–53 and 54–128 orders) record the highest average self-bundling rates, peaking around 30–40%. In contrast, the highest tenure group consistently self-bundles at lower rates, generally under 30%, suggesting greater reliance on platform-generated bundles and less need for manual coordination.

Overall, these trends point to distinct bundling strategies by tenure. All workers encounter bundling early in their tenure, but longer-tenure workers sustain higher use of platform-generated bundles, whereas mid-tenure workers experiment more with self-bundling. This divergence may reflect differences in efficiency, task coordination preferences, or strategic adaptation, and it raises the possibility that consistent reliance on platform bundles is linked to higher retention and performance.

6 Responding to Recommendations: *Orders to Select*

While gig workers can freely choose any available order on the platform, the system 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 appear 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 the 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 quantifies 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 between recommended and non-recommended tasks while conditioning on the full set of options available at the time of decision.

To capture how decision-making changes with experience, we divide each worker's tenure into the same five quantile-based *worker tenure groups* introduced in Section 5: 0–1 orders, 2–12 orders, 13–53 orders, 54–128 orders, and 129+ orders. Within each segment, we include a continuous measure of cumulative experience to capture within-bin variation. This approach maintains comparability with earlier analyses while allowing nonparametric heterogeneity in choice behavior over time. It also provides a clear framework for examining 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

We estimate separate multinomial logit (MNL) models for each of the five experience-based quantile segments defined earlier, allowing for nonparametric heterogeneity in choice behavior over time. For each completed order, the choice set consists of all tasks available to the worker in the one-hour window prior to acceptance, including both algorithmically recommended and non-recommended orders. Each alternative represents either a single delivery or a platform-generated bundle offered as one indivisible unit. Workers typically observe dozens of alternatives in a given choice occasion. All variables are normalized before estimation.

6.1.1 *Description of Key Variables.*

Dependent variable. CHOSEN: A binary indicator equal to 1 if the alternative was selected by the worker, and 0 otherwise.

Independent variables. We focus on order-specific and experience-based predictors; the full list is reported in Appendix D. Each variable enters the deterministic utility component V_{ij} for alternative j faced by worker i.

- *PlatformRecommended*: Equals 1 if the order is in the platform's recommended tab, 0 otherwise.
- *PastFrequency*: Proportion of the worker's prior orders completed at the same store (proxy for store familiarity; 0 if no prior visits).

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. 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. For worker i choosing among alternatives j in a given choice set, the utility is composed of a deterministic component V_{ij} and a stochastic component ϵ_{ij}:

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

where ϵ_{ij} follows a Gumbel distribution, consistent with the multinomial logit (MNL) framework.

The deterministic component V_{ij} is modeled as:

$$V_{ij} = \beta_0 + \sum_{g=1}^{G} \left[\beta_g X_{ijg} + \delta_g \left(X_{ijg} \cdot E_i \right) + \sum_{k=1}^{K-1} \eta_{gk} \left(X_{ijg} \cdot G_{ik} \right) \right], \tag{3}$$

where:

- X_{ijg}: Value of the g^{th} explanatory variable for alternative j faced by worker i (e.g., `PastFrequency`, `PlatformRecommended`).
- E_i: Cumulative number of orders completed by worker i at the time of the decision (`OrdersWorked`).
- G_{ik}: Dummy variable equal to 1 if worker i belongs to tenure group k, and 0 otherwise, with one group serving as the reference category.
- β_g: Coefficient for the main effect of X_{ijg}.
- δ_g: Coefficient for the interaction between X_{ijg} and the worker's continuous experience E_i.
- η_{gk}: Coefficient for the interaction between X_{ijg} and tenure group k.

This formulation allows us to capture (i) baseline effects of each explanatory variable, (ii) how these effects evolve continuously with experience, and (iii) differences across discrete tenure groups, providing a flexible specification for modeling both within-worker learning and cross-worker heterogeneity.

Choice probabilities. Given the utility specification above, the probability that gig worker i chooses alternative j in their choice set is given by the multinomial logit probability function:

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

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

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

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

where y_i denotes the alternative chosen by worker i. We estimate coefficients separately for each of the five cumulative order intervals: 0–1 (excluded due to sparsity), 2–12, 13–53, 54–128, and 129+ orders, allowing preferences to vary flexibly over the worker's tenure. Standard errors are clustered at the worker level.

This structure enables us to examine how choice behavior changes both within workers over time and across workers with different overall experience levels.

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

We estimate four separate multinomial logit (MNL) models, corresponding to the cumulative order ranges 2–12, 13–53, 54–128, and 129+. Workers are likewise segmented into four tenure groups based on their total completed orders, matching these intervals.

Our analysis focuses on two key predictors:

- *PastFrequency*: Measures store familiarity and serves as a proxy for exploitation behavior.
- *PlatformRecommended*: Indicates whether an order appeared in the platform's recommendation tab.

Each MNL model includes interactions between these predictors and (i) the worker's cumulative experience at the time of decision and (ii) tenure group indicators. This allows us to capture both gradual within-worker learning effects and systematic differences across experience levels in how workers respond to recommendations and familiarity cues.

6.2.1 Store Familiarity. Figure 6 reports estimated coefficients for *PastFrequency*, which measures a worker's tendency to select tasks from stores they have previously visited. For the reference group (workers with 129+ total orders), coefficients are consistently positive across all order ranges, indicating a strong preference for familiar stores and suggesting an exploitation-oriented strategy.

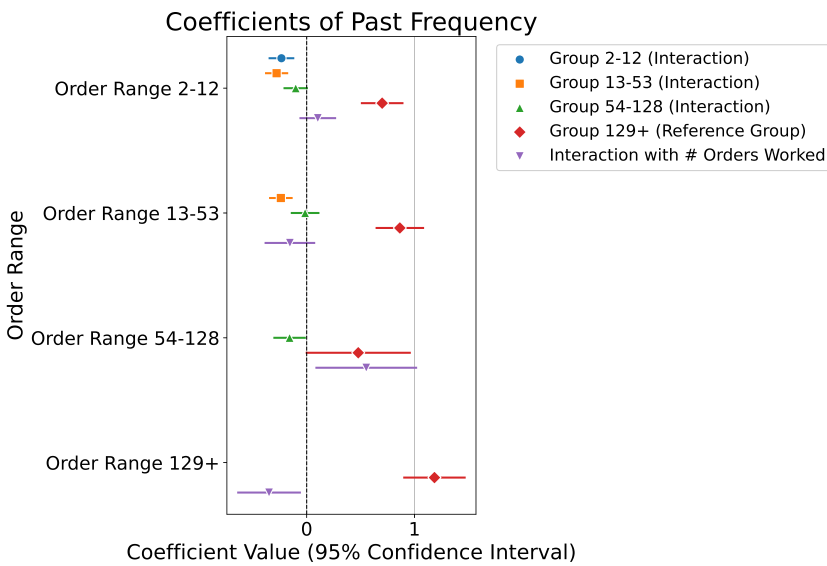


Fig. 6. Estimated MNL coefficients of *PastFrequency* by order range and tenure group.

Notes: Coefficients are from separate multinomial logit models for four cumulative order bins (y-axis: 2–12, 13–53, 54–128, 129+ orders). The red line shows the baseline coefficient for the reference group (129+ orders). Blue, orange, and green lines represent interaction terms for the 2–12, 13–53, and 54–128 groups, respectively. The purple line captures the interaction with cumulative experience within each bin. Error bars show 95% confidence intervals.

Relative to the 129+ group, interaction terms for lower-tenure groups (2–12, 13–53, 54–128) are generally negative, especially in early order ranges, indicating that less experienced workers place less weight on familiarity. For example, in the 2–12 range, both the 2–12 and 13–53 groups show a much weaker association between *PastFrequency* and choice probability.

The purple lines in Figure 6 reveal how this reliance shifts with cumulative orders within each bin. In the earliest ranges, there is little change in behavior as workers gain orders. In the 54-128 range, the interaction becomes significantly positive, suggesting a growing preference for familiar stores. In contrast, in the 129+ range, the coefficient turns negative, implying that even highly experienced workers may diversify away from familiarity after extensive tenure.

These results suggest that store familiarity is a key driver for experienced workers but is less influential early on, and may decline again after substantial tenure. This pattern contrasts with findings in Dai et al. (2022) [9], who inferred exploration from visiting new areas or stores and concluded that early-stage workers are more exploratory. Our choice-based model highlights that exploitation based on familiarity can dominate, particularly for high-tenure workers, when examined at the moment of decision.

6.2.2 Algorithmic Recommendations. For the 129+ reference group, *PlatformRecommended* has a strong and positive effect in the early stages, particularly in the 2–12 and 13–53 ranges, indicating that high-tenure workers initially follow the platform recommendations. This effect drops sharply

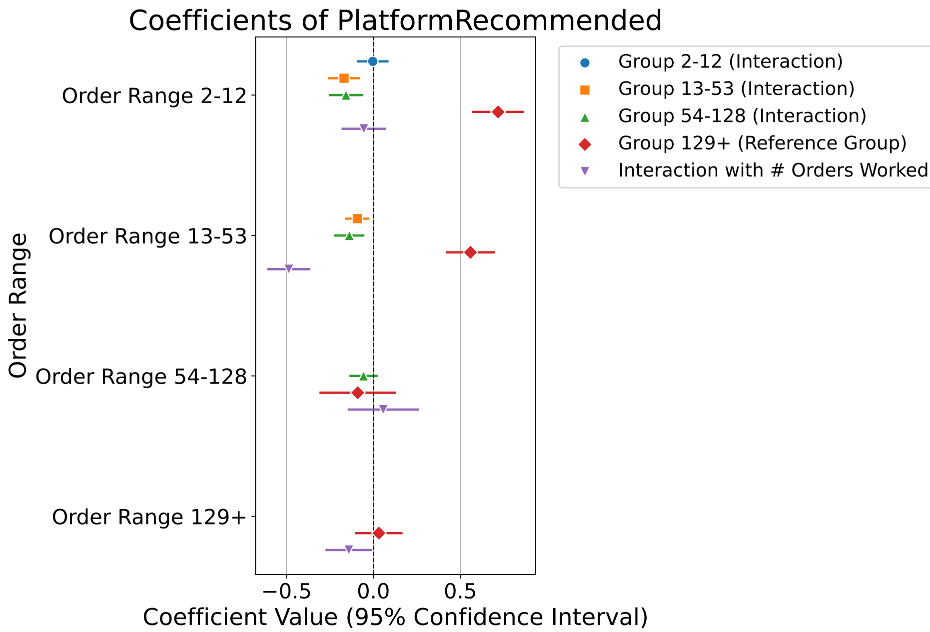


Fig. 7. Estimated MNL Coefficients Related to Recommendation Status *PlatformRecommended* by Order Range and Tenure Group.

Notes: This figure reports coefficients from separate multinomial logit models estimated across four cumulative order bins (y-axis: 2–12, 13–53, 54–128, and 129+ orders). The red line shows the baseline coefficient for *PlatformRecommended* in the reference group (workers with 129+ total orders). Blue, orange, and green lines represent interaction terms with tenure groups 2–12, 13–53, and 54–128, respectively. The purple line indicates the interaction between *PlatformRecommended* and cumulative orders worked. Error bars denote 95% confidence intervals. For non-reference groups, the total marginal effect equals the sum of the red line and the corresponding interaction coefficient.

in the 54–128 and 129+ ranges, where coefficients are statistically indistinguishable from zero, suggesting that experienced workers become more independent in their task selection.

Relative to the reference group, the 13–53 and 54–128 tenure groups show significantly negative interaction terms in the 2–12 range, implying weaker early reliance on recommendations. These group-level differences diminish in later order ranges.

The purple bars in Figure 7, capturing interactions between *PlatformRecommended* and cumulative orders, are consistently negative across bins, with statistically significant declines in the 13–53 and 129+ ranges. This provides robust evidence that reliance on recommendations decreases as workers gain experience.

Summary of Insights. Recommendation effects are strongest early in a worker's platform tenure and weaken with experience for all groups. High-tenure workers also reduce reliance on store familiarity over time, suggesting that both heuristics and algorithmic guidance become less influential as workers gain confidence. These results imply that adaptive recommendation systems, tailored to user experience level and strategic intent, could sustain engagement and improve efficiency.

7 Discussion and Implications

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.

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 use multinomial logit models with full choice sets, enabling estimation of how specific task attributes, such as store familiarity and recommendation status, influence selection. On bundling, we extend earlier work noting that workers often reject platform-generated batches [19]. We show that bundling behavior is dynamic: high-tenure workers increasingly adopt platform-generated bundles, while mid-tenure workers experiment more with self-bundling. These results suggest that bundling strategies evolve with experience and engagement. To our knowledge, this is among 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 *PlatformRecommended* is positive for early-stage workers but becomes statistically negligible as tenure increases. While our comparisons involve coefficients across models and interaction terms, the consistency of this decline indicates a robust behavioral shift. Notably, this pattern coincides with increased acceptance of platform-generated bundles (Section 5). These patterns are not contradictory: bundles are indivisible units optimized for efficiency and can be evaluated using observable features, while recommendations rely more on subjective fit and may be weighed 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 by showing that workers differentiate among algorithmic tools based on task type and perceived value. Our findings align with research on selective algorithm use [5, 14] and support calls for human-centered AI systems that adapt to user experience [16, 41].

Implications for adaptive platform design. The results 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, highlights an opportunity for platforms to modify the prioritization or presentation of recommendations. In contrast, the sustained adoption of platform-generated bundles by experienced workers indicates that algorithmic tools remain valuable when their benefits are clearly visible. Although this study focuses on a platform with high worker autonomy, similar patterns may occur on other gig platforms where workers have some choice: early reliance on platform guidance, increasing preference for familiar tasks, and selective use of algorithmic inputs as experience grows. In more constrained settings, adaptation may involve adjusting the timing, presentation, or framing of tasks rather than altering direct task selection. Overall, aligning algorithmic support with evolving worker preferences and capabilities could enhance both satisfaction and performance. Systems that adapt to experience level and behavioral signals may be more effective in sustaining engagement. Future work should evaluate these strategies across a wider 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 mitigate this concern. 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 cannot observe algorithm or interface changes, which could influence 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. Future research could integrate interface logs or worker interviews to contextualize decisions, study multiple platform models, or explore co-evolution of worker and algorithm behavior. Extending this work to other labor platforms will help assess generalizability and refine 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 issues in human–algorithm interaction and the design of decision-support systems in labor platforms.

Our analysis proceeds in three parts. First, we track learning curves using descriptive trends and two-way fixed effects regressions to quantify performance gains with experience. Second, we analyze bundling behavior, comparing engagement with platform-generated versus self-initiated bundles across tenure groups. Third, we apply a multinomial logit (MNL) model to examine task selection 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 trajectories are heterogeneous. High-tenure workers adopt platform-generated bundles more frequently, suggesting sustained reliance on optimization tools. At the same time, they rely less on platform-generated recommendations, increasingly favoring familiar stores and self-developed heuristics. This divergence in responses to different types of algorithmic support illustrates that 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. The results highlight the importance of tailoring platform support to workers' experience levels and revealed preferences. As platforms play a growing role in structuring work, understanding how workers develop expertise and autonomy in response to algorithmic systems will be essential for improving both platform performance 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. doi:10.1145/3109859.3109866

[5] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2025. Improving human sequential decision making with reinforcement learning. *Management Science* (2025).

[6] Hessam Bavafa and Jónas Oddur Jónasson. 2021. The variance learning curve. *Management Science* 67, 5 (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. *Available at SSRN 4106978* (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. doi: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. doi: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*. 163–167.

[20] Eric H Grosse and Christoph H Glock. 2015. The effect of worker learning on manual order picking processes. *International Journal of Production Economics* 170 (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 *(CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 181–184. doi: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. doi: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. doi: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*. 241–244.

[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 (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. doi: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. doi: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.

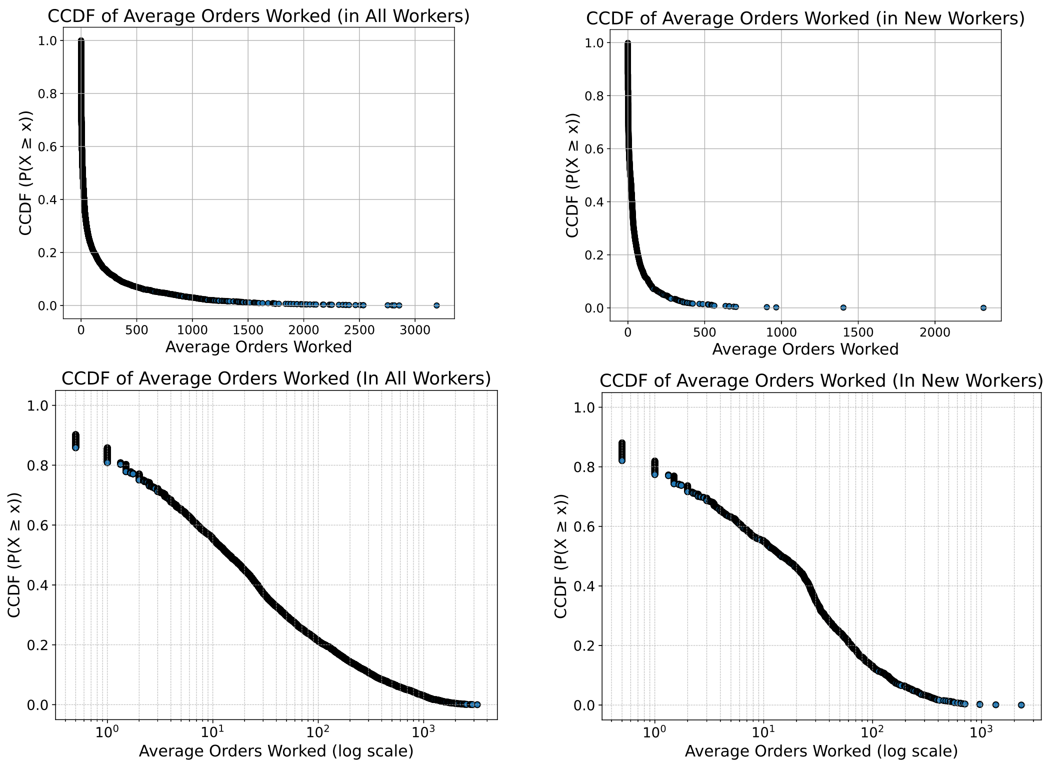
[41] Angie Zhang, Alexander Boltz, Jonathan Lynn, Chun-Wei Wang, and Min Kyung Lee. 2023. Stakeholder-centered AI design: Co-designing worker tools with gig workers through data probes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 63. ACM, New York, NY, USA, 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, et al. 2025. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945* (2025).

A Additional Statistics of Our Dataset for Section 3

To illustrate the distribution of worker activity on the platform, Figure 8 shows the Complementary Cumulative Distribution Function (CCDF) of total orders completed per worker during the one-year study period. Panel 8a reports this distribution for all approximately 5,000 active workers, while Panel 8b focuses on the 1,131 newcomers who joined during the study period and form the primary focus of our subsequent learning analyses. Both plots reveal a highly skewed distribution typical of many online platforms: a large share of workers completes only a small number of orders (indicated by the steep initial drop in the CCDF), while a long tail consists of a smaller set of highly active workers responsible for a disproportionate share of orders. This substantial heterogeneity in engagement is evident even within the newcomer cohort, motivating our later analyses that segment workers by activity levels (tenure groups) to examine variations in learning and strategy development.



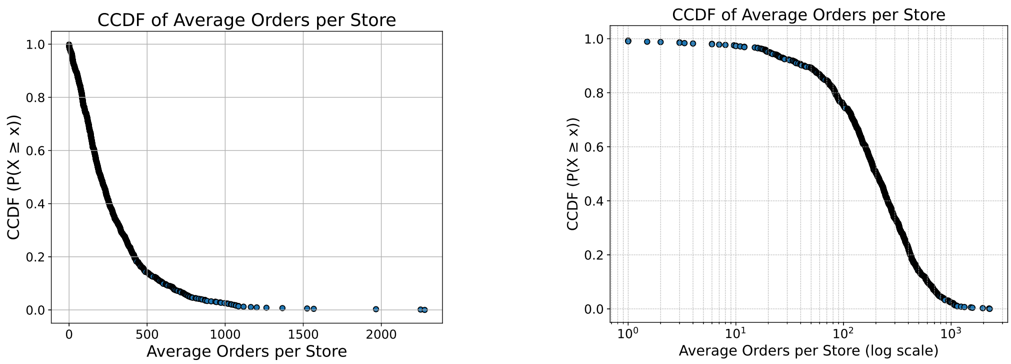
(a) All Workers ($N \approx 5{,}000$) (b) Newcomer Cohort ($N = 1{,}131$)

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 (and its log scale) completed per worker over the one-year study period. The Y-axis represents the probability $P(X \geq x)$, or the fraction of workers completing x or more orders. Panel (a) reports the distribution for all active workers in the dataset ($N \approx 5{,}000$). Panel (b) shows the distribution for newcomers ($N = 1{,}131$) 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 examined the distribution of order volume across the approximately 800 stores included in the dataset. Figure 9 shows the CCDF of the total number of orders processed per store over the study year. As with worker activity, store order volume is highly skewed. A large share of stores handled relatively few orders, as indicated by the steep initial decline in the CCDF curve. In contrast, a long tail represents a smaller group of high-volume stores that processed a disproportionately large share of total orders. This heterogeneity in store activity provides important context for understanding potential variations in worker experience, the role of store-specific learning (Section 4), and the broader operational landscape of the platform.

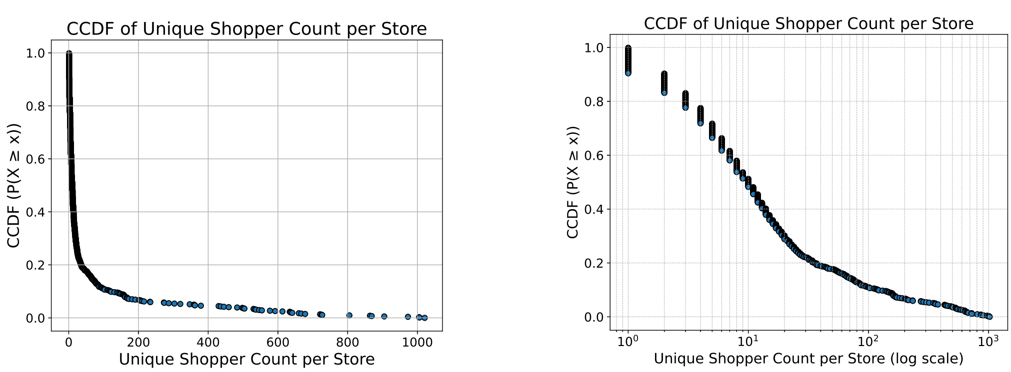
Complementing the order volume distribution, Figure 10 illustrates heterogeneity in worker traffic across stores by plotting the CCDF of the number of unique workers completing at least one order per store. Consistent with the patterns for order volume and worker activity, this distribution is also highly skewed. Many of the approximately 800 stores in the dataset were visited by only



(a) Natural Scale (b) Log Scale

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 for all participating stores ($N \approx 800$). 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: many stores handled relatively few orders, while a long tail of high-volume stores processed a disproportionately large share. This highlights heterogeneity in store importance and workload within the platform's operations.

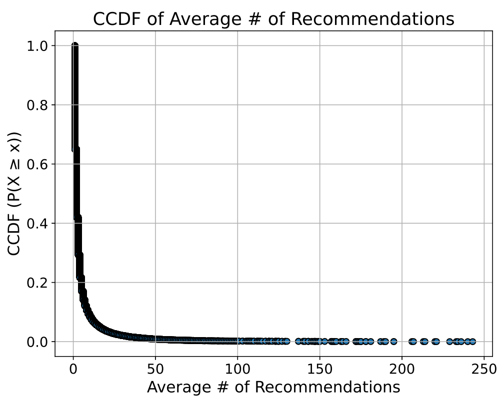


(a) Natural Scale (b) Log Scale

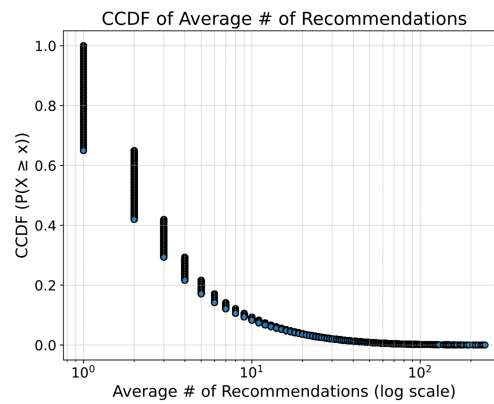
Fig. 10. Complementary Cumulative Distribution Function (CCDF) of Unique Worker Count per Store.

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

a small number of distinct workers over the year. In contrast, a long tail indicates that some high-traffic stores attracted a much larger and more diverse pool of workers. This variation in the number of unique workers interacting with each store provides further context for understanding the environment, including factors such as store-specific congestion and the diversity of worker experience at particular locations.



(a) Natural Scale



(b) Log Scale

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 included in the choice sets workers faced when selecting an order during the study period (mean = 2.47; quantiles (25%, 50%, 75%) = [0, 0, 2]). The Y-axis represents the probability $P(X \geq x)$, or the fraction of observed choice sets containing X or more recommended orders. The X-axis shows the number of recommended orders in the choice set (despite the axis label referring to “Average"). The plot reveals a highly skewed distribution: most choice sets contained very few recommended orders, while a small fraction included a large number. This provides context for the recommendation environment workers navigated.

To further describe the recommendation environment faced by workers, Figure 11 shows the distribution of algorithmically recommended orders available in workers' choice sets at the time of selection. The CCDF reveals a highly skewed pattern. In most decision instances, workers were presented with only a small number of recommended orders (often fewer than 10). However, the long tail shows that workers occasionally encountered choice sets containing a very large number of recommended options (up to roughly 200).

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

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

We first examined the distribution of join dates across our newcomer tenure quantile groups (defined in Section 5.1). The analysis revealed statistically significant differences in mean join dates across groups (ANOVA $p < 0.001$), confirming that workers with 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.

We also examined the practical distribution visually. As shown in Figure 12, while the median join dates differ (notably earlier for the 130+ group, corresponding to the 129+ label used elsewhere), there is substantial overlap in the interquartile ranges and overall distributions across the five quantile groups. This indicates that workers who joined at various points in the year are still represented across most tenure levels.

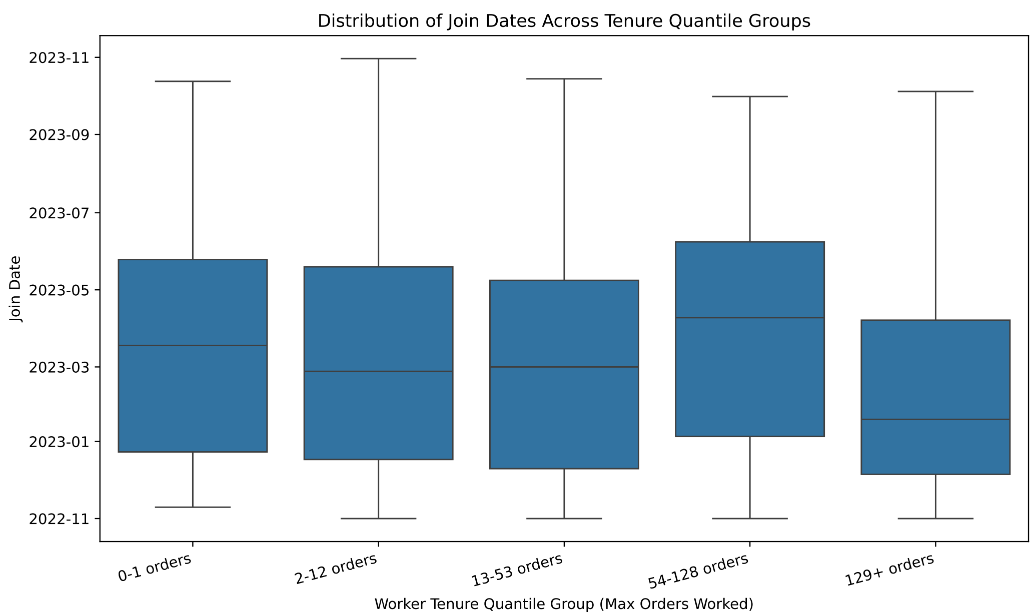


Fig. 12. Distribution of Worker Join Dates Across Tenure Quantile Groups.

Notes: 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 display the median, 25% and 75% quantile range, and overall range of join dates for workers in each group. Tenure groups are labeled as: '0–2 orders', '3–13 orders', '14–54 orders', '55–129 orders', and '130+ orders'.

To further assess the impact of this correlation, we performed a robustness check focusing only on 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 substantially reduce the potential confounding effect of join date.

Importantly, Figure 13 shows that even within this restricted cohort, significant differences in performance trajectories remain across eventual tenure groups. Workers who eventually completed the most orders (129+) exhibit 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 association between higher eventual tenure and better performance is not solely an artifact of earlier join dates. Instead, it suggests that differences in learning, strategy, or initial characteristics contribute to which workers persist and achieve higher engagement, even among those who started around the same time.

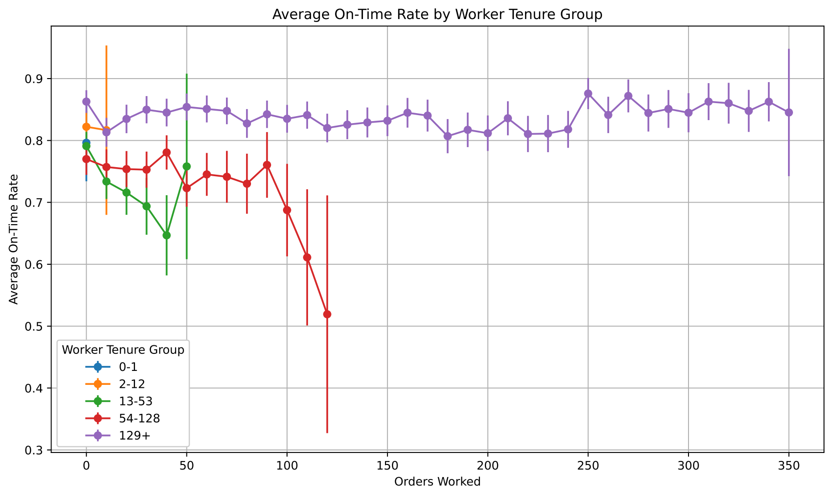


Fig. 13. Average On-time Rate by Worker Tenure Group (Restricted Cohort).

Notes: 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 ($N = 690$). Separate lines represent workers belonging to different eventual tenure groups, defined by quantiles of the maximum total orders completed over the full 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 show 95% confidence intervals for the mean on-time rate within each bin. This restricted analysis mitigates potential confounding by join date when comparing performance trajectories across eventual tenure groups.

Therefore, although the correlation between join date and final tenure is indeed a limitation, the robustness check supports the validity of using quantile-based tenure groupings for the 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 reports the results of the Tukey HSD (Honestly Significant Difference) post-hoc test (significance level = 0.05), conducted following a significant ANOVA result ($F = 8.114, p < 0.001$) that compared mean on-time rates across the five worker tenure quantile groups. This analysis focuses on workers' very earliest experience on the platform, specifically within the 0–1 order completed bin. The test performs pairwise comparisons between all tenure groups to identify which specific groups had significantly different mean on-time rates during this initial period.

The table columns are defined as follows:

- **Group 1 / Group 2**: The pair of worker tenure quantile groups being compared.
- **Mean Diff**: The difference between the mean on-time rate of Group 2 and Group 1, calculated using only data from the 0–1 order completed bin. A positive value indicates Group 2 had a higher average on-time rate in this 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 mean difference. If this interval does not contain zero, the difference is statistically significant at $\alpha = 0.05$.
- **Reject**: Indicates whether the null hypothesis (equal group means) should be rejected. (True = significant difference; False = no significant difference).

Key findings indicate that, even within their first two orders, workers who eventually achieved the highest tenure (129+ orders) had a significantly higher mean on-time rate than those in the lowest tenure group (0–1 order), 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 compared to the 13–53 orders group during this initial period. Other pairwise comparisons did not reveal statistically significant differences in on-time rate at this early stage. Overall, these results suggest that differences in performance trajectories between workers who ultimately achieve different tenure levels emerge very early in their engagement with the platform.

Table 2. Tukey HSD Test Results for Mean On-time Rate across Worker Tenure Groups (Period: 0-1 order). ANOVA F-statistic = 8.114, p-value = 1.745×10^{-6}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 order	129+ orders	0.1123	0.0010	0.0334	0.1912	True
0–1 order	13–53 orders	-0.0255	0.9010	-0.1037	0.0528	False
0–1 order	2–12 orders	0.0661	0.1798	-0.0159	0.1481	False
0–1 order	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

Table 3. Tukey HSD Test Results for Mean On-time Rate across Worker Tenure Groups (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

Table 3 reports the pairwise comparisons for the 2–12 orders period. In this stage, the 129+ orders group continued to significantly outperform the 13–53 orders and 54–128 orders groups. The 2–12 orders group also maintained significantly higher performance than both the 13–53 and 54–128 orders groups.

Table 4 presents the pairwise comparisons for the 13–53 orders period. Significant performance differences were observed among all three groups: the highest tenure group (129+ orders) significantly outperformed both the 13–53 and 54–128 orders groups, and the 54–128 orders group significantly outperformed the 13–53 orders group.

Table 5 provides the results for the 54–128 orders period. Here, the only possible comparison between tenure groups shows that the highest tenure group (129+ orders) continued to significantly outperform the 54–128 orders group. No Tukey HSD test was performed in this case since only two groups were available for comparison.

Table 4. Tukey HSD Test Results for Mean On-time Rate across Worker Tenure Groups (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 across Worker Tenure Groups (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

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

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 order	129+ orders	0.1607	<0.001	0.0792	0.2422	True
0–1 order	13–53 orders	0.0967	0.0096	0.0160	0.1775	True
0–1 order	2–12 orders	0.0515	0.4594	-0.0332	0.1362	False
0–1 order	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

The following tables report pairwise comparisons of mean bundle volume across tenure groups for different early order periods, each following a significant ANOVA.

Table 6 reports results for the initial 0–1 order period (ANOVA: F = 7.828, $p < 0.001$). At this very early stage, significant differences ($\alpha = 0.05$) show that the highest tenure group (129+ orders)

Table 7. Tukey HSD Test Results for Mean Bundle Volume across Worker Tenure Groups (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 across Worker Tenure Groups (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 across Worker Tenure Groups (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

bundled more than the 0–1, 2–12, and 54–128 orders groups. Additionally, the 13–53 orders group bundled more than the 0–1 orders group.

Table 7 presents comparisons for the 2–12 orders period (ANOVA: F = 37.250, p < 0.001). Significant differences ($\alpha = 0.05$) indicate that the 129+ orders group bundled more than all other groups, 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 8 summarizes comparisons for the 13–53 orders period (ANOVA: F = 35.257, p < 0.001). Here, significant differences ($\alpha = 0.05$) show that 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 provides the comparison for the 54–128 orders period (ANOVA: F = 13.840, p < 0.001). The only possible contrast 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 Proportion

The tables report post-hoc Tukey HSD comparisons of the average platform-bundled proportion 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 10. Tukey HSD Test Results for Platform-Bundled Proportion across Worker Tenure Groups (Period: 0-1 order). ANOVA F-statistic = 4.429, p-value = 1.45×10^{-3}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 order	129+ orders	0.0387	<0.001	0.0123	0.0650	True
0–1 order	13–53 orders	0.0199	0.2305	-0.0063	0.0461	False
0–1 order	2–12 orders	0.0210	0.2236	-0.0064	0.0484	False
0–1 order	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 Proportion across Worker Tenure Groups (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 Proportion across Worker Tenure Groups (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

Table 13. Mean Difference of Platform-Bundled Proportion across Worker Tenure Groups (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 presents comparisons for the initial 0–1 order period (ANOVA: F = 4.429, p < 0.01). At this earliest stage, the only statistically significant differences indicate that workers who ultimately reached 129+ orders had a higher average platform-bundled proportion 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 differences emerges. The 129+ group had significantly higher platform-bundled proportions than all other groups. In addition, the 13–53 group outperformed the 2–12 group, and the 54–128 group also bundled more than the 2–12 group.

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

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

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

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

Table 14 shows results for the initial 0–1 order period (ANOVA: F = 5.183, p < 0.001). At this early stage, workers in the 129+, 13–53, and 54–128 groups exhibited significantly higher self-bundled proportions than those in the 0–1 group. No other pairwise comparisons were significant.

In the 2–12 orders period (Table 15, ANOVA: F = 10.110, p < 0.001), several significant differences were observed. The 129+ group had a lower self-bundled proportion than the 54–128 group. In addition, the 54–128 group bundled more than the 2–12 group, and the 13–53 group bundled more than the 2–12 group.

Table 16 reports results for the 13–53 orders period (ANOVA: F = 46.612, p < 0.001). Here, workers in the 129+ group had significantly lower self-bundled proportions than both the 13–53 and 54–128 groups. The 13–53 group also bundled less than the 54–128 group.

Finally, Table 17 presents results for the 54–128 orders period (ANOVA: F = 84.727, p < 0.001). Even at this later stage, the 129+ group had a significantly lower self-bundled proportion than the 54–128 group.

Table 14. Tukey HSD Test Results for Self-Bundled Proportion across Worker Tenure Groups (Period: 0-1 order). ANOVA F-statistic = 5.183, p-value = 3.747×10^{-4}.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
0–1 order	129+ orders	0.0576	0.0198	0.0060	0.1092	True
0–1 order	13–53 orders	0.0619	0.0086	0.0108	0.1131	True
0–1 order	2–12 orders	0.0124	0.9697	-0.0412	0.0661	False
0–1 order	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 Proportion across Worker Tenure Groups (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 Proportion across Worker Tenure Groups (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 Proportion across Worker Tenure Groups (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

D Full Independent Variables in the MNL Model in Section 6

- **PlatformRecommended**: Indicator for whether the order is recommended by the platform's algorithm (1 = recommended, 0 = not recommended).
- **BUNDLED**: Indicator for whether the order is part of a bundle (1 = bundled, 0 = not bundled).
- **Past Frequency**: The proportion of orders a worker has previously completed from the same store, relative to their total completed orders across all stores, measured prior to the current choice. This value is 0 if the worker has not previously completed any orders from the 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 included in the order.
- **DOLLARS_BONUS**: The dollar amount of any bonus offered for completing the order.
- **DOLLARS_PAY**: The base 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 order (rounded to 0, 0.5, or 1 for data privacy).
- **PCT_EXPANDED_FOOD_ITEMS**: The percentage of expanded food items in the order (rounded to 0, 0.5, or 1 for data privacy).

- **PCT_GENERAL_MERCH_ITEMS**: The percentage of general merchandise items in the order (rounded to 0, 0.5, or 1 for data privacy).
- **MAX_CONTRIBUTIONS_CAT_PCT**: The proportion of the order accounted for by its largest item category (continuous from 0 to 1). This serves as a proxy for how concentrated versus distributed the order's items are across categories.

E MNL Key Variables Estimated Coefficients in Section 6

Table 18. Multinomial Logit Model Results (Order 2 - 12)

Variable	Coef.	S.E.	z	p-value	95% C.I. Lower	95% C.I. Upper
FREQ_PERCENTAGE	0.700***	(0.096)	7.277	0.000	0.512	0.889
FREQ_PERCENTAGE × ORDERS_WORKED	0.103	(0.082)	1.258	0.209	−0.058	0.264
FREQ_PERCENTAGE × GROUP_13–53	−0.279***	(0.050)	−5.547	0.000	−0.378	−0.180
FREQ_PERCENTAGE × GROUP_2–12	−0.234***	(0.056)	−4.216	0.000	−0.343	−0.125
FREQ_PERCENTAGE × GROUP_54–128	−0.102*	(0.053)	−1.909	0.056	−0.206	0.003
PlatformRecommended	0.717***	(0.073)	9.776	0.000	0.574	0.861
PlatformRecommended × ORDERS_WORKED	−0.055	(0.062)	−0.886	0.375	−0.177	0.067
PlatformRecommended × GROUP_13–53	−0.169***	(0.044)	−3.812	0.000	−0.256	−0.082
PlatformRecommended × GROUP_2–12	−0.004	(0.043)	−0.082	0.935	−0.088	0.081
PlatformRecommended × GROUP_54–128	−0.158***	(0.047)	−3.341	0.001	−0.250	−0.065
Model Statistics						
Observations	7,042					
Log-likelihood	−8,018.106					
Pseudo R^2	0.133					
Pseudo $\bar{R}^2$	0.126					
AIC	16,178.213					
BIC	16,665.248					

Notes: Standard errors in parentheses. Statistical significance:
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Model estimated using maximum likelihood estimation. Reference group for interactions is tenure group 129+.

Table 19. Multinomial Logit Model Results (Order 13–53)

Variable	Coef.	S.E.	z	p-value	95% C.I. Lower	95% C.I. Upper
FREQ_PERCENTAGE	0.864***	(0.110)	7.853	0.000	0.649	1.080
FREQ_PERCENTAGE × ORDERS_WORKED	−0.156	(0.115)	−1.350	0.177	−0.382	0.070
FREQ_PERCENTAGE × GROUP_13–53	−0.240***	(0.051)	−4.702	0.000	−0.339	−0.140
FREQ_PERCENTAGE × GROUP_54–128	−0.015	(0.063)	−0.230	0.818	−0.138	0.109
LIST	0.559***	(0.068)	8.248	0.000	0.426	0.691
LIST × ORDERS_WORKED	−0.487***	(0.060)	−8.109	0.000	−0.605	−0.369
LIST × GROUP_13–53	−0.093***	(0.032)	−2.880	0.004	−0.156	−0.030
LIST × GROUP_54–128	−0.139***	(0.041)	−3.376	0.001	−0.219	−0.058
Model Statistics						
Observations	15,482					
Log-Likelihood	−12,180.127					
Pseudo R^2	0.044					
Pseudo $\bar{R}^2$	0.040					
AIC	24,474.254					
BIC	24,910.157					

Notes: Standard errors in parentheses. Statistical significance:
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Model estimated using maximum likelihood estimation. Reference group for interactions is tenure group 129+.

Received October 2024; revised April 2025; accepted August 2025

Table 20. Multinomial Logit Model Results (Order 54–128)

Variable	Coef.	S.E.	z	p-value	95% C.I. Lower	95% C.I. Upper
FREQ_PERCENTAGE	0.479*	(0.244)	1.958	0.050	-0.001	0.957
FREQ_PERCENTAGE × ORDERS_WORKED	0.553**	(0.236)	2.339	0.019	0.090	1.016
FREQ_PERCENTAGE × GROUP_54–128	-0.159**	(0.072)	-2.203	0.028	-0.300	-0.017
LIST	-0.090	(0.109)	-0.829	0.407	-0.304	0.123
LIST × ORDERS_WORKED	0.057	(0.101)	0.562	0.574	-0.141	0.254
LIST × GROUP_54–128	-0.057*	(0.038)	-1.495	0.135	-0.131	0.018
Model Statistics						
Observations	15,426					
Log-Likelihood	−11,649.175					
Pseudo R^2	0.053					
Pseudo $\bar{R}^2$	0.049					
AIC	23,384.349					
BIC	23,713.033					

Notes: Standard errors in parentheses. Statistical significance:
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Model estimated using maximum likelihood estimation. Reference group for interactions is tenure group 129+.

Table 21. Multinomial Logit Model Results (Order 129+)

Variable	Coef.	S.E.	z	p-value	95% C.I. Lower	95% C.I. Upper
FREQ_PERCENTAGE	1.186***	(0.143)	8.269	0.000	0.905	1.467
FREQ_PERCENTAGE × ORDERS_WORKED	−0.349**	(0.146)	−2.387	0.017	−0.636	−0.062
LIST	0.032	(0.066)	0.483	0.629	−0.097	0.160
LIST × ORDERS_WORKED	−0.142**	(0.065)	−2.165	0.030	−0.270	−0.013
Model Statistics						
Observations	18,717					
Log-Likelihood	−15,157.627					
Pseudo R^2	0.050					
Pseudo $\bar{R}^2$	0.048					
AIC	30,373.254					
BIC	30,600.532					

Notes: Standard errors in parentheses. Statistical significance:
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Model estimated using maximum likelihood estimation. Reference group for interactions is tenure group 129+.