Balancing Flex and Non-Flex Labor to Reliably Meet On-Demand Capacity

Ramon Auad
Amazon.com
ramauad@amazon.com

Thomas Fillebeen
Amazon.com
fillebee@amazon.com

Roman Levkin
Amazon.com
rllevkin@amazon.com

Arkajit Rakshit
Amazon.com
rakshit@amazon.com

Martin Savelsbergh
Amazon.com
savelsma@amazon.com

Abstract

The 21st century workforce is increasingly characterized by more flexible labor models, particularly in e-commerce and supply chain operations. While previous research has focused mostly on last-mile, on-the-road settings, we focus on under-the-roof (UTR) environments, which present unique challenges due to their complex, varied tasks requiring training and experience. Our study addresses the need to better understand how a blended UTR workforce balances factors like structural efficiency and labor flexibility in complex logistics management. We present an optimization framework for determining an effective workforce composition of flexible and non-flexible associates in UTR environments. We validate insights from the optimization framework through an empirical study that increased flexible staffing at Amazon delivery stations. Our analysis includes measuring differences in productivity learning curves and examining impact on efficiency and the associate experience. Key findings reveal that while flexible associates take longer to achieve full proficiency, especially for complex tasks, these effects diminish over time. Importantly, a blended workforce improves structural staffing efficiency and makes it easier to accommodate demand shocks. We estimate that the upper bound of the efficiency improvement to be around 4%. Our research highlights the benefits of strategic workforce planning in the face of increasing demand volatility and a need for operational agility.

Keywords: E-commerce Logistics; Workforce Planning; Labor Flexibility; Part-Time Labor; Mixed-Integer Programming

1 Introduction

The traditional Supply Chain labor model for under-the-roof (UTR) operations is characterized by a fixed, full-time (non-flex) workforce that operates using standard recurring shift patterns. These associates are typically trained to handle complex tasks within the facility. An example would be an associate who works four 10-hour shifts from Monday through Thursday every week in a fulfillment facility on receiving, stowing, picking, and packing inventory. In contrast, the rise of the on-demand economy introduced the need for a more part-time (flex) workforce, where associates work a variable number of hours per week within a certain range of hours defined by lower and upper bounds. Allowing customers to order any time they want from anywhere they want and rapidly delivering to wherever they want creates fluctuating demand that poses significant operational challenges, often leading to tens of millions of dollars in incremental cost or lost revenue. Amazon delivery stations in the US exemplify this challenge. Using 2024 weekly demand data, we estimate the coefficient of variation for demand across stations. Figure 1 illustrates week-over-week demand fluctuations for stations at different demand volatility levels (25th percentile, median, and 75th percentile). The data reveal strong seasonality, with holiday season volumes more than double that of the first two quarters. In general, week-over-week demand shows substantial variability: when demand increases, the average demand growth is 7%, while when demand decreases the average decline is 8%. Accommodating this volatility by utilizing only a full-time (non-flex) workforce would require systematic over-staffing by 8%-15%. If we assume that at least half of this volatility can be accommodated by part-time (flex) employees, we can achieve up to 4%-7% cost savings. This highlights the importance of understanding what constitutes an optimal mix of full-time and part-time employees for logistics businesses.

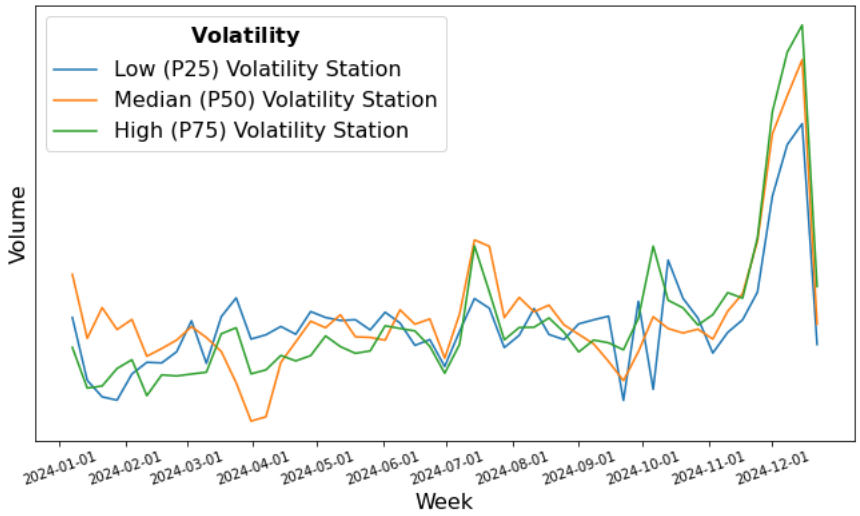

Figure 1: Weekly Volume For Low, Median, and High Volatility Stations By Week

While the flexibility offered by part-time associates enables businesses to adjust capacity to meet fluctuating demand, implementing it in UTR environments presents unique challenges. Unlike simpler gig economy models, UTR operations require associates to perform more intricate varied tasks, often necessitating longer training. For example, in the Amazon use case presented in Section 3, we estimate that it may take 4-8 weeks for new associates to achieve the same proficiency as the veteran associates and where the initial productivity gap can be up to 25%. The lengthy training period and the high training cost further increase the complexity and urgency of understanding the optimal workforce composition in the UTR settings. Furthermore, even when a business has the ability to bring in additional workforce, it needs to be planned carefully, with a gradual employee ramp-up period to avoid a degradation to productivity. Furthermore, due to the difference in working hours per week, it is reasonable to assume that there is a significant difference between the productivity learning

curves of full-time and part-time associates. This needs to be empirically tested and incorporated into optimizing UTR workforce composition decisions.

Striking the right balance between operational efficiency, maintaining high-quality performance, and service availability in complex UTR tasks represents a strategic challenge for supply chain operations in determining the optimal mix between non-flex and flex associates. We first provide an overview of related literature, which has previously focused mostly on non-UTR operations. Then, we motivate our study with the use case of Amazon delivery stations' UTR workforce. The paper details a two-component modeling framework to determine the optimal composition of non-flex and flex associates—the Structural Efficiency Model (SEM) and the Cost Model (CM). The SEM model focuses on demand variability and volatility on a finer scale, hour by hour throughout an operations day, aiming to optimize service level by minimizing underage and overage hours while maintaining a target range for non-flex hours. In other words, the SEM model focuses on structural efficiency differences between non-flex and flex workforce at a granular, hourly level. Flex associates typically have shorter, more variable shift lengths compared to the standard 40-hour non-flex weekly schedules. Hence, when labor need fluctuates throughout the day with spikes during certain hours, it can be more efficiently covered by shorter flex shifts. On the other hand, since flex associates work fewer hour per week, they may require longer training to achieve full proficiency. Furthermore, the fact that flex associates have the ability to choose their working hours within a specified range introduces more uncertainty into workforce planning. This creates additional challenges for planners as they need to hedge against uncertainty to minimize any risks of service availability. The SEM model captures these trade-offs to recommend efficient mix of flex and non-flex associates to optimize this structural efficiency throughout the operations day. In contrast, the CM addresses demand volatility over time, day over day or week over week. In this case, planners need to efficiently accommodate demand volatility over time. If they hire more associates to accommodate a temporary demand shock, this leads to overstaffing and inefficiency when demand decreases. Flex associates offer the agility to quickly scale capacity and meet fluctuating demand without the need to hire incremental non-flex associates. However, this may come at the expense of lower productivity and longer learning curves because flex associates develop expertise in UTR operations more gradually, and potentially higher uncertainty around their weekly working hours. In particular, non-flex, full-time associates provide a reliable foundation of productivity and operational capacity due to their consistent work hours, ensuring predictable service levels and proficiency in complex tasks. The CM model finds balance between these two associate types in order to enable high service availability to accommodate fluctuations in customer demand, lower labor planning uncertainty, and efficient labor cost management. The two models are solved in sequence: SEM first, CM second. A range of flex to non-flex ratios achieving high structural efficiency is passed to the cost optimization in the form of guardrails. The two models reflect that service and cost are both important for profitability. The sequence reflects Amazon's relentless focus on customer convenience and experience: operational planning seeks to minimize cost while ensuring that a (very) high service level is maintained.

We apply this two-component modeling framework at Amazon delivery stations by running an experiment which increased flex staffing at a select group of stations in 2023. This combination of optimization modeling and empirical validation lends credibility to the final recommendations for workforce composition in UTR environments. We conclude with a discussion of how the insights generated from our study can inform strategic workforce planning decisions across different supply chain miles, considering the unique challenges of integrating flex associates into complex UTR environments.

2 Literature Review

All companies strive to have the right people in the right place at the right time and at the right cost – a concept that is sometimes referred to as workforce readiness. Various aspects contribute to workforce readiness. Workforce planning is focused on determining the right workforce composition and is strategic in nature. Workforce optimization is focused on how to best use a given workforce and is operational in nature. Workforce planning becomes more challenging when dealing with a blended workforce, *i.e.,* a workforce comprised of different types of workers - for example a mix of full-time and part-time employees, or a mix of company employees and contractors, etc. Achieving workforce readiness is challenging, in part, because of the ubiquitous presence of demand uncertainty. Providing a comprehensive literature review of an area that encompasses so many different topics is

challenging. It is further complicated by the fact that alternative terms are often used, for example workforce planning and workforce optimization are sometimes referred to as staffing and scheduling. Therefore, we are selective and discuss a few articles that we have found helpful in developing our mental model of workforce composition optimization.

A strand of economic literature extensively analyzes the supply side of the blended workforce, in particular, productivity and wages of part-time employees. To explain productivity differences between part-timers and full-timers, the literature focuses on a few factors such as shift duration, the nature of the part-time contracts (*e.g.,* weekly work hours), and the labor pool that part-time jobs attract. First, assuming no intrinsic differences in full-time and part-time job seekers, productivity can be lower during the first few hours of work due to start-up cost (Barzel [1994]) making employees less productive in the beginning of the shift. However, longer work hours can lead to employees' fatigue (Brewster et al. [1994]), which can create a bell-shaped relationship between work hours and productivity. Devicienti et al. [2018] use firm-level observational data from Italy to measure the relationship between the percentage of firm's part-time workers and its productivity. They find that one standard deviation increase in the share of part-time labor decreases firm's productivity by 2%. Second, besides part-time shifts duration, the arrangement of part-time contracts seems to play a role as well. In particular, the literature distinguishes short part-timers, who work fewer weekly hours, and long part-timers, who work more weekly hours. Specchia and Vandenberghe [2013] find that the negative impact of short part-timers is more pronounced—a 10% increase in the share of the part-time labor can decrease firm's productivity by 1.3% for low weekly hours part-timers and 0.7% for high weekly hours part-timers. At the same time, Specchia and Vandenberghe [2013] note that the estimates are sensitive to a specific industry, finding positive impact of part-time employees in retail and trade. Finally, this strand of literature studies potential implications of differences in job seekers who are attracted to part-time jobs, specifically, women, and potential implications for wage penalty for part-time workers and women. Manning and Petrongolo [2008] show that a sizable part of the part-time wage penalty can be explained by the different characteristics of full-time and part-time women. Garnero et al. [2025] find a positive impact of part-time workers on firm's productivity, however, it is driven entirely by long part-timers and men while they find no impact of short part-timers and women on productivity. Kim et al. [2025] leverage a large-scale field experiment in Ethiopia to show that when part-time and full-time data entry positions are randomly offered to potential candidates, the former tend to attract applicants with lower productivity but stronger preference for shorter work hours. In some cases, for example, for the above the median productivity workers, this self-selection bias can lead to 14% productivity degradation for part-time employees compared to full-time employees. Therefore, despite extensive research in this area, the estimates of the relationship between part-time work and productivity remain inconclusive and depend on a specific industry and context.

Another strand of the literature studies the demand side of the blended workforce, trying to explain firms' decisions to hire part-time employees and how these decisions are made.April et al. [2014] focus on workforce planning, which they define as the business process that enables the identification and analysis of what an organization will need in terms of the future size, type, and quality of workforce to achieve its objectives. The authors present "OptForce," a software system that enables organizations to optimize workforce readiness and composition while accounting for critical workforce constraints. OptForce offers a broad range of predictive analytics capabilities including sophisticated workforce demand planning, fine-grained employee retention models, and agent-based simulation forecasting. Fuller [2020] shares his view that most successful companies are those that adopt an on-demand workforce model, one that blends talent sourced externally and employed temporarily with full-time workers. Such a model, they argue, offers the prospect of achieving every manager's goal of putting the right talent on the right project at the right time. Their discussion focuses on the battle for highly skilled workers and the use of digital talent platforms to build blended workforces. Mahato et al. [2021] also discuss the rise of blended workforces in the post-COVID era, which they characterize as a movement away from a homogeneous workforce towards a heterogeneous workforce of full-time employees working in tandem with gig talents connected via digital platforms. Both Fuller [2020] and Mahato et al. [2021] see a blended workforce as a way for a company to acquire the talent they need for the success of their business.

In our context, UTR operations of Amazon's fulfillment networks, we see a blended workforce as a way to more cost-effectively handle demand uncertainty and volatility. Boysen et al. [2019] and Rijal et al. [2021] do not discuss in detail how to choose an optimal workforce composition mix but

focus on warehouse operations in the era of e-commerce. Boysen et al. [2019] argue that e-commerce has to rely on more automated warehouse systems (*e.g.,* automated picking workstations, robots, and AGV-assisted order picking systems) and novel organizational adaptations (*e.g.,* mixed-shelves storage, dynamic order processing, and batching, zoning and sorting systems). The following two papers consider workforce composition and scheduling in the context of delivery platforms. Rather than having full-time and part-time associates, as our paper does in this context, there are scheduled and unscheduled couriers. The time at which scheduled couriers work (when they start and how long they work) is decided by the platform, but the time at which unscheduled couriers work is decided by the courier himself. Initially, many delivery platforms only relied on unscheduled couriers, but to be able to provide reliable service to their customers, they had to introduce scheduled couriers over which they have more control. Behrendt et al. [2023] consider a scheduling problem that arises in an environment in which a mix of scheduled and unscheduled couriers (they refer to these couriers as ad hoc couriers) serve dynamically arriving pickup and delivery orders. The platform seeks a set of shifts for scheduled couriers in order to minimize total courier payments (scheduled and unscheduled couriers) and penalty costs for expired orders (orders that cannot be delivered by their promise time). Leveraging crowdsourced delivery capacity, *i.e.,* unscheduled couriers who offer their vehicles and time for deliveries, enables companies to provide faster delivery options and readily accommodate demand fluctuations. However, the inherent uncertainty of this crowdsourced model presents challenges in maintaining consistent service quality. To mitigate potential negative effects stemming from this uncertainty, companies often opt to maintain a scheduled delivery workforce alongside their crowdsourced capacity. This dual approach allows for greater control over service quality and reliability. Essentially, this strategy represents a trade-off between service availability and productivity, with service quality serving as proxy for the latter. Ulmer and Savelsbergh [2020] investigate continuous approximation and value function approximation methods for scheduling such a workforce, *i.e.,* deciding their shifts (start time and duration), to achieve a service level target at minimum cost.

As mentioned above, achieving workforce readiness is challenging, in part, because of the ubiquitous presence of demand uncertainty. The following two papers focus on that aspect. Kim and Mehrotra [2015] study the problem of integrated staffing and scheduling under demand uncertainty. This problem is formulated as a two-stage stochastic integer program with mixed-integer recourse. The here-and-now decision is to find initial staffing levels and schedules. The wait-and-see decision is to adjust these schedules at a time closer to the actual date of demand realization. Ganesan et al. [2017] propose a mixed integer programming approach to solve the problem of scheduling cybersecurity analysts and the determining the right mix of analysts' experience levels. Altner et al. [2018] focus on staffing decisions in the face of demand uncertainty and volatility. Motivated by a cybersecurity workforce optimization problem, the authors investigate optimizing staffing and shift scheduling decisions given unknown demand and multiple on-call staffing options at a 24/7 firm with three shifts per day, three analyst types, and several staffing and scheduling constraints. The problem is modeled as a two-stage stochastic program and is solved with a column-generation-based heuristic. Altner et al. [2018] optimize scheduling decisions for on-call analysts but do not explicitly answer the question of optimal mix of on-call analysts and full-time analysts.

This paper brings brings novelty into a few strands of workforce planning research. None of the above articles capture all (or even most of the) aspects of the workforce composition problem that is the focus of this paper. To the best of our knowledge, this is the first paper that explicitly focuses on the strategic question of optimal workforce composition, e.g., determining the efficient percentage mix of part-time and full-time employees.

Also, we extend more standard last mile delivery contexts for part-time employment to more complex under-the-roof operations. We are dealing with a variety of attributes of associates, *e.g.,* productivity, reliability, and attrition, *e.g.,* hourly compensation rates, special compensation rates for overtime, hiring cost, etc., tenure-based learning, shift structures, and demand uncertainty and volatility. This complex environment requires a customized solution approach, which will be discussed in the next sections. Finally, we contribute to the strand of economics literature that analyzes the impact of part-time employment on productivity. Previous literature often had to rely on observational data studies and could not differentiate short-term and long-term impact of part-timers on productivity, whereas we were able to leverage regional experiment to consider different time horizons.

3 Non-Flex vs Flex UTR Associates: The Amazon Use Case

From an UTR operations perspective, the Amazon fulfillment network represents the following chain:

- The inventory is stored in large fulfillment centers (FC). Upon a customer placing an order (*e.g.,* South Seattle area), this order is picked and packed at the closest FC (*e.g.,* Oregon FC) and shipped to the corresponding geographical delivery area sort center.
- At the sort center (*e.g.,* Seattle area) the packages are sorted and shipped to a specific delivery station.
- Finally, at the corresponding delivery station (*e.g.,* South Seattle area), the packages are sorted to a specific delivery zone and are dispatched to the customer's home address.

Historically, the Amazon operations have utilized a blended workforce model, offering non-flex and flex positions to associates who self-select based on their personal preferences and circumstances. Both non-flex and flex associates are hired and trained ahead of time. Non-flex associates are hired to specific recurring weekly schedules, whereas flex shifts are posted any time between three weeks and one day out based on planning needs. Non-flex associates, who have selected more structured work arrangements, demonstrate several advantages: faster progression through productivity learning curves, and higher stability as they work fixed and predictable schedules week over week. For supply chain planners, this translates into higher productivity and lower uncertainty of the non-flex associates. From an operational perspective, flex associates allow for greater adaptability to demand fluctuations, both throughout the operation day and over time, *e.g.,* their shifts tend to be shorter and they are more willing to flex their weekly worked hours up and down when needed. This self-selection into different work models enables the creation of a balanced and stable, highly productive core workforce with the operational agility needed to respond to changing demands, while respecting associates' diverse work preferences and life circumstances. The blended model thus optimizes both workforce satisfaction and operational efficiency.

In this paper, we discuss the development of an intuitive optimization framework to provide strategic guidance on workforce composition for Amazon's UTR operations. An optimal workforce composition depends on multiple factors, including structural staffing efficiency, the ability to manage demand variability, predictability of the operations, and overall business strategy. Rather than prescribing a single workforce blend, our approach aims to recommend a range of workforce compositions that provide a good balance between structural efficiency, the need to accommodate demand fluctuations, stability and productivity of operations. We leverage multiple data inputs, and mixed integer programming optimization to estimate and provide guidance on the key trade-offs that businesses face when choosing a workforce mix. Business leaders, UTR planners, and operators can utilize this guidance and the resulting estimates to select their preferred composition within the recommended range, aligning with current business priorities and operational realities. We start by analyzing the key differences between non-flex and flex associates using Amazon delivery stations operations as a case study. For confidentiality reasons, we leverage approximate data that represent the true magnitude of differences in structural efficiency, flexibility, productivity, and predictability of shift acceptance and attendance.[1]

3.1 Work Hours, Flexibility, and Structural Efficiency

At Amazon delivery stations, non-flex (full-time) associates work the same 40-hour schedule week over week. For example, a typical schedule can be from 8 am to 6 pm four days a week. They can flex down their weekly hours by signing up for Voluntary Time Off (VTO), and flex up by taking Voluntary Extra Time (VET). Both options are voluntary, *e.g.,* have to be accepted by associates. Since non-flex associates already work 40 hours week over week, their ability to flex working hours by accepting VET is typically limited.

Flex (part-time) associates can choose to work as many hours as they like from shifts offered by the business. The median weekly worked hours for flex associates is 12 hours per associate. A typical flex shift length can vary between 4 and 6 hours depending on the type of operations. Hence, on average per week non-flex associates work around three times more than the hours worked by flex associates.

[1]We add noise to the financial data for confidentiality reasons but they represent the true ball park of differences between non-flex and flex associates.

However, flex associates are more willing to adjust their weekly worked hours when customer demand fluctuates without significant monetary incentive. Depending on a week and geographical location we find the flex associates may be willing to work between 8 and 20 hours per week.

Furthermore, typically the labor need at a delivery station fluctuates throughout the day. There may be more need to labor when inbound trucks arrive or when the packages are being dispatched. Shorter flex shift structure compared to non-flex shifts makes it easier to fit the varying throughout the day labor need with a rigid shift structure. However, relying on flex shifts may lead to higher planning uncertainty and less stability as planners need to predict flex associates' willingness to accept those flex shifts, and associates may have preferences of flex shifts during certain days of week and times of day.

3.2 Demand Volatility

Another aspect of structural efficiency is labor need volatility not throughout the day but rather across time. We refer to this labor need volatility day over day or week over week as demand volatility. As a measure of demand volatility, we use the coefficient of variation (the ratio of the standard deviation to the mean) based on weekly demand data. We leverage weekly demand data for Amazon delivery stations in 2024 to calculate the coefficient of variation for each station and use it as a proxy for a station's volatility. Then, we analyze the distribution of volatility at the station level. On average the Amazon delivery stations need to accommodate volatility of 6% for the volume ask. Table 1 shows the distribution of demand volatility across delivery stations.

Table 1: Demand Volatility Across Multiple Percentiles

Metric	$p10$	$p25$	$p50$	$p75$	$p90$
Demand Volatility	4.0%	4.0%	6.0%	24.0%	30.0%

3.3 Productivity Learning Curves

UTR associates at Amazon delivery stations perform different tasks varying from inducting delivery packages from long-haul trucks, sorting these packages and organizing them for delivery to performing quality checks and solving various operational problems. To analyze differences in productivity and corresponding learning curves between flex and non-flex associates, we focus on the two key direct processes of delivery stations, process A and process B. For confidentiality reasons, this paper does not go into details of the tasks that these processes require. At a high level, process A is more physically demanding and requires more associate time per package, whereas process B is considered to be a slightly easier process. Overall, process A and process B represent the most time-consuming processes at delivery stations - 25% and 20% of the overall worked hours, respectively. Most of the other processes are indirect, *e.g.,* are not directly related to processing a certain number of packages per hour, which makes it harder to directly observe and measure the corresponding associate productivity. Throughout this paper, we focus on the two direct processes above and extrapolate their results to the entire delivery station.

To measure differences in productivity between flex and non-flex associates at Amazon delivery stations, we used two years of historical associate level data (2023 and 2024). We ran linear regressions for each associate tenure group controlling for other available observable factors such as station and time fixed effects. Even though the absolute numbers varied depending on a time period and a specific station, we found that directionally the results are very robust. This allowed us to estimate the so-called learning curves, the differences in productivity between flex and non-flex associates by tenure. Figure 2 compares learning curves between non-flex and flex associates for process A and process B across different tenure periods. In particular, we show 50-th and 75-th percentiles (respectively denoted P50 and P75) of the productivity distribution across associates for a particular process and particular tenure period. Based on the median productivity in process A, non-flex associates initially demonstrate significantly faster learning rates. However, after 150 hours tenure in that process path, median productivities of both associate types converge, with differences becoming statistically insignificant (though point estimates remain slightly higher for tenured non-flex associates). The analysis of the 75-th percentile shows that among the most productive associates, a persistent productivity gap of 5% favors non-flex associates, even among veteran workers.

By contrast, Process B exhibits very similar productivity and learning rates for non-flex and flex new hires with overall tenure in the process below 100 hours. Beyond this point, the productivity of non-flex associates continues to grow, although at a slower rate, while flex associates' productivity plateaus. This divergence creates a more pronounced productivity gap in process B between non-flex and flex associates with tenure between 100 and 400 hours. Eventually, after 500 hours of tenure, median productivity levels re-converge between the two associate types.

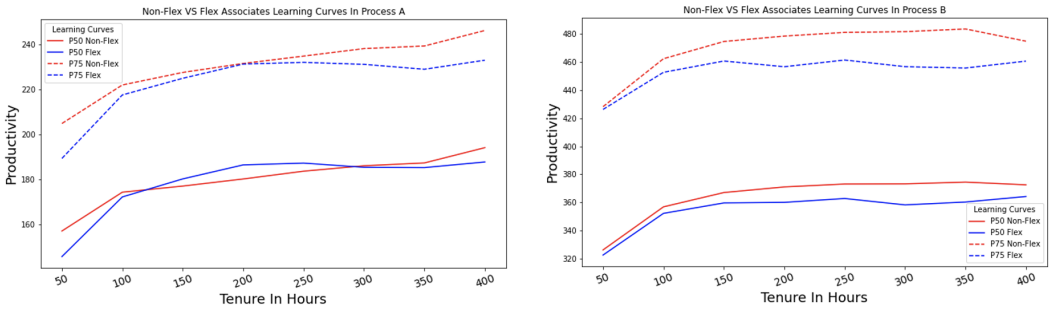

Figure 2: P50 and P75 Learning Curves for Flex and Non-Flex Associates in Key Processes

Overall, we find that a small difference in productivity persists even between highly tenured flex and non-flex associates. Since flex associates work fewer hours per week, it takes longer to achieve the same level of tenured productivity and in some cases a small gap remains over the long-term.

In order to estimate the average productivity difference between flex and non-flex labor pools, one needs to weigh the corresponding learning curve tenure productivity by the share of flex and non-flex associates with that tenure. Across both process A and process B we find that non-flex associates are nearly 6% more productive on average compared to flex associates (see Table 12 in Appendix A). Therefore, in a capacity constrained environment when a station operates at its maximum mechanical capacity, leveraging 100% non-flex labor pool may result in almost 6% more customer packages processed in delivery stations compared to leveraging a 100% flex labor pool. Our estimates are consistent with existing literature [Kim et al., 2025, Specchia and Vandenberghe, 2013, Devicienti et al., 2018], which finds that part-time workers tend to be slightly less productive, especially in the case of short part-timers, which better represent the flex associates in our case study. We leverage our learning curves estimates to inform assumptions about productivity for flex and non-flex associates in our simulations (see Section 4.3).

3.4 Planning Uncertainty

When planning labor for UTR operations, besides potential demand volatility planners have to account for possible supply volatility, which in this context refers to uncertainty around the labor pool's ability to fill and eventually work the shifts. For non-flex associates, the only uncertainty is around shift attendance. For flex associates, uncertainty is represented by the product of flex shift acceptance rate and flex attendance. Similarly to our approach for learning curves estimation above, we leverage a linear regression at the delivery station level to measure differences in uncertainty between flex and non-flex associates controlling for other observable factors such as shift characteristics, and station and week fixed effects. Our estimates indicate that flex uncertainty relative to non-flex is approximately 80%. From a planning perspective, this difference in relative reliability leads to a trade-off between potential overstaffing due to scheduling more flex associates than needed to account for uncertainty in shift acceptance versus potential risk of understaffing in case demand increases and there is no ability to flex up non-flex labor.

3.5 Summary

Based on the above analysis of Amazon's UTR environment, including relevant data, we can frame the trade-off between non-flex and flex associates as follows. Flex associates offer higher flexibility, allowing improved customer service in the presence of uncertain and fluctuating demand. Furthermore, flex shifts are typically shorter in length (4-5 hours) than non-flex shifts (7-10 hours), allowing reduced under- and over-staffing when labor need fluctuates throughout the day. However, these benefits

come with a degradation in productivity (up to 6%) and higher labor supply uncertainty (flex relative reliability of 80%). Hence, the advantages of flex labor in accommodating demand shocks over time and providing greater structural efficiency need to be balanced with potential downsides of lower productivity and planning uncertainty. These are the key aspects that an optimization framework has to balance in order to recommend an effective mix of flex and non-flex associates given business priorities.

4 Optimization Framework

We present a two-stage optimization framework for determining an optimal mix of non-flex and flex associates. Our framework optimizes workforce composition for individual delivery stations independently, with each station's decisions based on its own operating conditions. The approach separates decisions by time dimension and source of volatility: SEM optimizes hour-level staffing patterns and shift structures given labor need volatility throughout an operations day, while CM determines weekly-level hiring and employment decisions that accommodate week over week volatility in demand. This intentional decomposition serves multiple purposes: First, it makes the problem computationally tractable by avoiding the combinatorial complexity that would result from an integrated model. An integrated approach would require simultaneously modeling both intra-week (*i.e.,* hourly) demand variability (to capture structural efficiency) and week-to-week volume fluctuations, leading to a multiplicative increase in scenario requirements and a substantially larger integer program, which can be computationally prohibitive at network scale. Second, it naturally aligns with different business decisions: operational efficiency through shift structure optimization (via SEM) and strategic workforce planning through cost-optimal hiring decisions (via CM). Third, the decomposition enhances model interpretability, allowing planners to focus on specific optimization aspects, *i.e.,* structural efficiency versus cost optimization, facilitating practical implementation and adoption. Finally, it addresses the two key distinct aspects of flex labor: structural efficiency due to shorter flexible shifts and the ability to better accommodate demand shocks over time due to higher flex-up potential.

Non-flex associates provide more stability and less planning uncertainty. On average, each non-flex associate works a recurring schedule with a total of H_f hours per week. Compared to a flex associate, a non-flex associate typically achieves a higher productivity rate π_f due to their consistent schedules and longer tenures. Their reliability rate ρ_f, given by their attendance rate, is generally higher, providing more reliable labor need coverage. Non-flex associate's rigid schedules can lead to overstaffing during low-demand periods or understaffing during peak times. To mitigate this limitation, a planner can enact VET and VTO to, respectively, increase or decrease hours worked by non-flex associates in a week; past experience indicates that each non-flex associate may accept up to $H^{\max}_{\text{VET}}$ VET hours and $H^{\max}_{\text{VTO}}$ VTO hours per week, with associated per-hour costs for the business of C^{VET} and C^{VTO}.

On the other hand, flex associates provide flexibility in labor planning. They pick up shifts on demand, allowing for more precise matching of labor to variable demand patterns. On a weekly basis, a flex associate is required to work a minimum of $H^{\min}_r$ hours, and (on average) works a maximum of $H^{\max}_r$ hours. Both flex and non-flex associates incur the same wage rate C^{wage}, recruiting costs C^{hire}, and weekly attrition rate γ. However, flex associates work fewer hours per week ($H^{\max}_r < H_f$), ultimately providing a lower overall productivity ($\pi_r < \pi_f$) due to slower tenure progression. Furthermore, given that flex associates must pick up their shifts and also attend, there is more uncertainty around their attendance ($\rho_r < \rho_f$).

Planners consider several key trade-offs when determining the optimal mix of flex and non-flex associates. Flex associates provide enhanced flexibility to adapt to demand fluctuations, but at the expense of the consistency, higher productivity, and lower uncertainty offered by non-flex associates. Non-flex associates' higher productivity is crucial during high-demand periods or when operating near capacity constraints, but maintaining a large non-flex workforce during low-demand periods can lead to costly overstaffing. Lastly, flex associates' shorter, more adaptable shifts allow for better alignment with hourly demand patterns when labor need fluctuates throughout the day, while non-flex associates' longer shifts may result in misalignment due to hourly demand variability. By carefully balancing these factors and leveraging the strengths of both associate types, a planner seeks to determine an optimal workforce composition that minimizes total labor costs while meeting service level requirements and accommodating demand variability.

4.1 Structural Efficiency Model (SEM)

We start by defining the relevant sets used in SEM to represent the various components of workforce planning. The set Ω_1 represents a collection of weekly demand scenarios, capturing the variability in labor needs across different demand realizations. To model time granularity, for any $\delta > 0$ we define $\mathcal{T}_\delta$ as the set of δ-hour time periods comprising a week. This allows for flexible time discretization to match operational needs. The set $\mathcal{F}$ encompasses all possible weekly recurring non-flex associate cohorts, while $\mathcal{F}_t \subseteq \mathcal{F}$ represents the subset of cohorts that overlap with time period $[t, t+\delta) \in \mathcal{T}_\delta$. Similarly, $\mathcal{R}$ denotes the set of all flex shift structures that can be planned throughout a week, with $\mathcal{R}_t \subseteq \mathcal{R}$ being the subset of flex shifts that overlap with time period $[t, t+\delta) \in \mathcal{T}_\delta$. For non-flex cohort or flex shift $q \in \mathcal{F} \cup \mathcal{R}$ and time period $[t, t+\Delta)$, we denote by $T_{q,t}$ the duration of the time overlap between them.

SEM incorporates several key parameters to capture the nuances of workforce planning. The scenario set Ω_1 is specifically designed to capture intra-week labor need variability at a sub-daily granularity. Each scenario $\omega \in \Omega_1$ represents a different realization of labor requirements across time windows throughout a week, enabling SEM to evaluate how different shift structures and staffing patterns perform under various within-week demand patterns. We assume that in all scenarios, the labor need is distributed throughout the week across Δ-hour time windows, $\Delta > 0$. Thus, d_t^ω represents the labor need in hours associated with time window $[t, t+\Delta)$ in scenario $\omega \in \Omega_1$. The average weekly total labor need is given by $\overline{W} = |\Omega_1|^{-1} \sum_{\omega \in \Omega_1} \sum_{(t,t+\Delta) \in \mathcal{T}_\Delta} d_t^\omega$. For non-flex associates, h_p denotes the total number of shift hours per week scheduled for an associate in cohort $p \in \mathcal{F}$. In addition, the model incorporates demand coverage guardrails at the time window level, $f_t, [t, t+\Delta) \in \mathcal{T}_\Delta$, to enforce that at least a certain fraction f_t of demand d_t^ω is covered by non-flex associate hours, for each week scenario. The guardrails can be used for risk mitigation, *e.g.,* to ensure the presence of high-productivity non-flex associates during certain time periods.

The model balances understaffing and overstaffing, with τ indicating the relative importance of understaffing. Understaffing captures service and overstaffing captures cost considerations. Let $\pi^* = \max\{\pi_f, \pi_r\}$; since labor need is already in hours, SEM accounts for associate type productivity as factors relative to the best productivity among associate types, *i.e.,* the flex and non-flex productivity factors correspond to $\hat{\pi}_r = \pi_r/\pi^*$ and $\hat{\pi}_f = \pi_f/\pi^*$. Importantly, the productivity rates are averaged over different tenure level of associates (which depend on attrition and hours worked), and thus account for associates' learning. Finally, $HC_{\max}$ sets the maximum total headcount allowed in the facility at all times, to account for the capacity constraints in the facility. These parameters collectively enable the model to account for various operational constraints and workforce characteristics in optimizing the staffing levels.

The SEM model employs several decision variables to determine the optimal workforce composition. The variable x_p represents the number of non-flex associates hired to work cohort $p \in \mathcal{F}$. For the flex workforce, y_s^ω denotes the number of flex associates assigned to flex shift $s \in \mathcal{R}$ in scenario $\omega \in \Omega_1$. Note that non-flex cohort decisions x_p do not depend on the scenario as cohorts repeat their labor pattern every week; by contrast, flex staffing decisions y_s^ω can vary across scenarios, hence the use of the index over Ω_1. To account for mismatches between staffing and labor needs, the model uses o_t^ω and u_t^ω to represent the number of overstaffed and understaffed hours, respectively, during time window $[t, t+\Delta) \in \mathcal{T}_\Delta$ in scenario $\omega \in \Omega_1$. For convenience, an overview of the notation used in the presentation of the SEM model is provided in Appendix B (Table 17). Model (1) shows the complete formulation of SEM; to simplify the presentation of the model, we abuse of notation and

use $t \in \mathcal{T}_\delta$ to mean $[t, t+\delta) \in \mathcal{T}_\delta$.

$$\min \sum_{\omega \in \Omega_1} \sum_{t \in \mathcal{T}_\Delta} (o_t^\omega + \tau u_t^\omega) \tag{1a}$$

$$\text{s.t. } d_t^\omega + o_t^\omega - u_t^\omega = \sum_{p \in \mathcal{F}_t} \rho_f \hat{\pi}_f T_{t,p} x_p + \sum_{s \in \mathcal{R}_t} \rho_r \hat{\pi}_r T_{t,s} y_s^\omega \qquad \forall \omega \in \Omega_1,\ \forall t \in \mathcal{T}_\Delta \tag{1b}$$

$$\sum_{p \in \mathcal{F}_t} x_p + \sum_{s \in \mathcal{R}_t} y_s^\omega \leq HC_{\max} \qquad \forall \omega \in \Omega_1,\ \forall t \in \mathcal{T}_{0.25} \tag{1c}$$

$$\sum_{p \in \mathcal{F}_t} \rho_f \hat{\pi}_f T_{t,p} x_p \geq f_t \max_{\omega \in \Omega_1} d_t^\omega \qquad \forall t \in \mathcal{T}_\Delta \tag{1d}$$

$$x_p \in \mathbb{Z}_0^+ \qquad \forall p \in \mathcal{F} \tag{1e}$$

$$y_s^\omega \geq 0 \qquad \forall \omega \in \Omega_1,\ \forall s \in \mathcal{R} \tag{1f}$$

$$o_t^\omega, u_t^\omega \geq 0 \qquad \forall \omega \in \Omega_1,\ \forall t \in \mathcal{T}_\Delta \tag{1g}$$

Objective (1a) minimizes the overstaffing and understaffing cost across hours and weeks. Constraints (1b) determine overstaffing and understaffing hours given the choices of non-flex cohorts and flex shifts. Constraints (1c) prevent planning associate headcount beyond the maximum headcount allowed by the facility. Constraints (1d) enact the demand coverage guardrails for non-flex associates. Constraints (1e)–(1g) specify the domains of the decision variables.

It is important to note that SEM does not directly optimize workforce composition, but rather minimizes overstaffing and understaffing. This approach can result in multiple optimal solutions that achieve the same level of staffing efficiency but with different workforce compositions. Consequently, a single solution for Model (1) may not provide sufficient insight into the range of acceptable workforce compositions.

To gain a comprehensive understanding of viable workforce compositions, we introduce the following constraint in the model:

$$\theta_{\min} \overline{W} \leq \sum_{p \in \mathcal{F}} h_p x_p \leq \theta_{\max} \overline{W} \tag{2}$$

Constraint (2) ensures that the number of associate hours from non-flex cohorts falls within specified minimum and maximum targets $\theta_{\min}$ and $\theta_{\max}$, relative to the average total weekly need $\overline{W}$. By doing so, we effectively optimize structural efficiency for a limited set of workforce compositions.

To identify the range of optimal workforce compositions, we perform a search over small intervals $[\theta_{\min}, \theta_{\max}]$ by solving Model (1) for each interval. Specifically, we identify the intervals with understaffing and overstaffing less than ε_1, where $\varepsilon_1 \geq 0$. Let Θ_{ε_1} be the set of such intervals, let $\Theta^* = \{\theta : \exists [\theta_1, \theta_2] \in \Theta_{\varepsilon_1} : \theta \in \{\theta_1, \theta_2\}\}$ be the set of boundary points of these intervals, and let $\theta^*_{\min} = \min_\theta \Theta^*$ and $\theta^*_{\max} = \max_\theta \Theta^*$. The values $\theta^*_{\min}$ and $\theta^*_{\max}$ inform CM of the range of workforce compositions that ensure low understaffing and overstaffing.

It is important to clarify the information flow between SEM and CM. The optimal values of the decision variables x_p and y_s^ω in SEM are not used in CM. Instead, CM operates with its own decision variables, but these are constrained by the range $[\theta^*_{\min}, \theta^*_{\max}]$ to ensure that we search for a cost-optimal workforce composition among the workforce compositions with high structural efficiency identified by SEM. This design allows the two-stage approach to incorporate operational efficiency information without requiring CM to replicate SEM's detailed shift-level decisions.

4.2 Cost Model (CM)

CM builds upon SEM to optimize the overall WFC and associated costs. The combination of demand uncertainty, discrete decisions, and complex operational constraints, leads us again to using a scenario-based optimization model to determine a workforce composition that minimizes overall cost (while maintaining structural efficiency). The model uses a set of weekly demand scenarios Ω_2, with d_ω representing the weekly demand (in units, representing the volume to be processed in a given week) in scenario $\omega \in \Omega_2$. The sample average weekly demand is denoted by $\overline{D}$ and we enforce minimum and maximum non-flex coverage percentages relative to this average weekly demand, $\theta^*_{\min}$ and $\theta^*_{\max}$, respectively (obtained from SEM) to ensure that low understaffing and overstaffing can be

achieved. Unlike SEM's scenario set Ω_1 which focuses on sub-daily labor patterns within weeks, Ω_2 captures week-to-week demand variability at an aggregate level. Each scenario $\omega \in \Omega_2$ represents the total weekly processing volume, allowing CM to optimize WFC decisions that must accommodate longer-term demand fluctuations while relying on SEM's insights about operational efficiency within any given week.

CM employs decision variables that determine the optimal workforce structure and scheduling. Variables x^f and x^r represent the number of non-flex and flex associates to hire, respectively. For scenario $\omega \in \Omega_2$, h^f_ω and h^r_ω denote the number of planned regular hours for non-flex and flex associates. h^{VET}_ω and h^{VTO}_ω represent the number of planned VET and VTO hours, respectively, for non-flex associates in scenario ω. For convenience, an overview of the notation used in the presentation of the CM model is provided in Appendix B (Table 18). The mathematical formulation of CM is given by Model (3).

$$
\begin{aligned}
\min\ & C^{\text{hire}}\gamma(x^f + x^r) \\
& + \tfrac{1}{|\Omega_2|} \sum_{\omega\in\Omega_2} \left(C^{\text{wage}}(\rho_f h^f_\omega + \rho_r h^r_\omega) + C^{\text{VET}}\rho_f h^{\text{VET}}_\omega + C^{\text{VTO}} h^{\text{VTO}}_\omega\right) && && \text{(3a)}\\
\text{s.t.}\ & \rho_f \pi_f (h^f_\omega + h^{\text{VET}}_\omega) + \rho_r \pi_r h^r_\omega \geq d_\omega && \omega \in \Omega_2 && \text{(3b)}\\
& x^r H^{\min}_r \leq h^r_\omega \leq x^r H^{\max}_r && \omega \in \Omega_2 && \text{(3c)}\\
& h^f_\omega = x^f H_f - h^{\text{VTO}}_\omega && \omega \in \Omega_2 && \text{(3d)}\\
& h^{\text{VET}}_\omega \leq H^{\max}_{\text{VET}} && \omega \in \Omega_2 && \text{(3e)}\\
& h^{\text{VTO}}_\omega \leq H^{\max}_{\text{VTO}} && \omega \in \Omega_2 && \text{(3f)}\\
& \theta^*_{\min}\overline{D} \leq \rho_f \pi_f H_f x^f \leq \theta^*_{\max}\overline{D} && && \text{(3g)}\\
& x^f, x^r \in \mathbb{Z}^+_0 && && \text{(3h)}\\
& h^f_\omega, h^r_\omega, h^{\text{VET}}_\omega, h^{\text{VTO}}_\omega \geq 0 && \omega \in \Omega_2 && \text{(3i)}
\end{aligned}
$$

Objective (3a) minimizes the total cost from hiring non-flex and flex associates and hourly cost of non-flex and flex shifts across scenarios. Constraints (3b) ensure that a sufficient number of hours is available to process the required volume in each scenario. The computation of volume processed uses the productivity of associates and the actual hours provided by non-flex and flex schedules, *i.e.,* it accounts for the reliability rates. Constraints (3c) dictate that the number of hours assigned to Flex associates is within the lower and upper limits based on contractual minimum and maximum hours worked in each scenario ω. Constraints (3d) calculate the number of regular non-flex scheduled hours in scenario ω, as the regular non-flex weekly hours minus the enacted VTO hours in that scenario. Constraints (3e) and (3f) restrict the maximum number of VET and VTO hours that may be enacted in scenario ω, respectively. Constraints (3g) incorporate the structural efficiency bounds derived from SEM, as percentages of the average total weekly demand that must be covered by non-flex hours; this restricts the solution space to solutions that preserve structural efficiency. Constraints (3h) and (3i) specify the domain of the model's decision variables.

Given optimal solution $(x^{f*}, x^{r*}, h^{f*}_\omega, h^{r*}_\omega, h^{\text{VET}*}_\omega, h^{\text{VTO}*}_\omega)$, the optimal flex percentage is calculated as $(\rho_r \sum_{\omega\in\Omega_2} h^{r*}_\omega)/(\sum_{\omega\in\Omega_2} \rho_r h^{r*}_\omega + \rho_f h^{f*}_\omega + \rho_f h^{\text{VET}*}_\omega)$. In addition, it may be of interest for users to determine the minimum and maximum flex percentage that preserves structural efficiency and total cost. This can be achieved by, respectively, solving Model (3) with either of the Objectives (4a) (to minimize flex hours) or (4b) (to minimize non-flex hours) as secondary objective (in goal programming fashion, with a threshold $\varepsilon_2 \geq 0$).

$$
\min \sum_{\omega\in\Omega_2} h^r_\omega \tag{4a}
$$

$$
\min \sum_{\omega\in\Omega_2} (h^f_\omega + h^{\text{VET}}_\omega) \tag{4b}
$$

Observe that even though the primary objective of the optimization framework is to recommend a workforce composition, the solution to SEM for the recommended workforce composition provides the non-flex cohorts that should be employed and the flex shifts that complement the non-flex hours.

4.3 Simulations

In this section, we present the simulation results from applying our two-stage optimization framework to Amazon delivery stations. For illustration, we apply the framework for two hypothetical stations, Station 1 and Station 2, chosen to demonstrate the applicability and effectiveness of our approach across different operational contexts. Station 1 is located in a major US metropolitan area served by multiple stations. As a result, the average daily volume is relatively low, yet demand is characterized by high volatility, and customers tend to order fewer (and smaller) items, leading to a relatively smaller average package sizes. In contrast, Station 2 is located in a suburban Midwestern area and is the largest station in the area with no nearby delivery stations. Hence, compared to Station 1, it has a 70% higher capacity and processes around 50% higher daily volume on average. Demand is less volatile in this suburban area, and customers prefer ordering more items per order, resulting in larger average package sizes.

Regarding labor supply patterns, both stations experience similar shift attendance patterns. The primary difference is that Station 1's urban location facilitates easier recruitment of flex labor, and flex acceptance of shifts is higher compared to Station 2.

For SEM, we consider $|\Omega_1| = 9$ weeks of station-specific labor need. Moreover, we assume $\Delta = 4$ (since working periods are roughly 4-hour long), a relative understaffing importance factor $\tau = 3$, and guardrails $f_t = 20\%$, $\forall t \in \mathcal{T}_4$. To construct the set of non-flex coverage percentage intervals Θ, we partition the interval [50%,100%] into intervals of size $\alpha = 2.5\%$, and employ a tolerance $\varepsilon_1 = 5\%$. The 50% lower bound reflects the (current) business requirement that at least 50% of needed hours has to be covered with non-flex labor.

For CM, we generate $\Omega_2 = 3000$ weekly scenarios to ensure a robust representation of demand variability. These scenarios are constructed for each station assuming normally distributed weekly volumes, based on station-specific historical data. We compute each station's weekly average volume $\overline{D}$ and coefficient of variation ξ to sample the scenarios. We assume non-flex associates work $H_f = 40$ hours per week, with VET and VTO weekly limits of $H_{\text{VET}}^{\max} = 2$ and $H_{\text{VTO}}^{\max} = 4$ hours, respectively. For flex associates, we assume a minimum weekly requirement of $H_r^{\min} = 4$ hours and a maximum of $H_r^{\max} = 24$ hours per associate.

We assume hiring costs, wage costs, VET, and VTO hourly costs satisfy $C^{\text{hire}} = 84C^{\text{wage}}$, $C^{\text{VET}} = 1.3C^{\text{wage}}$ and $C^{\text{VTO}} = 0.25C^{\text{wage}}$. For differences in productivity and reliability, we performed a linear regression analysis using two years of historical associate level data (see Sections 3.3 for more details). Our assumptions for productivity differences are in line with existing literature [Kim et al., 2025, Specchia and Vandenberghe, 2013, Devicienti et al., 2018]. The overall flex and non-flex productivities are estimated at $\pi_f = 376$ and $\pi_r = 364$ demand units per hour, respectively. We consider a non-flex reliability of $\rho_f = 85\%$. Lastly, CM's goal programming scheme uses a threshold of $\varepsilon_2 = 0.5\%$. Employed station-specific parameter values are listed in Table 2.

Table 2: Station-Specific Parameter Values

Station	$HC_{\max}$	γ	ρ_r	$\overline{D}$	ξ
1	200	1.7%	69.6%	267,000	6.0%
2	400	5.3%	55.0%	405,000	3.3%

Figures 3 and 4 show, respectively, for Stations 1 and 2, the obtained values of structural efficiency, given by the sum of average weekly understaffing and overstaffing hours relative to the average weekly labor needs, for the different non-flex coverage intervals in Θ. Structural efficiency deteriorates in two cases. For a high non-flex coverage relative to the average total weekly need, the rigid weekly non-flex schedules can lead to overstaffing during periods with fluctuating or below-average need. This is observed for both stations: for Station 1, when non-flex coverage exceeds 87.5-90%, it causes overstaffing as the high non-flex labor surpasses the labor need for certain weekly scenarios. The same is observed for Station 2 when the non-flex coverage exceeds 90-92.5%. Specifically, for Station 1, operating beyond 90% non-flex coverage generates overstaffing equivalent to about 1% of average weekly labor requirements. For Station 2, exceeding 92.5% non-flex coverage results in overstaffing of up to 0.6% of average labor needs. Although not observed for either station, structural efficiency can deteriorate when non-flex coverage falls too low; in these cases, capacity constraints (*i.e.,* maximum headcount) may prevent adding enough flex associates to meet labor needs effectively,

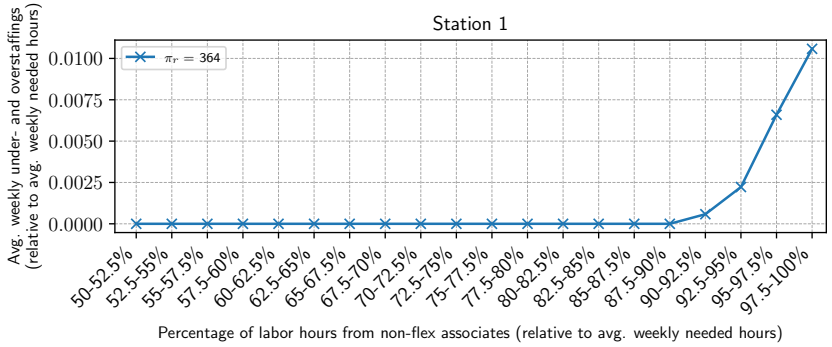

Figure 3: SEM Results for Station 1 - Sum of Average Weekly Understaffed and Overstaffed Hours (Relative to Average Weekly Total Needed Hours) for Different Non-Flex Coverage Ranges

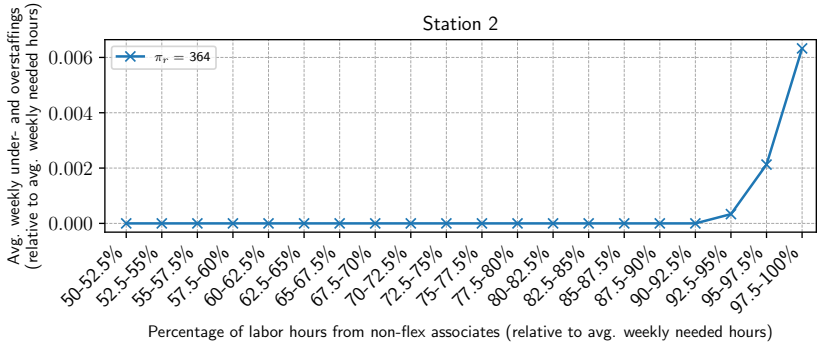

Figure 4: SEM Results for Station 2 - Sum of Average Weekly Understaffed and Overstaffed Hours (Relative to Average Weekly Total Needed Hours) for Different Non-Flex Coverage Ranges

incurring understaffing. The main reason for this is potentially lower productivity due to longer transition through learning curves.

Each curve defines an optimal non-flex coverage percentage range where structural efficiency is maximized, see Table 3. Note that the ranges for both stations are similar, with $\theta^*_{\min}$ equal to 50% for both, and $\theta^*_{\max}$ equal to 90% and 92.5% for Stations 1 and 2, respectively. This is the optimal range for non-flex hours recommended by the SEM model from a structural efficiency perspective.

Table 3: Structural Efficiency Bounds Derived from SEM, as Percentages of Average Total Weekly Demand that Must Be Covered by Non-Flex Hours by Station

Station	$\theta^*_{\min}$	$\theta^*_{\max}$
1	50%	90%
2	50%	92.5%

Figure 5 presents the optimal flex associate percentage ranges for Stations 1 and 2, each range derived by solving the CM with secondary objectives (4a) and (4b). Although the SEM ranges for both stations are closely aligned, these results show that stations exhibit different optimal flex ranges based on their operational characteristics. Station 1 exhibits comparatively higher weekly and hourly demand volatility, as well as less uncertainty in non-flex associates attendance, overall favoring a higher flex workforce to accommodate fluctuating labor needs. The CM captures these factors, recommending a higher flex percentage between 12% and 19%. In contrast, Station 2 illustrates how stability in operations can drive workforce composition decisions. The station experiences comparatively lower fluctuations in both weekly and hourly demand patterns. This stability reduces the need for flexible labor, as non-flex associates can effectively cover the more predictable demand patterns with less planning uncertainty. In addition, Station 2's flex associates exhibit more uncertainty in their acceptance and attendance of flex shifts. The CM confirms the cost-effectiveness of a non-flex-

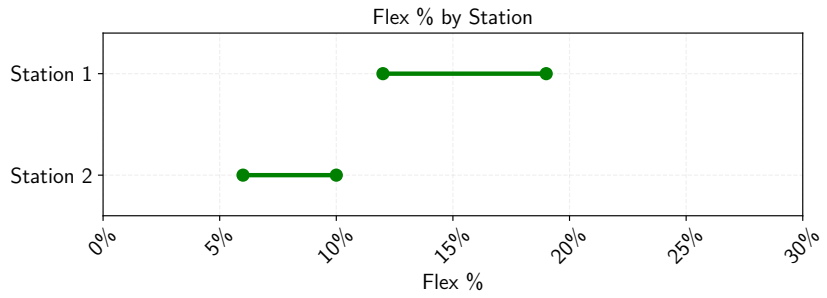

Figure 5: CM Results for Stations 1 and 2 - Recommended Flex Percentage Ranges from Solving CM Considering Secondary Objectives (4a) and (4b), Respectively

heavy workforce composition. As a result, the models recommend a lower optimal flex percentage for Station 2, ranging from 6% to 10%. Intuitively, these results are congruent with the market conditions where these stations are located. Being located in a major US city with relatively low but volatile demand and higher flexible labor supply, Station 1 benefits from leverage more flexible labor. Whereas Station 2 located in a suburban area in the Midwest faces a much higher but more stable demand and doesn't benefit as much from flexible labor as Station 1.

To quantify the potential benefits of increasing the flex workforce from 1% to 10%, which was done in the experimental trial to be discussed in the next section, we conduct a network-wide analysis comparing scenarios with 1% and 10% flex percentages using average data across all stations. In SEM, we set a high penalty ratio for the understaffing importance factor τ to strongly discourage understaffing hours, aligning this with a no-understaffing constraint in CM. This is in line with the goal to serve all customer demand possible. Overstaffing costs are then calculated by multiplying the excess labor hours by the hourly cost. CM is then executed with a constraint that forces the ratio between the flex recommended hours and the total recommended associate hours to be 1% and 10%, respectively, allowing deviations of at most $\varepsilon_3 = 0.25\%$. To obtain the total cost impact, we combine each scenario's SEM overstaffing costs with its corresponding CM total cost, after removing CM's overstaffing costs to prevent double-counting. The results suggest that increasing the flex percentage from 1% to 10% yields average efficiency improvements up to 3.9%.

4.4 Sensitivity Analysis

This section provides a sensitivity analysis on the results from SEM with respect to flex associate productivity π_r, and on the results from CM with respect to flex associate productivity π_r and demand volatility ξ.

4.4.1 SEM Sensitivity to Flex Associate Productivity

We analyze the effect of varying flex productivity π_r on the optimal non-flex coverage percentage range $[\theta^*_{\min}, \theta^*_{\max}]$ determined by SEM. Table 4 presents the values of π_r considered for this analysis, expressed relative to non-flex productivity π_f (note that these include the base case $\pi_r = 364$ from Section 4.3). All other parameters are held constant at values from Section 4.3.

Table 4: Sensitivity Analysis - Considered Values of π_r

Station	Values of π_r
1 2	$\left\{\frac{i}{10}\pi_f, i \in \{1, 2, \ldots, 10\}\right\} \cup \{364\}$

Figures 6 and 7, respectively, show, for Stations 1 and 2, the resulting overstaffed and understaffed hours for different values of π_r and non-flex coverage percentages. The results show that low flex productivity leads to understaffing due to capacity constraints when non-flex coverage percentage decreases.

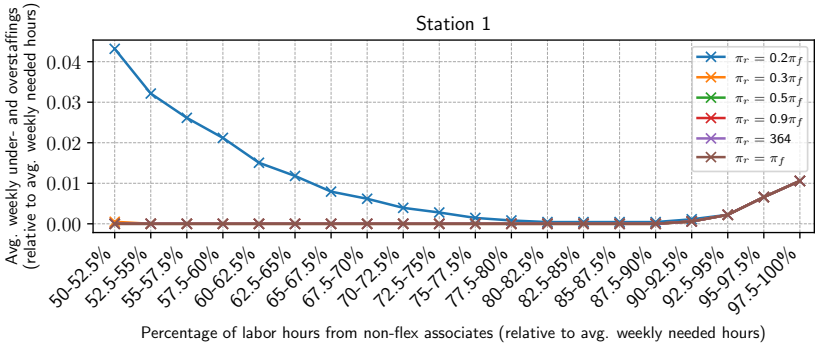

Figure 6: SEM Results for Station 1 - Sum of Average Weekly Understaffed and Overstaffed Hours (Relative to Average Weekly Total Needed Hours) for Different Non-Flex Coverage Ranges and Flex Productivity Values

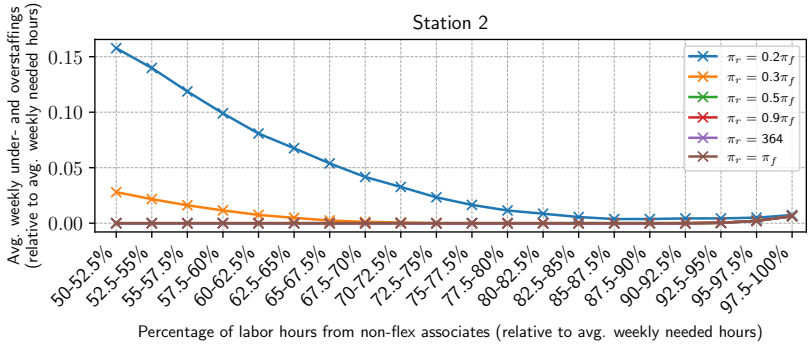

Figure 7: SEM Results for Station 2 - Sum of Average Weekly Understaffed and Overstaffed Hours (Relative to Average Weekly Total Needed Hours) for Different Non-Flex Coverage Ranges and Flex Productivity Values

For Station 1, when π_r drops to $0.2\pi_f$, the station is unable to meet labor requirements within 50-80% non-flex coverage range, incurring understaffing equivalent to approximately 4% of average weekly labor needs. Across all π_r values tested, exceeding 90% non-flex coverage results in overstaffing of up to 1% of the average weekly labor requirements.

Station 2 exhibits greater sensitivity to productivity reductions. At $\pi_r = 0.2\pi_f$, understaffing occurs across all non-flex coverage levels below 85%, reaching 15% of average weekly labor needs at 50% of non-flex coverage. At $\pi_r = 0.3\pi_f$, understaffing is still observed although to a lower degree, occurring within the 50-67.5% non-flex coverage range, and peaking at 3% of the average weekly needs at 50% non-flex coverage. Higher values of π_r produce results consistent with the base case ($\pi_r = 364$) reported in Section 4.3.

For each station and tested value of π_r, the structural efficiency bounds determined by SEM are listed in Table 5. Both stations show that low productivity levels ($\pi_r \leq 0.2\pi_f$) constrain workforce composition options by requiring higher non-flex coverage requirements, and that this constraint becomes less binding for higher productivity values ($\pi_r \geq 0.4\pi_f$). Despite Station 2 having a larger max headcount, the relatively larger labor needs result in overall higher understaffing, compared to Station 1, for low flex productivity levels. These results show that productivity differentials affect workforce composition flexibility, with greater effects observed under capacity constraints.

4.4.2 CM Sensitivity to Flex Productivity

Next, we study how varying flex productivity π_r affects the optimal flex coverage percentage and optimal total cost yielded by CM. Again, we consider the values of π_r shown in Table 4. All other parameters are set to the values indicated in Section 4.3. The results can be found in Figures 8 and 9.

Table 5: Structural Efficiency Bounds Derived from SEM, as Percentages of Average Total Weekly Demand that Must Be Covered by Non-Flex Hours by Station, for Multiple Flex Associate Productivity Values

Station	π_r	$\theta^*_{\min}$	$\theta^*_{\max}$
1	$0.1\pi_f$	87.5%	100.0%
	$0.2\pi_f$	75.0%	97.5%
	$0.3\pi_f$	52.5%	90.0%
	$\geq 0.4\pi_f$	50.0%	90.0%
2	$0.1\pi_f$	90.0%	100.0%
	$0.2\pi_f$	80.0%	100.0%
	$0.3\pi_f$	67.5%	97.5%
	$0.4\pi_f$	52.5%	92.5%
	$\geq 0.5\pi_f$	50.0%	92.5%

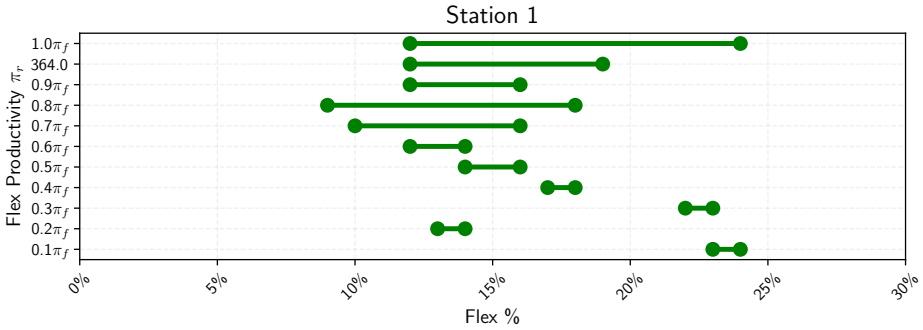

Figure 8: CM Results for Station 1 for Multiple Flex Associate Productivity Values - Recommended Flex Percentage Ranges from Solving CM Considering Secondary Objectives (4a) and (4b), Respectively

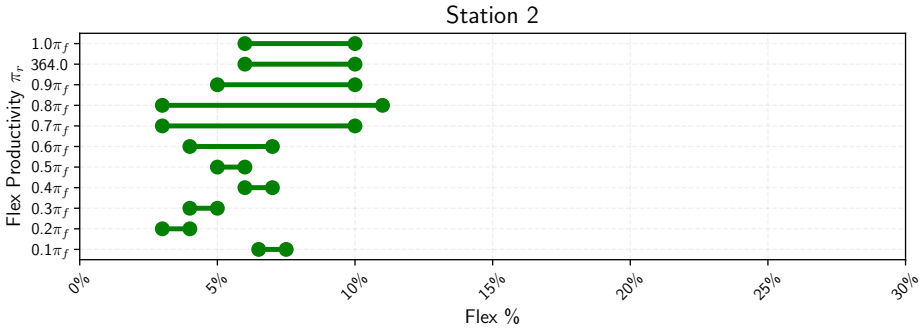

Figure 9: CM Results for Station 2 for Multiple Flex Associate Productivity Values - Recommended Flex Percentage Ranges from Solving CM Considering Secondary Objectives (4a) and (4b), Respectively

To interpret these results, it helps to divide the flex productivity values into three groups: low ($0.1\pi_f$ thru $0.3\pi_f$), moderate ($0.4\pi_f$ thru $0.8\pi_f$) and high ($0.9\pi_f$ thru π_f). When flex productivity is low, a large number of flex associates is required to process even a small portion of demand. As a consequence, even modest demand coverage by flex associates becomes infeasible due to the maximum headcount allowed in a facility. Therefore, when productivity is low the recommended flex percentages are strongly influenced by structural efficiency considerations. When flex productivity is moderate, a wide range of flex percentage are structurally efficient and cost considerations drive the decisions. We observe the expected behavior: when flex productivity increases, the lower end of the structural efficiency-based range decreases, which reflects that fewer flex associates are needed and each flex associate works a smaller number of hours. In fact, the total number of flex associate hours as well as the total number of non-flex associate hours reduces. When flex productivity is high (*i.e.,* little productivity difference between flex and non-flex associates), the decisions are driven by the differences in the demand scenarios. A careful trade-off is made between the value of having one non-flex associate versus two flex associates. Having one non-flex associate incurs only one hiring charge, but is also a commitment to at least 36 hours a week, whereas two flex associates incur two hiring charges, but can work as little as 8 hours a week and as much as 48 hours a week. Although not shown in the figures, the total cost monotonically decreases when the flex associate productivity increases.

4.4.3 CM Sensitivity to Demand Volatility

To evaluate how demand volatility affects cost-optimal flex percentages and optimal total cost determined by CM, we conduct a sensitivity analysis on the volatility parameter ξ. We examine a range of volatility levels around each station's baseline value from Section 4.3: for Station 1, we vary ξ from 3.0% to 15% around the base level of 6.0%, while for Station 2, we vary ξ from 1.7% to 8.3% around the base level of 3.3%. The specific values tested for each station are provided in Table 6.

Table 6: CM Sensitivity Analysis - Considered Values of ξ

Station	Values of ξ
1	$3.0\%, 6.0\%, 9.0\%, 12.0\%, 15.0\%$
2	$1.7\%, 3.3\%, 5.0\%, 6.7\%, 8.3\%$

Figures 10 and 11 show the cost-optimal flex percentages for Stations 1 and 2, respectively, over different volatility levels.

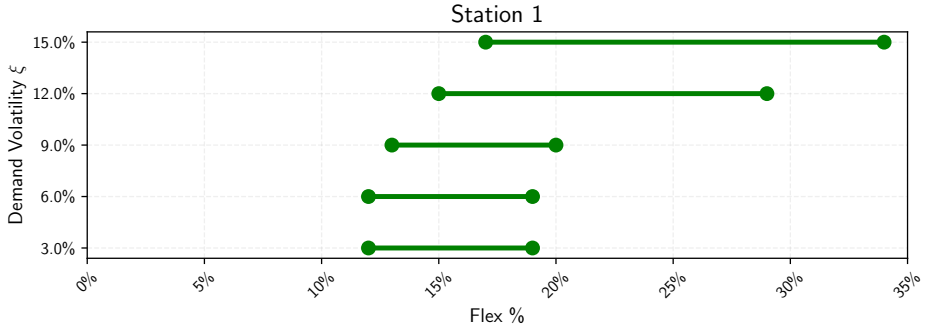

Figure 10: CM Results for Stations 1 Over Multiple Demand Volatility Scenarios - Recommended Flex Percentage Ranges From Solving CM Considering Secondary Objectives (4a) and (4b), Respectively

For Station 1, optimal flex percentages remain stable at 12-19% for low volatilities ($\xi \in \{3.0\%, 6.0\%\}$), with a slight increase to 13-20% at $\xi = 9.0\%$. However, a significant shift occurs at higher volatility values: 15-29% at $\xi = 12.0\%$ and 17-34% at $\xi = 15.0\%$ volatility. This suggests that once volatility exceeds 9.0-12.0%, the flexibility benefits of flex labor are more pronounced, leading to both higher recommended flex percentages and wider optimal ranges.

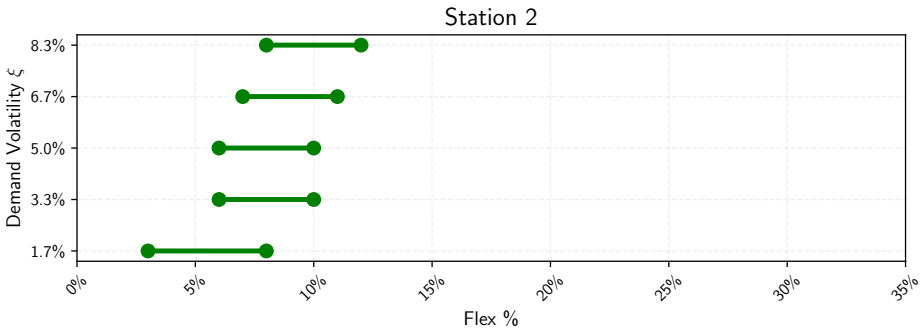

Figure 11: CM Results for Stations 2 Over Multiple Demand Volatility Scenarios - Recommended Flex Percentage Ranges From Solving CM Considering Secondary Objectives (4a) and (4b), Respectively

Station 2 exhibits a less pronounced sensitivity to changes in demand volatility. Flex percentages increase steadily from 4-8% at $\xi = 1.7\%$ to 8-12% at $\xi = 8.3\%$, with the base case ($\xi = 3.3\%$) yielding a flex percentage of 6-10%. Station 2 maintains relatively narrow and stable ranges across all tested volatility levels, indicating more predictable workforce composition requirements even under varying demand conditions.

Total costs increase with demand volatility for both stations. For Station 1, using the $\xi = 3\%$ scenario as baseline (relative cost = 1.0), costs increase to 1.007 (+0.7%) at $\xi = 6\%$, 1.017 (+1.7%) at $\xi = 9\%$, 1.034 (+3.4%) at $\xi = 12\%$, and 1.048 (+4.8%) at $\xi = 15\%$. Station 2 exhibits a similar pattern, with costs relative to the $\xi = 1.7\%$ baseline increasing to 1.015 (+1.5%) at $\xi = 3.3\%$, 1.027 (+2.7%) at $\xi = 5.0\%$, 1.054 (+5.4%) at $\xi = 6.7\%$, and 1.071 (+7.1%) at $\xi = 8.3\%$. This increase in total cost occurs since higher demand volatility requires greater workforce flexibility to accommodate demand fluctuations. As volatility increases, CM recommends increased total associate headcount and higher flex percentages to maintain service levels, resulting in higher total workforce costs. Both stations show comparable cost increase per unit of volatility change, with Stations 1 and 2's costs increasing approximately 1.08% and 0.9% per percentage point increase in volatility, respectively.

5 Empirical Evidence

5.1 Flex Experiment

In order to validate and support the findings of the workforce composition optimization model, we ran an experiment at a group of Amazon delivery stations in 2023. In particular, we dialed up flex (part-time) employees from 1-2% to 10% of total worked hours in 11 pilot delivery stations. The flex employees were primarily used in two key direct processes—process A and B. Prior to the experiment only a small group of Amazon stations had part-time labor and the share of part-time hours never exceeded 1-2% of the overall worked hours. In order to run the experiment, we randomly chose 14 delivery stations within each region of operations in the US and Canada. Due to concerns about the experiment in stations with tight operational capacity constraints and with staffing shortages, we were able to align with operational leaders on launching the experiment at scale in 11 delivery stations out of initial 14 randomly selected stations. The initial hiring ramp up of new flex associates in the treatment stations started in June 2023 and accelerated in July 2023. We selected the period from March-May as the baseline pre-treatment period for measurement (from week 9 through week 22). This is because January-February data are often impacted by post-holiday season of lower demand. We defined treatment period from end of July through end of October (week 29 through week 42). This is because we wanted to avoid the initial period of non-uniform hiring of new flex associates in June and July as well as the Prime week of deals at Amazon. At the same time, we did not want to include the noisy holiday season of 2023 into our measurement because all stations throughout the network increase the number of associates to accommodate the high demand period. Given quasi-randomness of treatment stations selection and high volatility of delivery stations operations, we selected a group of 75 control stations that satisfied the following criteria: (1) the percentage

of weekly flex hours in each key process was below 5% of total worked hours during the baseline pre-treatment period and (2) the week-over-week change in the percentage of flex hours in the total worked hours was no more than 1% during both the baseline and experiment period. Figure 12 shows the average percentage of flex hours in total weekly hours for the treatment and control sites before and after the experiment. The experiment led to significant increase in the percentage of flex hours from 1% to 8%-9% in the treatment stations versus virtually no change in the percentage of flex hours in the control stations.

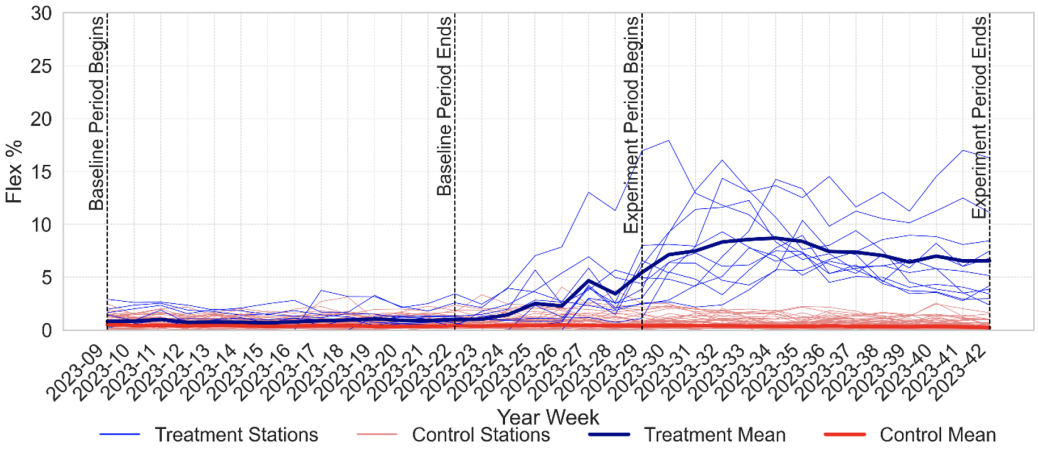

Figure 12: Percentage of Flex Hours in Total Hours Before and After Experiment in the Treatment and Control Stations

5.2 Measurement Methodology

Since the choice of stations into treatment was quasi-random (required approval from Operations given our random intent to treat) and delivery stations metrics tend to be noisy, we decided to leverage a difference-in-difference measurement strategy, *e.g.,* compare the metrics of interest before and after the experiment for the treatment and control sites. Table 7 shows that before the experiment the treatment stations had slightly higher percentage of flex hours (0.5% vs 0% in process A and 1.4% vs 0% in process B). The control stations did not have any significant change in the percentage of flex hours. Whereas the treatment stations experienced the average increase in flex hours by 9% in process A and 8% in process B. Given that the stations did not use flex labor in other processes, and together these two processes represent nearly half of the overall hours worked at a station, we chose to measure the impact at the process level rather than at the station level. We focus our measurement on productivity (packages per hour), and proxies for efficiency—non-flex hours worked per 1,000 process packages and flex hours worked per 1,000 packages. Overall, despite slight differences, these metrics were comparable between treatment and control stations during the baseline (pre-treatment) period. Comparing these metrics before the experiment (March-May 2023) and during the experiment (July-October) between the treatment and control sites in processes A and B enables us to measure the causal impact of higher percentage of flex labor.

Table 7: Key Metrics in Control and Treatment Stations Before and After Experiment

Process Path	Treatment/Control Group	% Of Flex Hours		Productivity Rate		Non-Flex Hours Per 1000 Packages		Flex Hours Per 1000 Packages	
		Baseline Period	Experiment Period	Baseline Period	Experiment Period	Baseline Period	Experiment Period	Baseline Period	Experiment Period
Process A	Control Stations	0.00%	0.00%	281.6	275.4	3.5	3.6	0.0	0.0
	Treatment Stations	0.51%	9.74%	280.2	263.3	3.5	3.4	0.0	0.4
Process B	Control Stations	0.00%	0.00%	520.8	490.7	1.9	2.0	0.0	0.0
	Treatment Stations	1.42%	9.27%	513.4	493.4	1.9	1.8	0.0	0.2

We employ the following econometric model:

$$\pi_{pst} = \gamma C_{st} + \alpha_{ps} + \beta_{pt} + \varepsilon_{pst} \tag{5}$$

where:

- π_{pst} is the outcome metric for process p, station s and week t.

- C_{st} is a binary dummy variable equal to 1 if station s is in the treatment group and week t falls between 07/22/23 and 10/30/23, and 0 otherwise.
- α_{ps} represents station and process fixed effect, controlling for pre-existing differences across stations.
- β_{pt} denotes week and process fixed effect, accounting for week-specific time shocks.
- ε_{pst} is the error term, capturing unobserved factors affecting the outcome metric.

We cluster standard errors at the site-week level and estimate this model separately for the two key processes A and B. The coefficient γ is our primary parameter of interest, representing a difference-in-differences estimate, *i.e.,* the average change in the outcome metric between treatment and control groups before and after the experiment. As a robustness check, we also estimate an econometric model with intensity of treatment:

$$\pi_{pst} = \mu Flex_{pst} + \alpha_{ps} + \beta_{pt} + \varepsilon_{pst} \tag{6}$$

where $Flex_{pst}$ is the percentage of flex hours worked in process p in station s and week t. In this case, μ estimates the elasticity of flex hours on the outcome metric. This model provides greater statistical power by directly measuring the intensity of treatment for each station in the treatment group. However, it may introduce potential endogeneity bias if operational managers utilized flex employees differently from non-flex employees (*e.g.,* scheduling more flex hours during lower volume periods or for simpler tasks).

5.3 Impact on Productivity

Table 8 reports the impact of flex labor on productivity. In order to understand whether the data are too noisy or there is no effect, we also estimated a 95% confidence lower bound impact of flex. This should be interpreted as the worst possible impact that dialing up flex could have on productivity. We find that after the initial ramp-up of flex hours up to 8%, productivity in process A decreased 4.5%, the productivity in process B decreased by 0.6%, and the overall station-level productivity decreased by 5.3%. The causal impact on process A overall productivity is statistically significant, the impact on process B overall productivity is not statistically significant but a high 90% confidence lower bound indicates a potential impact of up to a 4% decrease. These results are congruent with our estimates based on differences in learning curves between flex and non-flex employees (see Section 3; we estimated an upper bound of 6%). They also confirm that part-time labor has a more pronounced impact on productivity in more complex processes. These results are robust to using weekly-level data as well as using the statistical model (6) with intensity of treatment (see Tables 13 and 14 in Appendix A).

Table 8: Short-Term Impact of Flex Labor on Productivity

Metric	**(1)** Process A Productivity	**(2)** Process B Productivity	**(3)** Total Site Productivity
Impact	**-11.34**	-2.89	**-2.78**
Std. Error	4.78	10.63	1.65
Impact %	**-4.46%**	-0.59%	**-5.28%**
Impact % 90% CI LB	-7.56%	-4.15%	-10.46%
Impact % 90% CI UB	-1.36%	2.98%	-0.11%
Data Frequency	Daily	Daily	Daily
Observations	16828	16828	16828

There can be different types of negative impact of flex labor on productivity in process A: longer transition of new flex associates through learning curves, tasks allocation to new flex employees, suboptimal flex schedules, etc. To the best of our knowledge, there were no changes in task allocations or scheduling during the initial phase of our experiment. At the same time, as new flex associates transition through their learning curves, we expect the negative impact on productivity to decrease over time. Hence, to indirectly test the primary cause of the negative impact on productivity, we

leverage a synthetic control difference-in-difference methodology to construct a "synthetic" twin of all treatment stations during the baseline period. This allows us to estimate daily difference between the treatment stations and their synthetic twin in absence of increased part-time labor. Figure 13 shows that the negative causal impact of flex labor on productivity between treatment and control stations (*e.g.,* difference between treatment stations and their synthetic twin) is largest in magnitude during the first few weeks of the experiment. The magnitude of the impact decreases over time and becomes minimal by the end of the experiment (late October). These findings show that the negative effects of part-time labor on productivity are primarily driven by on-boarding new hires and their slower transition through the learning curve (due to lower number of weekly worked hours compared to non-flex employees). This negative effect can likely be mitigated by efficient training programs of flex associates. In particular, the businesses can consider training programs where fully tenured associates help teach new hires, and more gradual and staggered hiring of new flex associates, which allows them to get enough proficiency ahead of time. However, these strategies come with the corresponding costs of coordination in scheduling by pairing highly tenured associates, potentially decreasing their productivity, and staggered hiring of flex labor ahead of time can increase the cost of flex labor introduction as it may lead to initial surplus labor.

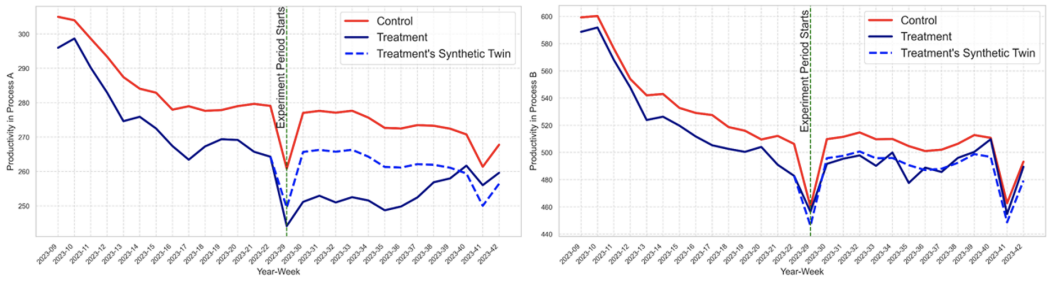

Figure 13: Daily Difference Between Productivity in Processes A and B and their Synthetic Twin in Absence of Intervention

5.4 Impact on Efficiency

We do not have access to Amazon UTR labor cost at the granularity which is needed for a rigorous causal measurement. In order to proxy the impact on efficiency, we focus our analysis on estimating the impact on non-flex (full-time) hours per package and flex (part-time) hours per package in the two major process paths. Table 9 reports the estimated impact. In process A, we find a significant decrease in non-flex (full-time) hours per package and a significant increase in flex (part-time) hours per package. However, the magnitude of flex hours change is almost twice the magnitude of non-flex hours (-0.19 full-time hours vs 0.37 part-time hours per 1,000 packages). Even accounting for 4.5% negative effect on productivity from the new flex hires ramping up period, this result indicates that full-time hours did not decrease adequately enough in process A. This is probably due to operators hedging on the uncertainty of flex labor performance in the complex process by intentionally over-staffing. In addition, this may reflect the incremental time managers and experienced non-flex associates had to spend in these processes to onboard the new flex associates. We performed extensive robustness checks and found that the estimated short-term effects on productivity are robust to using daily data and using specifications with the percentage of flex labor as the main explanatory variable (Table 15 and 16, respectively).

We again leverage a synthetic control difference-in-difference methodology to construct a "synthetic" twin of all treatment stations during the baseline period and compare non-flex hours per package in process A between treatment stations and their twin during the experiment. Figure 14 shows that the negative causal impact of flex labor on non-flex per package in process A increases by the end of the experiment (see 16 for process B). This signifies that as part-time associates transition through their learning curves, the operators may feel more comfortable to schedule fewer and fewer full-time hours per each package they need to process. Hence, in a steady state we may expect closer to a 1 to 1 substitution between non-flex and flex labor.

In process B, we find a 1 to 1 substitution non-flex and flex labor. Non-flex hours per 1,000 packages decreased by 0.17 hours whereas flex hours per 1,000 packages increased by 0.17 hours. These results

Table 9: Impact of Flex Labor on Efficiency. Difference in Difference with Treatment Dummy for Pilot Stations

Metric	(1) Process A	(2) Process A	(3) Process B	(4) Process B	(5) Total Site	(6) Total Site
	Non-Flex Hours Per 1,000 Packages	Flex Hours Per 1,000 Packages	Non-Flex Hours Per 1,000 Packages	Flex Hours Per 1,000 Packages	VET Hours Per 1,000 Packages	VTO Hours Per 1,000 Packages
Impact	**-0.19**	**0.37**	**-0.17**	**0.17**	**-0.34**	**0.78**
Std. Error	0.08	0.04	0.05	0.01	0.19	0.35
Impact %	**-5.29%**	**91.13%**	**-9.01%**	**84.67%**	**-22.01%**	**24.84%**
Impact % 90% CI LB	-8.90%	76.07%	-13.19%	72.53%	-42.42%	6.23%
Impact % 90% CI UB	-1.68%	106.19%	-4.82%	96.82%	-1.59%	43.45%
Data Frequency	Daily	Daily	Daily	Daily	Daily	Daily
Observations	16828	16828	16828	16828	16828	16828

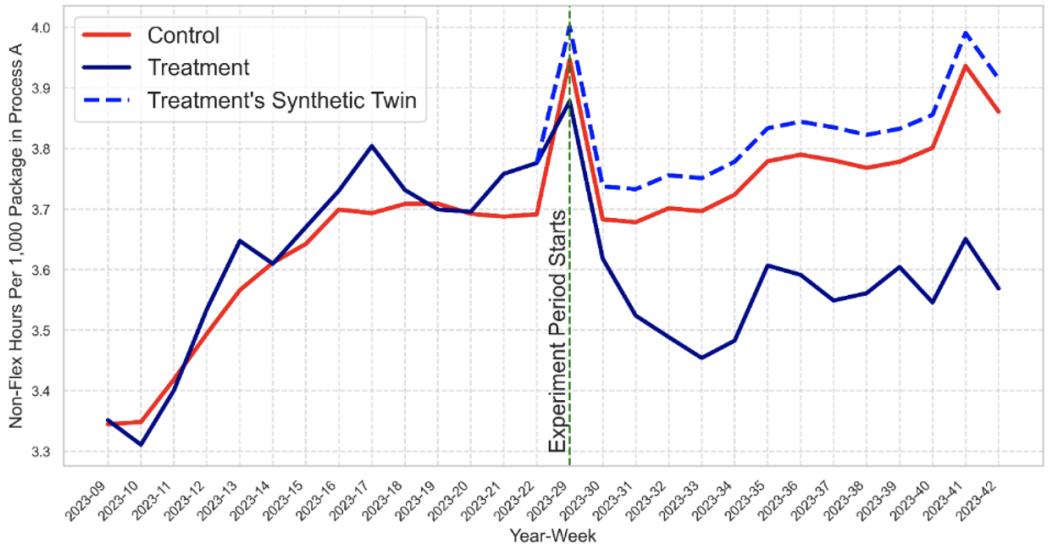

Figure 14: Daily Difference Between Non-Flex Hours Per Package in process A and its Synthetic Twin in Absence of Intervention

are congruent with relatively small decrease in productivity in this process (-0.6%). This finding shows that it is much easier to onboard part-time labor in a relatively simple processes versus a more complex one. We also observe inconclusive results for full-time flexibility levers. In particular, VET for non-flex associates decreased 22%, reflecting more active use of flex hours instead. However, VTO for non-flex associates increased by almost 25%. This probably illustrates the effect of not decreasing the non-flex associates' headcount during the experiment.

Overall, we conclude that significant increase in flex labor did not lead to material efficiency improvements in the short term. We see potential evidence for efficiency improvements in process B due to one to one substitution between hours and efficiency deterioration in process A due to inadequate decrease in full-time hours and a more pronounced negative initial impact of part-time labor on productivity. However, we do not expect this effect to last in the long-term as we see indicative signals that non-flex hours per package will further decrease and productivity of flex labor will further improve as operations get used to part-time labor and the new associates become more tenured. In this case, we can expect significant impact on efficiency longer-term due to structural staffing improvements and ability to accommodate demand shocks with higher flexibility of the part-time labor.

5.5 Long-Term Impact of Part-Time Labor

Most of the treatment stations in our experiment gradually decreased the percentage of flex labor in early 2024 after the 2023 holiday season. However, as a second wave of the experiment, twelve stations in one region (outside of the original experiment) dialed up part-time labor during the 2023

holiday season and continued to operate at above 10% of flex hours through the first half of 2024. This enables us to estimate the long-term impact of flex labor. In particular, we use these twelve Amazon delivery stations from this region as treatment and eight Amazon delivery stations from the same region as control stations to evaluate the causal impact of higher percentage of flex labor on operations over the long-term. As shown in Figure 15, the flex percentage stayed at near 0% levels for control stations in the experiment period, while at treatment stations flex percentage stayed high even after the holiday peak. Compared to the baseline, on average the percentage of flex hours increased by 10% overall at the station level, and by 13% in process A and 12% in process B.

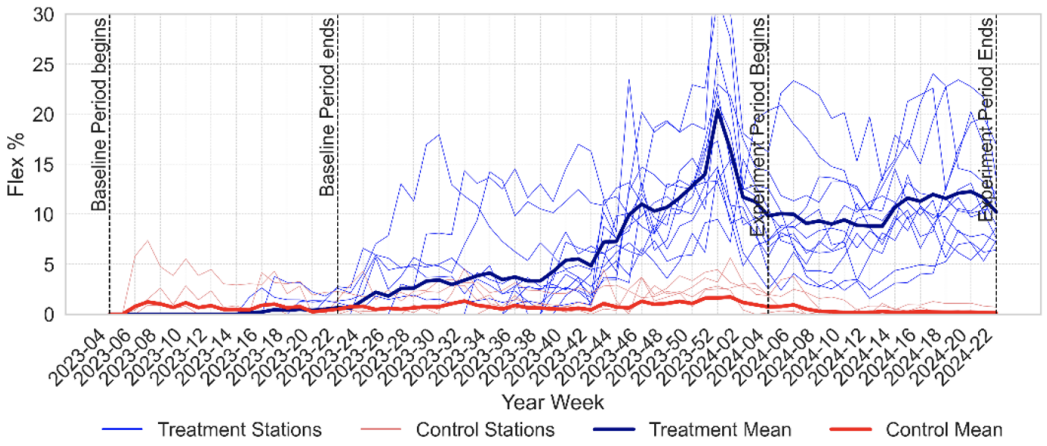

Figure 15: Flex Percentage at Treatment and Control sites Long-Term Analysis

Table 10 shows the long-term impact on process and site-level productivity. We find that a higher flex percentage leads to smaller and statistically insignificant negative impact (-3.91%) on the productivity in process A compared to the short-term impact (-4.46%). At overall station level, the productivity gap decreased from -5.28% in the short-term to an insignificant -0.98% in the long-term. Note that these reductions in productivity gap are despite the fact that flex percentage increase are significantly higher in the long-term than they were in the short-term. The improvements in productivity gap indicate that, over time, as flex associates transition through the learning curve their productivity improves and differences in productivity between flex and non-flex associates become negligible.

Table 10: Long-Term Impact of Flex Percentage Increase on Productivity

Metric	(1) Process A Productivity	(2) Process B Productivity	(3) Total Site Productivity
Impact	-11.88	24.45	-0.58
Std. Error	8.90	32.37	1.04
Impact %	-3.91%	4.75%	-0.98%
Impact % 90% CI LB	-8.74%	-5.63%	-3.85%
Impact % 90% CI UB	0.92%	15.13%	1.90%
Data Frequency	Daily	Daily	Daily
Observations	5320	5320	5320

Table 11 shows the substitution pattern between non-flex and flex hours across the two processes. Results suggest that it is difficult to replace full-time headcounts with part-time in process A where we observe that one flex hour replaces only 0.62 non-flex hours (-0.28 non-flex hours per 1,000 package vs 0.45 flex hours per 1,000 package). Nevertheless, in this process, non-flex vs flex substitution in the long term is still better than the short term where only 0.51 non-flex hours were replaced by one flex hour. When compared to process B where the substitution is 1.16 (-0.28 non-flex hours per 1,000 package vs 0.24 flex hours per 1,000 package), we see that the simpler process continues to provide a better substitution which results in more significant efficiency improvements. These findings are also

in line with the productivity estimates that show that flex hardly affects productivity of process B and thus flex is able to substitute non-flex hours 1 to 1 or more.

Table 11: Long-Term Impact of Flex Percentage Increase on Worked Hours

Metric	(1)	(2)	(3)	(4)	(5)	(6)
	Process A		Process B		Total Site	
	Non-Flex Hours Per 1,000 Packages	Flex Hours Per 1,000 Packages	Non-Flex Hours Per 1,000 Packages	Flex Hours Per 1,000 Packages	VET Hours Per 1,000 Packages	VTO Hours Per 1,000 Packages
Impact	**-0.28**	**0.45**	**-0.28**	**0.24**	-0.53	**-0.78**
Std. Error	0.10	0.06	0.09	0.03	0.33	0.38
Impact %	**-9.83%**	**100.59%**	**-16.46%**	**99.79%**	-33.12%	**-39.02%**
Impact % 90% CI LB	-15.78%	78.09%	-24.94%	82.48%	-67.10%	-70.58%
Impact % 90% CI UB	-3.88%	123.10%	-7.98%	117.09%	0.87%	-7.45%
Data Frequency	Daily	Daily	Daily	Daily	Daily	Daily
Observations	5320	5320	5320	5320	5320	5320

Finally, we see flex associates ultimately help reduce the need for non-flex levers to manage demand uncertainty. An increase in flex percentage decreases VET hours by 33% suggesting that stations can rely on part-time labor during high demand periods. The results also show improved structural efficiency as VTO hours reduce by 39%, which is a significant improvement from short-term where VTO hours increased (by 25%) when sites had to shed excess of non-flex hours. The decrease in VET and VTO hours suggest increased staffing efficiency and higher labor flexibility to accommodate demand shocks. At station level, with closer to 1 to 1 substitution between non-flex and flex hours and reduction in the VET and VTO hours, we see evidence for overall structural efficiency improvements that are likely to be congruent with our estimates of up to 3.9% efficiency improvements predicted in Section 4 by the two-stage optimization framework for going from 1% flex hours to 10% flex hours.

6 Conclusion and Future Research

This study presents a comprehensive framework for optimizing workforce composition at Amazon UTR operations, balancing flex and non-flex labor to meet on-demand capacity needs. Through the application of SEM and CM, coupled with empirical validation, we demonstrate that increasing flex labor from 1% to 10% of the workforce can lead to improved operational efficiency up to 3.9%. These improvements reflect the two main advantages of flex labor - the improvements in structural efficiency due to shorter and more flexible nature of flex shifts, and the ability to accommodate temporary demand shocks with higher flexibility. The insights obtained from this research offer a template for navigating the complexities of modern Supply Chain workforce management. Our findings suggest that flex associates can effectively integrate into UTR operations, particularly for less complex tasks, leading to improved structural efficiency without compromising service quality or reliability. This optimization framework and empirical study helped inform Amazon delivery stations strategy on workforce composition.

Importantly, our current framework assumes that delivery stations operate independently, with each station managing its own workforce composition. While Amazon has begun exploring flex workforce sharing across multiple delivery stations, this currently represents only a small fraction of total working hours. Therefore, our model recommendations remain valid under current operational conditions. As the practice of sharing associates across multiple stations becomes more prevalent, our framework will require adaptation to account for multi-station workforce allocation and the associated complexities of coordinating shared labor resources.

Another promising future research direction involves addressing short-term operational planning challenges in dynamic workforce environments. Short-term decisions may benefit from explicitly modeling individual associate tenure progression and week-over-week workforce evolution. A multi-period model that tracks associate experience levels could optimize hiring and scheduling based on expected learning curve progression. This dynamic framework would be particularly valuable in high-turnover environments where capturing tenure transitions becomes important for operational planning.

Future research can explore the impact of emerging trends on our workforce composition framework. This includes investigating how technological transformation such as automation would alter the optimal flex and non-flex balance and redefine worker roles in UTR environments. Studies on changing labor markets, evolving associate preferences, and potential regulatory shifts are crucial

to consider in order to ensure the model's long-term viability. Additionally, longer-term studies on the organizational impact of increased flex staffing, examination of how different workforce compositions affect sustainability goals could provide important insights. The fast development of large language models and artificial intelligence assistants has the potential to accelerate the learning curve for flex associates. This could help bridging the productivity gaps currently observed in complex processes. As a result, the utilization of flex labor for UTR operations may become increasingly viable, potentially driving higher efficiency for businesses.

References

D. S. Altner, A. C. Rojas, and L. D. Servi. A two-stage stochastic program for multi-shift, multi-analyst, workforce optimization with multiple on-call options. *Journal of Scheduling*, 21:517–531, 2018.

J. April, M. Better, F. Glover, J. P. Kelly, and G. Kochenberger. Strategic workforce optimization: Ensuring workforce readiness with optforce tm. *Ann. Optim.*, 2014.

Y. Barzel. The Determination of Daily Hours and Wages. *Quarterly Journal of Economics, 87(2), 220-238*, 1994.

A. Behrendt, M. Savelsbergh, and H. Wang. A prescriptive machine learning method for courier scheduling on crowdsourced delivery platforms. *Transportation Science*, 57(4):889–907, 2023.

N. Boysen, R. De Koster, and F. Weidinger. Warehousing in the e-commerce era: A survey. *European Journal of Operational Research*, 277(2):396–411, 2019.

C. Brewster, A. Hegewisch, and L. Mayne. Flexible Working Practices: The Controversy and the Evidence. *In Brewster Chris, Hegewisch Ariane, (Eds), Policy and Practice in European Human Resource Management: The Price Waterhouse Cranfield Survey. Routledge, London.*, 1994.

F. Devicienti, E. Grinza, and D. Vannoni. The Impact of Part-Time Work on Firm Total Factor Productivity: Evidence from Italy. *Industrial and Corporate Change, Vol. 7(2)*, 2018.

J. B. Fuller. Why a blended workforce may be key to lasting competitive advantage, November 2020. Retrieved from https://www.library.hbs.edu/working-knowledge/the-blended-workforce-imperative. Accessed in March 2025.

R. Ganesan, S. Jajodia, and H. Cam. Optimal Scheduling of Cybersecurity Analysts for Minimizing Risk. *ACM Transactions on Intelligent Systems and Technology (TIST), 8(4):52, 1 - 32*, 2017.

A. Garnero, S. Kampelmann, , and F. Rycx. Part-time Work, Wages and Productivity: Evidence from Belgian Matched Panel Data. *Industrial and Labor Relations Review, vol. 67 (3), pp. 926-954*, 2025.

H. Kim, H. Kim, and Y. Zhu. The Selection Effects of Part-Time Work: Experimental Evidence from a Large-Scale Recruitment Drive. *Review of Economics and Statistics, 1-46*, 2025.

K. Kim and S. Mehrotra. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research*, 63(6):1431–1451, 2015.

M. Mahato, N. Kumar, and L. K. Jena. Re-thinking gig economy in conventional workforce post-covid-19: A blended approach for upholding fair balance. *Journal of Work-Applied Management*, 13(2):261–276, 2021.

A. Manning and B. Petrongolo. The Part-Time Pay Penalty for Women in Britain. *The Economic Journal, 118(526)*, 2008.

A. Rijal, M. Bijvank, A. Goel, and R. De Koster. Workforce scheduling with order-picking assignments in distribution facilities. *Transportation Science*, 55(3):725–746, 2021.

G. Specchia and V. Vandenberghe. Is Part-time Employment a Boon or Bane for Firm Productivity? *Unpublished paper, Université Catholique de Louvain*, 2013.

M. W. Ulmer and M. Savelsbergh. Workforce scheduling in the era of crowdsourced delivery. *Transportation Science*, 54(4):1113–1133, 2020.

Appendix A

Table 12: Learning Curves and Productivity in Amazon Delivery Stations by Tenure in Process

Process	Metric	Associate Type	Hours Tenure In Process								Average
			<50	50-100	100-150	150-200	200-250	250-300	300-350	>350	
Process A	Productivity	Non-Flex	157.2	174.4	177.1	180.3	183.7	186.1	187.4	194.1	185.7
	% Of All Non-Flex Associates		9%	7%	7%	7%	7%	7%	6%	51%	
	Productivity	Flex	145.8	172.3	180.3	186.5	187.3	185.4	185.3	187.8	173.5
	% Of All Flex Associates		24%	17%	15%	11%	8%	6%	4%	14%	
Process B	Productivity	Non-Flex	326.2	356.9	367.1	371.1	373.2	373.3	374.5	372.6	365.5
	% Of All Non-Flex Associates		11%	9%	10%	9%	8%	7%	7%	38%	
	Productivity	Flex	322.6	352.3	359.7	360.1	362.9	358.3	360.4	364.3	346.5
	% Of All Flex Associates		32%	23%	14%	9%	6%	4%	3%	9%	

Table 13: Short-Term Impact of Flex Labor on Productivity Using Difference-In-Difference and Panel Fixed Effects Regression Based on Daily Data

Impact of Flex On Productivity During The Initial Experiment (July - October 2023)				
Dependent Variable		**(1)** Process A Productivity	**(2)** Process B Productivity	**(3)** Total Site Productivity
Panel A: Regression With % Of Flex As Treatment				
% Of Flex Hours in Total Hours	Impact	**-47.83**	-3.02	**-14.29**
	Std. Error	16.40	54.04	9.03
	Mean(Outcome)	254.22	491.57	52.68
	Impact %	-1.76%	-0.05%	-1.75%
	Impact % 90% CI LB	-2.76%	-1.60%	-3.57%
	Impact % 90% CI UB	-0.77%	1.49%	0.07%
Panel B: Difference in Difference With Treatment Dummy For Pilot Station				
Treatment Dummy	Impact	**-11.34**	-2.89	**-2.78**
	Std. Error	4.78	10.63	1.65
	Mean(Outcome)	254.22	491.57	52.68
	Impact %	-4.46%	-0.59%	-5.28%
	Impact % 90% CI LB	-7.56%	-4.15%	-10.46%
	Impact % 90% CI UB	-1.36%	2.98%	-0.11%
Data Frequency		Daily	Daily	Daily
Observations		16828	16828	16828

Table 14: Short-Term Impact of Flex Labor on Productivity Using Difference-In-Difference and Panel Fixed Effects Regression Based on Weekly Data

Impact of Flex On Productivity During The Initial Experiment (July - October 2023)				
Dependent Variable		**(1)** Process A Productivity	**(2)** Process B Productivity	**(3)** Total Site Productivity
Panel A: Regression With % Of Flex As Treatment				
% Of Flex Hours in Total Hours	Impact	**-68.20**	15.63	-16.50
	Std. Error	28.09	90.53	17.39
	Mean(Outcome)	253.45	488.35	52.53
	Impact %	-2.53%	0.27%	-2.04%
	Impact % 90% CI LB	-4.25%	-2.34%	-5.59%
	Impact % 90% CI UB	-0.81%	2.89%	1.51%
Panel B: Difference in Difference With Treatment Dummy For Pilot Station				
Treatment Dummy	Impact	**-11.44**	-3.15	-2.64
	Std. Error	4.92	10.76	1.77
	Mean(Outcome)	253.45	488.35	52.53
	Impact %	-4.51%	-0.64%	-5.02%
	Impact % 90% CI LB	-7.72%	-4.28%	-10.59%
	Impact % 90% CI UB	-1.31%	2.99%	0.55%
Data Frequency		Weekly	Weekly	Weekly
Observations		2408	2408	2408

Table 15: Short-Term Impact of Flex Labor on Efficiency Using Difference-In-Difference and Panel Fixed Effects Regression Based on Daily Data

Impact of Flex On Efficiency During The Initial Experiment (July - October 2023)							
		Process A		Process B		Total Site	
Dependent Variable		**(1)** Non-Flex Hours Per 1,000 Packages	**(2)** Flex Hours Per 1,000 Packages	**(3)** Non-Flex Hours Per 1,000 Packages	**(4)** Flex Hours Per 1,000 Packages	**(5)** VET Hours Per 1,000 Packages	**(6)** VTO Hours Per 1,000 Packages
Panel A: Regression With % Of Flex As Treatment							
% Of Ready Hours In Total Hours	Impact	**-3.30**	**3.68**	**-2.01**	**1.73**	-1.07	3.48
	Std. Error	0.21	0.06	0.21	0.13	1.64	3.02
	Mean(Outcome)	3.58	0.40	1.84	0.20	1.56	3.12
	Impact %	-8.65%	85.43%	-9.31%	75.39%	-4.43%	7.18%
	Impact % 90% CI LB	-9.53%	82.95%	-10.89%	66.19%	-15.59%	-3.11%
	Impact % 90% CI UB	-7.76%	87.91%	-7.73%	84.60%	6.74%	17.48%
Panel B: Difference in Difference With Treatment Dummy For Pilot Station							
Treatment Dummy	Impact	**-0.19**	**0.37**	**-0.17**	**0.17**	**-0.34**	**0.78**
	Std. Error	0.08	0.04	0.05	0.01	0.19	0.35
	Mean(Outcome)	3.58	0.40	1.84	0.20	1.56	3.12
	Impact %	-5.29%	91.13%	-9.01%	84.67%	-22.01%	24.84%
	Impact % 90% CI LB	-8.90%	76.07%	-13.19%	72.53%	-42.42%	6.23%
	Impact % 90% CI UB	-1.68%	106.19%	-4.82%	96.82%	-1.59%	43.45%
Data Frequency		Daily	Daily	Daily	Daily	Daily	Daily
Observations		16828	16828	16828	16828	16828	16828

Table 16: Short-Term Impact of Flex Labor on Efficiency Using Difference-In-Difference and Panel Fixed Effects Regression Based on Weekly Data

Impact of Flex On Efficiency During The Initial Experiment (July - October 2023)							
		Process A		Process B		Total Site	
Dependent Variable		**(1)** Non-Flex Hours Per 1,000 Packages	**(2)** Flex Hours Per 1,000 Packages	**(3)** Non-Flex Hours Per 1,000 Packages	**(4)** Flex Hours Per 1,000 Packages	**(5)** VET Hours Per 1,000 Packages	**(6)** VTO Hours Per 1,000 Packages
Panel A: Regression With % Of Flex As Treatment							
% Of Ready Hours In Total Hours	Impact	**-3.15**	**3.72**	**-2.14**	**1.91**	-3.62	6.20
	Std. Error	0.33	0.07	0.31	0.04	2.32	4.36
	Mean(Outcome)	3.58	0.40	1.85	0.20	1.60	3.05
	Impact %	-8.27%	86.67%	-9.93%	81.16%	-14.63%	13.17%
	Impact % 90% CI LB	-9.72%	84.03%	-12.33%	78.07%	-30.15%	-2.13%
	Impact % 90% CI UB	-6.83%	89.30%	-7.54%	84.24%	0.88%	28.48%
Panel B: Difference in Difference With Treatment Dummy For Pilot Station							
Treatment Dummy	Impact	**-0.19**	**0.37**	**-0.16**	**0.17**	**-0.36**	**0.77**
	Std. Error	0.08	0.04	0.05	0.02	0.20	0.35
	Mean(Outcome)	3.58	0.40	1.85	0.20	1.60	3.05
	Impact %	-5.17%	91.15%	-8.91%	85.31%	-22.41%	25.28%
	Impact % 90% CI LB	-8.84%	75.88%	-13.09%	71.77%	-43.39%	6.38%
	Impact % 90% CI UB	-1.51%	106.42%	-4.73%	98.84%	-1.42%	44.18%
Data Frequency		Weekly	Weekly	Weekly	Weekly	Weekly	Weekly
Observations		2408	2408	2408	2408	2408	2408

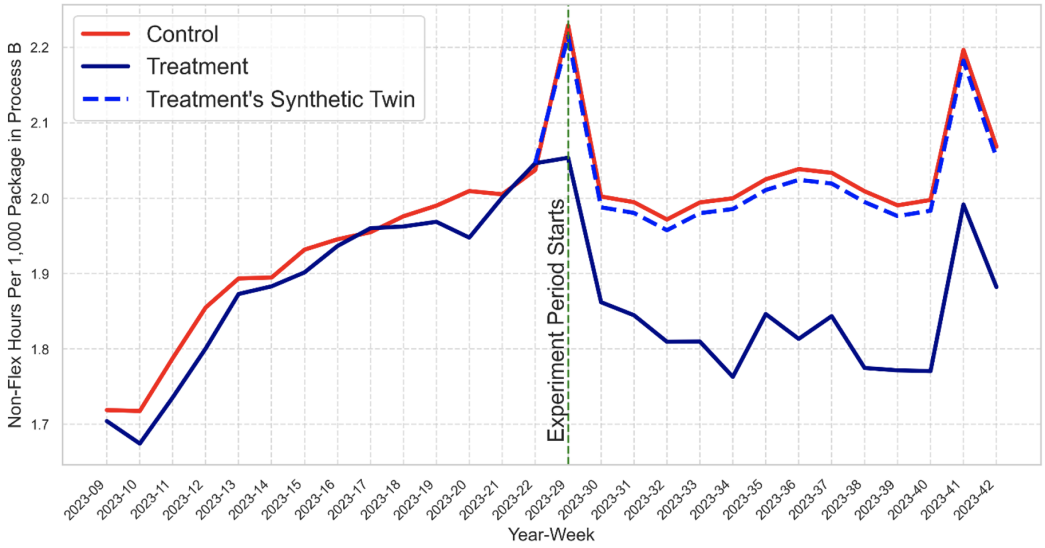

Figure 16: Daily Difference Between Non-Flex Hours Per Package in Process B and its Synthetic Twin in Absence of Intervention

Appendix B

Both SEM and CM are formulated as two-stage stochastic programs. First-stage decisions are made before scenario realization and remain constant across all scenarios (scenario-independent variables), while second-stage decisions are made after observing the scenario and vary accordingly (scenario-dependent variables indexed by ω). In SEM, x_p (non-flex cohort assignments) are first-stage decisions, while y_s^ω (flex shift assignments), o_t^ω (overstaffed hours), and u_t^ω (understaffed hours) are second-stage decisions. In CM, x^f and x^r (hiring decisions) are first-stage decisions, while all hour-related variables (h_ω^f, h_ω^r, h_ω^{VET}, h_ω^{VTO}) are second-stage decisions.

Table 17: SEM Notation Summary

Type	Symbol	Description
Sets	Ω_1	Set of weekly demand scenarios (hourly granularity)
	$\mathcal{T}_\delta$	Set of δ-hour time periods comprising a week
	$\mathcal{F}$	Set of weekly recurring non-flex associate cohorts
	$\mathcal{F}_t$	Set of weekly recurring non-flex associate cohorts that overlap with $[t, t+\Delta), t \in \mathcal{T}_\Delta$
	$\mathcal{R}$	Set of flex shift structures
	$\mathcal{R}_t$	Set of flex shift structures that overlap with $[t, t+\Delta), t \in \mathcal{T}_\Delta$
Parameters	Δ	Time window length for labor need (hours)
	$T_{t,q}$	Time overlap between non-flex cohort or flex shift structure $q \in \mathcal{F} \cup \mathcal{R}$ and time window $[t, t+\Delta), t \in \mathcal{T}_\Delta$
	d_t^ω	Labor needed hours in time window $[t, t+\Delta), t \in \mathcal{T}_\Delta$, scenario $\omega \in \Omega_1$
	$\overline{W}$	Average weekly total labor need (hours)
	h_p	Hours per week in non-flex cohort $p \in \mathcal{F}$
	f_t	Minimum fraction of demand covered by non-flex in window $t \in \mathcal{T}_\Delta$
	τ	Relative importance of understaffing vs. overstaffing
	π_f	Non-flex associate productivity
	π_r	Flex associate productivity
	ρ_f	Non-flex associate reliability rate
	ρ_r	Flex associate reliability rate
	$HC_{\max}$	Maximum total headcount in facility
Decision variables	x_p	Number of non-flex associates in cohort $p \in \mathcal{F}$
	y_s^ω	Number of flex associates in shift $s \in \mathcal{R}$, scenario $\omega \in \Omega_1$
	o_t^ω	Overstaffed hours in window $t \in T_\Delta$, scenario $\omega \in \Omega_1$
	u_t^ω	Understaffed hours in window $t \in T_\Delta$, scenario $\omega \in \Omega_1$

Table 18: CM Notation Summary

Type	Symbol	Description
Sets	Ω_2	Set of weekly demand scenarios (weekly granularity)
Parameters	π_f	Non-flex associate productivity
	π_r	Flex associate productivity
	ρ_f	Non-flex associate reliability rate
	ρ_r	Flex associate reliability rate
	d_ω	Weekly demand (units) in scenario $\omega \in \Omega_2$
	$\overline{D}$	Average weekly demand (units)
	H_f	Non-flex weekly hours
	$H_r^{\min}, H_r^{\max}$	Min/max flex weekly hours
	$H_{\text{VET}}^{\max}$	Max VET hours per non-flex associate per week
	$H_{\text{VTO}}^{\max}$	Max VTO hours per non-flex associate per week
	C^{wage}	Hourly wage rate
	C^{VET}	VET hourly cost
	C^{VTO}	VTO hourly cost
	C^{hire}	Per-associate hiring cost
	γ	Weekly attrition rate
	$\theta^*_{\min}, \theta^*_{\max}$	Min/max non-flex coverage percentages from SEM
Decision variables	x^f	Number of non-flex associates to hire
	x^r	Number of flex associates to hire
	h_ω^f	Planned non-flex regular hours in scenario $\omega \in \Omega_2$
	h_ω^r	Planned flex hours in scenario $\omega \in \Omega_2$
	h_ω^{VET}	Planned VET hours in scenario $\omega \in \Omega_2$
	h_ω^{VTO}	Planned VTO hours in scenario $\omega \in \Omega_2$