We study capacity management in a service system in which workers determine their own schedules. Our study is motivated by recent innovations in service delivery such as work-from-home call centers and ride-sharing services. Providers in these settings promise their customers a given level of service but do not directly assign agents to time intervals. Agents choose when to work based on the compensation offered and their individual availability thresholds in the relevant time interval. A provider must recruit enough agents and set the period-by-period compensation so as to generate enough participation to meet the desired service level.

We examine how a service provider should manage self-scheduling agents in two settings. The first is inspired by work-from-home call centers and assumes a service firm must staff to achieve contracted service level targets. The second is a newsvendor setting in which the service level is endogenous. In both settings, the firm prefers having a large pool of agents. However, this creates a tension between the firm’s profit and agents’ earnings. If the firm must guarantee agents a minimal level of compensation, the firm recruits a smaller pool of agents and limits the number of agents that can work in some time periods.

Finally, we show that allowing self scheduling leads to non-conventional cost structures and can be costly. Specifically, the average cost to serve a customer may increase with the volume of transactions. In other words, the service system does not exhibit the economies of scale one would typically expect. Further, we show that when the service level is endogenous, self scheduling leads to less staffing and hence worse service than a conventional benchmark.

Key words: strategic servers, independent capacity, distributed systems, service operations.

1. Introduction
Staffing in service environments is a challenging problem as managers seek to control costs while assuring adequate capacity to serve demand that varies over time. In tackling this problem, managers have always maintained an important trump card: the ability to tell workers when to work.
The overall construction of the schedule might involve worker preference, union rules, or government regulations but at the end of the process, the manager could inform each worker when she was expected to begin and end her shift. Further, there are generally implicit (and often explicit) consequences for not adhering to an assigned schedule.

Many novel service providers, however, are surrendering this power. Instead of ordering agents to clock in and punch out at appointed times, they are setting up systems in which agents create their own schedules and choose for themselves when to work. We are not speaking here of professional knowledge workers who are given flexible schedules as long as projects are completed on time. Rather, we are focusing on services such as work-from-home call centers (e.g., LiveOps and Arise Virtual Solutions) and ride-sharing services (e.g., Uber and Lyft) which must have capacity available to service demand as it arises.

These service providers have thus put themselves in a tenuous position. On the one hand, they have promised their customers good service. In the case of call centers, their relationships with clients (generally major corporations) are governed by contracts that target specific service levels. Ride-sharing services are not bound by the same contractual obligation, but they compete against conventional taxis and public transportation in part by emphasizing their availability. In the words on Uber’s chief executive, “Uber is ALWAYS a reliable ride.” (Kalanick 2012). Delivering on these commitments requires capacity; without adequate staffing, these service providers will fail to honor their obligations.

On the other hand, these service providers promise their agents flexibility. For example, the LiveOps’ agent recruitment web site promises the following:

“As a LiveOps independent agent, you can benefit from a highly flexible and rewarding opportunity. ... As an independent contractor providing services to LiveOps’ clients, you are your own boss!”

Flexibility and control of one’s schedule are important to agents. Work-from-home call centers, for example, often recruit people whose obligations outside of work (e.g., child or elder care) make taking a conventional job with little scheduling flexibility difficult. Consequently, the firm cannot simply renege on allowing agents to self schedule. It must instead offer compensation that induces the right number of agents to make themselves available at the right time.

The service provider must also assure that its agents have adequate earnings over time. Many of these firms recruit agents on-line as they have no facility for potential agents to drop in and

---

1 Describing the people serving customers for these firms requires some finesse. Generally, those answering calls or driving customers are not employees of the firm. Rather, they are independent contractors whose continued relationship with the service provider is dependent on achieving a minimal level of performance (e.g., an Uber driver rating) over time. We will generally refer to those serving customers as agents.

visit. They are consequently very concerned when web sites ask whether a particular firm is a “work-from-home scam”\(^3\) or when former agents complain in public forums that a firm is “the worst company ever” offering “below average” pay\(^4\).

The service provider’s problem can thus be understood as managing agent participation on two different time scales. On a short-term (hourly or less) basis, it must ensure that it attracts enough – but not too many – agents for each time interval over some horizon to maximize its profit while achieving a desired service level. On a longer-term basis, it must also ensure that agents earn enough that continuing with the firm is an attractive opportunity.

The goal of this paper is to examine how a firm that allows its agents to self schedule solves this problem. We consider a firm that must staff a service system facing time varying arrivals over a horizon. The firm recruits a pool of agents who are allowed in each period to choose whether or not to work. A given agent’s willingness to work varies with each period as she draws an availability threshold at the start of each period. Thus given the terms the firm offers, an agent may want to work this morning but then be unavailable this afternoon.

The firm has three control levers at its disposal. First, it can set the pool size – that is, how many agents it recruits and qualifies to serve customers. Since training agents takes time, the pool size is set at the start of the horizon and cannot be adjusted based on the demand in a given period. The second lever is the compensation offered to agents who work in a period. This can vary from time period to time period. For most of our analysis, we assume the firm offers a fixed compensation for each time interval (e.g., $15 per hour). However, we demonstrate that the firm can achieve identical results if, instead, it uses compensation schemes such as a piece rate that depend on the number of customers an agent serves. Finally, we allow the firm to cap the number of agents that are active in a period. That is, we allow the firm to tell an agent she cannot work in a given time interval even though she is willing to do so.

We consider two basic settings. The first is motivated by call centers. The service provider has contractually committed to deliver a given level of service, which translates to having a minimum staffing level. The required staffing level is increasing in customer volume so higher demand periods require more staff. The second is motivated by ride-sharing services and assumes that the firm is free to determine the appropriate service level. Consequently, it may set any staffing level it wants. Here, we employ a newsvendor setting and find that the optimal decision is an intuitive variant of the classical critical fractile solution.

We show that allowing self scheduling is costly to the firm and results in service systems with unconventional cost structures. In our newsvendor setting, the firm chooses a lower service level

\(^3\) See workathomemoms.about.com/od/callcenterdataentry/a/arise.htm accessed on Aug 28, 2014.

and hence a lower staffing level than it would if it could hire as many agents as it wanted at the prevailing wage and order them to work. The firm thus earns a lower profit – and customers experience worse service – because of self-scheduling agents.

In our call-center setting, the firm does not enjoy the typical economies of scale one expects in service systems. In many service models, it is cheaper to deliver a given service level (on a per customer basis) as demand volumes goes up. That does not happen when the firm allows self-scheduling. The firm must offer higher pay to attract more workers in high volume periods which (subject to some regularity conditions) results in the cost per customer served increasing.

We also demonstrate the firm needs to use all three control levers – particularly capping the number of active agents - when it must satisfy a nontrivial constraint on agent earnings. Absent an earnings constraint (i.e., when the firm only needs to consider gaining adequate agent participation in each period), the firm has incentive to make its pool of agents as large as possible. It is then able to offer relatively low wages in both high and low volume periods and still attract enough agents. Once there is a constraint on agent earnings, however, the firm cannot slash wages. This both increases costs and creates structural problems. In particular, low-volume periods will be overstaffed. Capping the number of active agents eliminates this problem. The firm limits the number of agents active in low-volume periods so it can guarantee agent earnings without inducing overstaffing and excessive labor costs. Further, we show that once the firm caps the number of workers, a firm allowing self scheduling offers the same service level as a conventional firm (in the newsvendor setting) and enjoys economies of scale (in the call-center setting).

Our work is related to the literature on the principal-agent model (see Salanie 1997 and Laffont and Martimort 2009 for reviews). Classical principal-agent models focus on hiring an agent to exert effort for the benefit of the principal when the agent’s actual effort cannot be observed. The principal must consequently be concerned with both directing the agent’s action as well as gaining the agent’s participation. Our model does not consider explicit effort in serving customers. In effect, we assume monitoring is sufficient to assure that agents provide the appropriate level of effort. Consequently, our attention is squarely on assuring agent participation.

There has also been some work in the operations literature looking at two-sided markets that match tasks with service providers (e.g., see Allon et al. 2012 and Moreno and Terwiesch 2014). In these papers, individual clients arrive looking to buy a specific service (e.g., coding a smart phone app) that can be carried out by one individual. The question then is how different rules or information structures affect market performance. In our case, the service provider commits to serving the stream of work while meeting a specified service level. The question is then not how one job gets matched with one service provider, but how the firm can assure it has sufficient capacity to meet its commitment.
2. A General Model

In this section, we present a general model highlighting the control variables available to the firm and the incentives of agents. We assume the firm faces a horizon composed of $T$ time intervals. In period $t$ (for $1 \leq t \leq T$), the firm’s revenue is determined by the number of available agents and by the market conditions. Let $A_t$ denote the number of agents available in period $t$ and $A = (A_1, \ldots, A_T)$. We assume that each agent can, on average, serve one customer per period so the firm’s staffing level is equivalent to its capacity. Let $\lambda_t$ denote the market conditions in period $t$ and $\Lambda = (\lambda_1, \ldots, \lambda_T)$. It is useful to think of $\lambda_t$ as a parameter to a probability distribution or stochastic process. For example, in a call center setting $\lambda_t$ could represent the rate of a Poisson process. $\Lambda$ is commonly known by the firm and the agents. This will be the case in markets with obvious day of the week or time of day seasonality (e.g., call volume is higher on weekday mornings than weekend afternoons or demand for ride sharing services is highest at rush hour).

Let $R(A_t, \lambda_t)$ denote the firm’s revenue in a period with $A_t$ available agents and market conditions $\lambda_t$. Its revenue over the horizon is $R_T(A, \Lambda) = \sum_{t=1}^{T} R(A_t, \lambda_t)$. We assume that the firm pays agents $\eta_t$ for being available in period $t$. For now we assume that compensation is implemented through a per-interval compensation (e.g., paying $20 per hour). We discuss other schemes such as a piece rate later. Letting $\eta'A = \sum_{i=1}^{T} \eta_i A_i$, we have that the firm’s profit over the horizon is

$$\Pi_T(A, \Lambda) = R_T(A, \Lambda) - \eta'A.$$

While the firm might like to have $A_t$ agents available to serve customers in period $t$, it cannot directly order $A_t$ agents to work. Instead it must offer sufficient compensation to induce that many agents to choose to work. We suppose that the firm has a pool of $N$ qualified agents. Interpret $N$ as the number of agents that are affiliated with (or belong to) the network of a firm, who have been trained to serve customers. Thus $N$ is the maximum number of agents that could potentially work in a given period. However, it is not the case that all pool members will work; some may find the firm’s offered compensation insufficient.

We model variation in agents’ availability to work by assuming that each agent has an availability threshold for each period. An agent may thus be available for work this morning because they have drawn a low threshold but be unavailable this afternoon or tomorrow morning because they have drawn a significantly higher threshold. More formally, each agent draws an availability threshold $\tau$ from a distribution $F$ at the start of each period. Agents are assumed to be statistically identical and independent of each other. The distribution does not vary over time, and a given agent’s draw for period $t$ is independent of her draw for any other period. We assume that $F$ is continuous with a strictly positive density $f$ on a support $(0, \Phi)$. Let $\bar{F}(\tau) = 1 - F(\tau)$. We assume that $F$...
is log-concave, a condition that holds for many common distributions (see Bergstrom and Bagnoli 2005).

Agents are risk neutral and seek to maximize their earnings subject to earning more than their availability threshold. Thus an agent with realized availability threshold $\tau$ in period $t$ makes herself available to work if the firm offers compensation $\eta_t$ greater than $\tau$. The total number of agents interested in working in period $t$ is then

$$I(\eta_t) = NF(\eta_t).$$

Note that we are implicitly appealing to the law of large numbers by assuming that the pool of qualified agents is sufficiently large that working the with average number of available agents is a reasonable approximation of the actual number of available agents.

The firm's problem is then to maximize $\Pi_T(A, \Lambda)$ by manipulating its available control levers. We consider three. The first is the pool size $N$. Since training agents takes some time, this decision must be made upfront. The pool size is thus constant over the horizon. The second variable is agent compensation which is allowed to vary from period to period. Finally, the firm may impose a cap $K_t$ on the number agents allowed to work in period $t$. If, under the offered compensation, the number of interested agents, $NF(\eta_t)$, is more than the number the firm needs, it can choose to limit access and allow entry only to the number it needs. With a cap $K_t$, the staffing level is $A_t = NF(\eta_t) \wedge K_t$.

To this basic problem we can add a variety of constraints. First, the firm's revenue may be contingent on providing a targeted service level. This is typically the case in the call center industry in which contracts routinely contain service level agreements. For example, a service provider may be obliged to answer a specified fraction of calls in less than a given number of seconds. We represent the measure of service level by a parameter $\alpha$. We denote by $C_\alpha(\lambda)$ the minimal staffing required to meet the service level when the arrival rate is $\lambda$. We will be more detailed about $C_\alpha$ when needed. We interpret $C_\alpha(\lambda) \equiv 0$ as leaving the choice of service level to the firm (which we will term an endogenous service level).

A second possible set of constraints deals with worker welfare. As we noted above, firms have an interest in assuring that they are seen as providing good opportunities for workers. We model this by imposing a constraint on the agents’ compensation. Such a constraint can take one of two forms. A per-period earnings constraint requires $\eta_t \geq \beta$ for all $i = 1, \ldots, T$, i.e., that the compensation offered on each interval exceeds a certain value. In a setting with a long repetitive horizon (as in virtual call centers where consecutive weeks are similar), this is equivalent to requiring that agents
get sufficient earnings on any “type” of interval on which they work. Agents that, say, can work only afternoons are not punished. If $\beta \equiv 0$ the firm faces no earnings constraints.

Alternatively, we may have an average earnings constraint, which requires that agents make more than a certain amount $\bar{\beta}$ in an average period. If the firm does not impose an access cap, the likelihood of an agent to work in period $t$ is $NF(\eta_t)$. Over a horizon of length $T$, the agent’s average per-period earnings are $\frac{1}{T} \sum_{t=1}^{T} \eta_t F(\eta_t)$. An average earnings constraint is less restrictive than a per-period one since the firm may effectively underpay in some periods as long as it offers sufficiently generous pay in others.

If the firm imposes an earnings cap, the calculation of average per-period earnings becomes more complicated and requires some assumption regarding how the firm chooses from among interested agents. We will assume random rationing: the agents who work in an interval are chosen randomly from amongst those that are interested.

Given these considerations, we can write the general form of the firm’s optimization problem as

$$\max_{\eta, N, K} \Pi_T(A, \Lambda)$$

s.t. $A_t \geq C(\lambda_t)$, $t = 1, \ldots, T$, \hspace{1cm} (Service level)

$A_t = NF(\eta_t) \land K_t$, $t = 1, \ldots, T$, \hspace{1cm} (Participation)

$\eta_t \geq \beta$, $t = 1, \ldots, T$. \hspace{1cm} (Earnings)

We will use the term uncapped formulation when referring to the formulation above where $K$ is dropped as a decision variable or, equivalently, set to $\infty$.

### 3. An Exogenously Specified Service Level: A Call-Center Model

What follows is a specialization of our general model to a call-center setting. We assume the call center has signed a contract to process a client’s calls over a horizon of length $T$. The horizon is split into $T_h$ high-demand periods (i.e., with market conditions $\lambda_h$) and $T_l = T - T_h$ low-demand periods (i.e., with market conditions $\lambda_l < \lambda_h$). We will interpret market conditions as the expected arrival rate in a period, and assume that customers neither balk or renege. $\lambda_i$ is then also the expected number of served calls.

The firm charges the client a price $p_i$ per call served on an interval of type $i = l, h$. The contract specifies a service level $\alpha$ that has to be met on all periods independent of market of conditions. Higher values of $\alpha$ correspond to more demanding service level requirements.

---

5 This is arguably important in virtual call center settings where different agents reside in different time zones.

6 Other rationing mechanisms are possible; for example, Netessine and Yakubovich (2012) discuss several settings in which better workers are given priority.
Let $C_\alpha(\lambda_i)$ denote the minimum capacity necessary to meet service level $\alpha$ given market conditions $\lambda_i$. The function $C_\alpha(\lambda)$ is assumed to be differentiable in $\lambda$. We impose several natural assumptions on $C_\alpha(\lambda)$. First, we assume that $C_\alpha(\lambda) \geq \lambda$ and that it increases with $\alpha$ – the higher the service level, the more capacity is required. Further, $C_\alpha(\lambda)$ increases with $\lambda$ – the higher the demand level, the more capacity is required.

We also assume that $C_\alpha(\lambda)$ exhibits economies of scale: $(C_\alpha(\lambda) - \lambda)/\lambda$ is a strictly decreasing function of $\lambda$ such that $C_\alpha(\lambda)/\lambda$ approaches one as $\lambda$ increases. In other words, excess capacity is increasing at a slower rate than the arrival rate. This is the case, for example, for the stationary $M/M/n$ queue if the service-level is expressed through the average-speed of response. Note that economies of scale imply that resources are more heavily utilized in high-demand periods. Assuming a constant wage for agents, this would in turn imply that the average cost of serving a call falls as demand volume goes up. Such economies of scale are one of the key reasons firms prefer processing customer transactions through call centers as opposed to dispersed, in-person locations.

Since there are two types of periods, it suffices to consider two compensation levels $\eta_l$ and $\eta_h$. We say that pool size $N$ is service-level feasible for market conditions $\Lambda$ if there exists an $\eta_h$ such that $NF(\eta_h) \geq C_\alpha(\lambda_h)$. We can then write $R_T(A, \Lambda) = p_l\lambda_l T_l + p_h\lambda_h T_h$ and the firm’s optimization problem is given by

$$\max_{N, \eta, K} p_l\lambda_l T_l + p_h\lambda_h T_h - (\eta_l A_l T_l + \eta_h A_h T_h)$$

s.t. $A_i \geq C_\alpha(\lambda_i), \ i = l, h,$

$A_i = NF(\eta_i) \land K_i, \ i = l, h,$

$\eta_i \geq \beta, \ i = l, h.$

Let $N^*, \eta^* = (\eta^*_l, \eta^*_h)$, and $K^* = (K^*_l, K^*_h)$ denote the firm’s optimal decisions.

Beyond the optimal actions, we are interested in the firm’s resulting cost structure and in particular its cost per call served. Understanding the cost per call is necessary for the firm to properly set its prices, $p_l$ and $p_h$. More generally there is the question of whether the firm still enjoys economies of scale. Given arrival rate $\lambda$, per-interval compensation $\eta$, cap $K$, and pool size $N$, the cost per call is given by

$$c(\lambda, \eta, N, K) = \eta \frac{A}{\lambda} = \eta \frac{NF(\eta) \land K}{\lambda}.$$  

For uncapped formulations (i.e., when $K_i = \infty$), we will drop $K$ as an argument of $c$. Assuming that the firm always staffs to at least meet its service level obligations, $c(\lambda, \eta, N, K)$ is always (weakly) greater than $\eta \frac{C_\alpha(\lambda)}{\lambda}$. Since $C_\alpha(\lambda_i) \geq \lambda_i$, rewriting the firm’s objective as

$$\lambda_l T_l (p_l - c(\lambda_l, \eta, N, K)) + \lambda_h T_h (p_h - c(\lambda_h, \eta, N, K)),$$

shows that profitability requires that $\eta^*_i \leq p_i$. 
3.1. Analysis without an earnings constraint

In analyzing the firm’s problem, we first consider the setting $\beta = 0$: the firm maximizes profit subject to service-level targets without any regard for the agents’ earnings. In this case, for a given $N$, the optimal compensation $\eta^*$ is specified by

$$NF(\eta^*_i) = C_\alpha(\lambda_i), \ i = l, h.$$ 

The firm thus sets $\eta^*_i$ to attract the minimum number of agents needed to meet its service level obligations and there is no need to impose any access caps (i.e., $K_i = \infty$). It is straightforward to show that both $\eta^*_h$ and $\eta^*_l$ are decreasing in $N$. As the pool of agents increases, $C_\alpha(\lambda_i)$ represents a smaller and smaller fraction of $N$ and the firm needs to induce an ever decreasing share of agents to be active. The firm’s profit thus increases with the pool size and all else being equal it will choose the largest possible $N$.

Since the required capacity increases with $\lambda$, we immediately have $\eta^*_h \geq \eta^*_l$. This suggests that economies of scale are in jeopardy. The need to pay more in a high-volume period to attract a sufficient number of agents assures that total labor costs will be higher. However, that does not guarantee average costs will be higher because required capacity rises more slowly than demand, increasing agent utilization.

A simple example illustrates this tension. Suppose availability thresholds follow a power function distribution $F(x) = (x/\Phi)^k$, and let $C_\alpha(\lambda) = \lambda + \gamma \sqrt{\lambda}$ for $\gamma > 0$. Then, the required compensation, $\eta$, solves $N(\eta/\Phi)^k = C_\alpha(\lambda) = \lambda + \gamma \sqrt{\lambda}$ with an associated cost per call of

$$\eta \frac{C_\alpha(\lambda)}{\lambda} = \Phi \left(1 + \frac{\gamma}{\sqrt{\lambda}}\right) \left(\frac{\lambda + \gamma \sqrt{\lambda}}{N}N\right)^{1/k},$$

which is increasing in $\lambda$ if $\sqrt{\lambda} \geq \gamma (k-1)$. Whether the cost increases in $\lambda$ or not depends then on the parameter $k$. For $k < 1$, $F$ is concave with a mode at zero. Hence, when $k$ is less than one, there are many agents that have low availability thresholds and the compensation offered when market conditions are weak is very low. An increase in demand consequently leads to a high percentage change in the offered compensation. The increase is so high that the increase in agent utilization is not enough to offset the higher compensation. The situation is initially reversed for $k \geq 1$. Now the mode of distribution is at $\Phi$ and the agent pool is dominated by agents with high availability

---

7 In practice, there is no infinite supply of agents so the firm will choose a finite value. Further, we have ignored any costs associated with maintaining the pool. If there is any overhead associated with carrying an agent in the pool, this would be weighed against the ability to pay a lower wage.

8 For intuition, it useful to note that the cost per call can be written as $\eta/\rho(\lambda)$ where $\rho(\lambda) = \lambda/C_\alpha(\lambda)$ is the utilization of the agents. Cost per call served then falls with $\lambda$ if the resulting percentage increase in utilization is greater than the percentage increase in compensation.
thresholds. The initial compensation is high and subsequent percentage increases are small. Thus increases in utilization are at first sufficient to offset the rise in compensation and reduce the cost per call served. However, once staffing requirements are high, percentage increases in utilization are small and cost per call now increases.

As the following theorem shows, these results generalize. A concave threshold distribution assures that compensation increases in volume while for non-concave distributions, volume must be sufficiently high. In either setting, letting agents set their own schedule means the firm must forgo economies of scale.

**Theorem 1 (diseconomies of scale for $\beta = 0$)** Fix $\alpha$, $N$ and a range $0 < \underline{\lambda} < \bar{\lambda} < \infty$ and suppose that $N$ is service-level feasible for all $\lambda$ in this range. Let $c_\alpha (\lambda, N)$ be the cost per call served given the market condition and the pool size and assuming that the firm sets compensation optimally, i.e.,

\[ c_\alpha (\lambda, N) = c(\lambda, \eta^*, N). \]

If either $F$ is concave or $\bar{\lambda}$ is sufficiently large, it holds that

\[ c_\alpha (\lambda_h, N) \leq c_\alpha (\lambda_l, N), \]

for all $\underline{\lambda} \leq \lambda_l \leq \lambda_h \leq \bar{\lambda}$.

The following example further illustrates the theorem. It is an example we will re-visit when studying $\beta > 0$.

**Example 1 ($\beta = 0$)** Suppose that $F = U[0, 1]$: $NF(\eta) = N(\eta \wedge 1)$ for all $\eta \geq 0$. Let $\lambda_h = 100$ and $\lambda_l = 70$, $T_l = T_h = 10$ and suppose the service level target requires that the probability of delay not exceed 22%. We use the $M/M/n$ queue model to find $C_\alpha (\lambda_h) = 110$ and $C_\alpha (\lambda_l) = 79$. The firm’s revenue per call is $p_l = p_h = 2$. Next, solving for $NF(\eta_l) = C_\alpha (\lambda_l)$, we get

\[ \eta_h = 110/N \] and $\eta_l = 79/N$,

and

\[ c_\alpha (N, 100) = \eta_h C_\alpha (\lambda_h) / \lambda_h = 110^2 / (100N) \] and $c_\alpha (N, 70) = 79^2 / (70N)$,

for each $N$ such that $(N, \lambda)$ is service-level feasible.

The solution is depicted in Figure 1. The x-axis corresponds to the pool size $N$, the left vertical axis corresponds to profit and the right vertical axis to earnings and cost per call.

Evidently, for each value of $N$, the cost per call for high demand period is, in fact, higher. Also, observe that the firm optimally increases $N$ as much as it can. Even if forced to use 110 agents (the minimum required for the high demand period) the firm makes 1730 over the horizon.
Assuming $F$ is concave is obviously a restrictive assumption but it does capture, for example, gamma and Weibull distributions with shape parameters less than one. We will show next that the result changes once the firm faces a constraint on the agent earnings (i.e., $\beta > 0$).

### 3.2. Analysis with an earnings constraint

We next bring in the per-period earnings constraint, $\eta_i \geq \beta$, $i = l, h$. The optimal solution to the uncapped formulation has $\eta^*_l = \eta^*_h = \beta$ with $N^*$ chosen so that $N^*F(\beta) = C_\alpha(\lambda_l)$ (which guarantees in particular that $N^*F(\beta) \geq C_\alpha(\lambda_l)$). The cost per call in the two periods then satisfies

$$c_\alpha(\lambda, N^*) = \beta \frac{C_\alpha(\lambda)}{\lambda^*}$$

for $N^*$.

The above derivations prove the following – the average cost per call served now falls with the arrival rate, restoring economies of scale.

**Theorem 2** Consider the uncapped formulation and suppose that $\beta > 0$. Then, the cost per call served is higher in the low demand period.

**Example 2** ($\beta > 0$) We keep the data from Example 1 and solve the firm’s optimization problem with the earnings constraint $\eta_i \geq \beta = 0.85$, $i = l, h$. See Figure 2. The firm has a finite optimal pool

---

9 When the firm is indifferent between several values of $N$ that yield equal profits, we break the ties by choosing the smallest value of $N$ that maximizes profits. This reflects some costs of managing the worker pool that we do not explicitly model.
size $N^* = 130$ and a maximum profit of 1520. Note that the difference between the cost per call levels across low and high demands periods is significant. This poses a challenge in terms of setting the prices $p_l$ and $p_h$. Finally, at the optimal point $N^*$ and $\eta^*$ are such that the low demand period is overstaffed. There are 110 agents joining instead of the required $C_{\alpha}(\lambda_l) = 79$.

Overstaffing in low demand periods creates a role for capping access. The firm can limit the number of agents that are allowed to sign-up for a period. The number of interested agents $NF(\eta_i)$ may exceed the cap, but at most $K_i$ agents are allowed to sign-up. Among those that are interested, $K$ agents are chosen randomly.

**Example 3 ($\beta > 0$ + access cap)** Consider the data from Example 1. If the firm is allowed limit access at the required level, namely letting at most 79 agents sign up at low demand periods and 110 at the high demand periods, the outcome is as in Figure 3. The firm’s profit has increased significantly from 1520 to 1800 as a result of the cap. The optimal pool size is the same as without the use of a cap and, at the optimal decision, the cost per call is lower in high demand periods (i.e., the system exhibits economies of scale) although the absolute difference between low and high demand costs is small ($c_{\alpha}(\lambda_l, N^*) = 0.96$ and $c_{\alpha}(\lambda_h, N^*) = 0.94$). Finally, note that the firm sets the optimal cap at $K_l = C_{\alpha}(\lambda_l) = 79 < 110 = NF(\beta)$. 

Figure 2: A numerical example for the uncapped formulation with $F = U[0, 1]$ and $\beta > 0$
In principle, the firm can choose to set the cap at various values. It is obvious from the structure of the problem that, if $K_i > C_\alpha(\lambda_i)$, the cap can be further decreased while increasing the objective function value. The following characterizes the firm’s optimal solution.

**Theorem 3** (*scale economies recovered*) The following holds if $\beta > 0$:

(i) There exists an optimal solution with $N^* = C_\alpha(\lambda_h)/F(\beta)$, $K_i^* = C_\alpha(\lambda_i)$ and $K_h^* = \infty$. In fact, any optimal solution has $K_i^* = C_\alpha(\lambda_i)$.

(ii) Under any optimal solution the costs per call $c_\alpha(N^*, \lambda_h)$ and $c_\alpha(N^*, \lambda_l)$ satisfy $c_\alpha(N^*, \lambda_l) \geq c_\alpha(N^*, \lambda_h)$.

(iii) The difference $c_\alpha(N^*, \lambda_l) - c_\alpha(N^*, \lambda_h)$ shrinks as the volumes $\lambda_l$ and $\lambda_h$ increase.

The first part of the theorem states that limiting access is a necessity to maximize profits but it suffices to limit access only in low demand periods. The second part states that the intuition of economies of scale is recovered once access is restricted through an access cap. The last item of the theorem is useful for pricing purposes, i.e., for determining $p_l$ and $p_h$ – the cost per call served is relatively independent of the demand volume which suggest that little is lost if the firm charges a *static* per call price $p = p_l = p_h$.

The intuition behind this last result rests on the access cap’s ability to control *utilization*. When a per-period earnings constraint is introduced, the firm must for some periods offer compensation above the mere level needed to guarantee the service level (if the pool size is sufficiently high). If it
does not limit access, the effect of this increase on the staffing cost is magnified because additional agents join. The resulting drop in utilization increases the cost per call.

The cap negates the utilization effect. Additional agents beyond those needed to meet the service level obligation cannot join. The increase in total labor cost is driven solely by the need to offer adequate compensation and is not exacerbated by having a surplus of agents. When access is capped at $C_a(\lambda_i)$, the costs per call differences between intervals are purely determined by the effects of scale economies and thus shrink as the arrival rates grow.

We conclude this section with a study of the average earnings constraint: the period earning constraint $\eta_i \geq \beta$, $i = l, h$ is replaced with

$$\frac{1}{T} \left( \eta_l T_l \frac{A_l}{N} + \eta_h T_h \frac{A_h}{N} \right) \geq \bar{\beta}.$$

In contrast to Theorem 3, the following shows that limiting access is not a necessary tool when facing average earning constraints. The firm can regulate capacity by using only its power to set the pool size in the beginning of the horizon and vary compensation on a period-by-period basis.

**Theorem 4** Consider the firm’s problem with an average earnings constraint. Then, there exist an optimal solution with $K^*_l = K^*_h = \infty$.

**Example 4** *(Period vs. average earnings)* To compare the effect of period-earnings and average-earnings constraints on firm’s profit let $(N^*, \eta^*, K^*)$ be a solution to the period earnings constraint problem (if there are multiple solutions with different value of $N^*$ we take one with the smallest value of $N^*$). This solution generates an average earnings per agent of

$$\bar{\beta} = \eta^*_l n^*_l T_l + \eta^*_h n^*_h T_h \frac{T^*}{N^*}$$

where $n^*_i = NF(\eta^*_i) \land K^*_i$, $i = l, h$.

Consider next the firm’s problem with the average constraint with $\bar{\beta}$ using the data of Example 1. An optimal solution to the period-earnings formulation with $\beta = 0.85$ has

$$N^* = 130, \ \eta^*_l = \eta^*_h = 0.85, \ K^*_l = 79, \ K^*_h = \infty,$$

with an objective function value of 1793. The resulting total earnings during the horizon computed as above are 12.35. The equivalent value of $\bar{\beta}$ would then be $\bar{\beta} = 12.35/20 = 0.6175$. The optimal solution to the average constraints formulation is

$$N^* = 122, \ \eta^*_l = 0.65 \ \eta^*_h = 0.9, \ K^*_l = K^*_h = \infty,$$

with an objective function value of 1895. The firm utilizes its flexibility in setting the compensation above 0.85 in high demand periods and below 0.85 in low demand periods. This allows it also to decrease the pool size to 122 from 130.
4. Alternative compensation mechanisms

Thus far we assumed that the firm implements its optimal compensation $\eta^*$ by paying agents a fixed per-period amount: an agent that signs-up to work Monday 10:00-10:30 gets $x$ regardless of the number of calls served. Call centers like LiveOPS and Arise use piece-rate compensation or some combination of piece rate and a guaranteed per-interval minimum. Similarly, ride-sharing service such as Lyft and Uber compensate drivers by splitting fares with them. There are several reasons why a firm might prefer some sort of volume dependent compensation. For example, a piece rate may address moral hazard issues (that we have left unmodeled) and induce agents to exert more effort. Additionally, a piece rate is easier on a firm’s finances since it only pays agents when it has been paid by the client; such a consideration may be important for a nascent firm with limited resources.

Below we show that within our model with risk neutral agents many reasonable compensation mechanisms are equivalent. Namely, there exists a translation from one to the other that generates the same outcomes in terms of staffing levels and profits.

In all the compensation schemes we consider, the amount an agent earns in a period depends in part on the number of transactions she processes. Let $X^j_t$ denote the number of calls agent $j$ serves in period $t$. The distribution of $X^j_t$ depends on a number of factors including the number of active agents, the arrival process of calls, and how the firm allocates calls across agents. If some agents are given higher priority in the allocation procedure, they will earn more money than those with lower priority. Here we assume that calls are distributed uniformly among the agents that are working. Thus, if $x_t$ calls arrive during interval $t$ and $A_t$ agents are active, each will receive roughly $x_t/A_t$ calls. We will make the assumption that the arrival rate $\lambda_t$ is sufficiently large that $\lambda_t/A_t$ is a good approximation for the expected number of calls each agent serves in period $t$ (i.e., $E[X^j_t] = \lambda_t/A_t$ for all $j$). For a setting in which routing is non-uniform see Stouras et al. (2013).

4.1. Piece rate compensation

We first consider paying agents solely a piece rate $\phi$ per transaction. (While we allow $\phi$ to vary with the time period, here we will suppress the dependence on $t$ for notational simplicity.) If $A$ agents are active and the arrival rate of calls is $\lambda$, each agent’s expected earnings are

$$\mu = \frac{\phi \lambda}{A}.$$ 

The number of interested agents is then be $NF(\phi \lambda / A)$. Recall that the number of interested agents when the firm paid a fixed amount $\eta$ per interval was $NF(\eta)$. In comparing these values, note that under a fixed rate scheme, an agent can determine whether or not to work by considering only her availability threshold. However, under a piece rate scheme, an agent must consider both
her threshold and what other agents are doing. That is, we need to consider an equilibrium among
the agents so that the number of interested agents is equal to the number that join. Hence, we
require that
\[ NF\left( \frac{\phi \lambda}{A^e} \right) = A. \]

**Lemma 1** Fix \( N, \lambda \) and \( \phi \) such that \( NF(\phi) > \lambda \). There then exists an equilibrium \( A^e \in (0, NF(\phi)] \),
characterized by the unique solution to the equation
\[ NF\left( \frac{\phi \lambda}{A^e} \right) = A^e. \] (1)

With \( F = U[0, \Phi] \) for \( \Phi > 0 \) and provided that \( NF(\phi) = N\phi/\Phi > \lambda \) the equilibrium equation is
given by \( N\phi \lambda/A^e = A^e \), so that \( A^e = \sqrt{\lambda N\phi/\Phi} \) and agents earn \( \mu = \phi \lambda/A^e = \sqrt{\lambda \Phi \phi/N} \) per hour worked.

A priori, it is conceivable that this equilibrium structure introduces additional constraints into
the firm’s optimization problem or, in other words, that an optimal solution \( (N^*, \eta^*, K^*) \) to the
firm optimization problem is not implementable via a piece rate. It is a matter of a simple argument
to show that this is not the case, i.e., that the firm can move from per-interval compensation
to piece-rate compensation without compromising its profits or changing the pool size or staffing
levels.

Suppose the firm is using a feasible solution \((N, \eta, K)\) with \( K_i \geq NF(\eta_i), i = l, h \) (so that access
is not really limited). The firm should offer the piece rate \( \phi' = (\phi'_l, \phi'_h) \) such that
\[ NF\left( \frac{\phi'_i \lambda_i}{A_i} \right) = A_i, \]
where \( A_i = NF(\eta_i) \). Since \((N, \eta, K)\) is a feasible solution to the firm’s problem, it must be that
\( N \geq NF(\eta_i) = A_i \) so that the existence of \( \phi' \) follows from the continuity of \( F \). With this choice of
\( \phi' \), the number of agents that sign-up in equilibrium is (using Lemma 1) the unique solution to
\( NF(\phi'_i \lambda_i/x) = x \) which must equal \( A_i \) by construction.

If \((N^*, \eta^*, K^*)\) is an optimal solution to the firm’s optimization problem, then by Theorem 3
we have that \( A_l = C_\alpha(\lambda_l) \) and \( A_h = C_\alpha(\lambda_h) \) so that the firm can set the piece rate at \( \phi'_i = \eta'_i C_\alpha(\lambda_i)/\lambda_i \)
or equivalently \( \phi'_i \lambda_i/C_\alpha(\lambda_i) = \eta'_i \). With this translation, the optimal solution \((N^*, \eta^*, K^*)\) to the
firm’s problem with interval compensation is equivalent to the solution \((N^*, \phi^*, K^*)\) with piece rate compensation: (i) the number of agents interested for each interval is the same (and using the same cap, so is the number of people actually signing-up), (ii) the expected staffing costs are the same
and, hence, (iii) the expected profits are the same.
There are, however, some interesting differences between the piece rate and fixed compensation. First, the firm’s staffing costs are deterministic under fixed compensation but these costs are variable under a piece rate if there is any uncertainty in the demand volume. Thus a properly chosen piece rate delivers the same expected profit as the optimal fixed interval compensation but the realized profit for a given demand outcome differs.

Second, piece rate compensation lessens the impact of increasing the pool size. Under fixed rate compensation, doubling the pool size while holding the compensation rate constant will double the number of interested agents. The response to an increase in the pool size is less elastic when a piece rate is used. If the pool sized is doubled while the piece rate is unchanged, the number of interested agents increases but does not double. Competition between agents dissuades some agents with availability thresholds less than $\phi$ from making themselves available.

4.2. Two-part tariff

Having considered fixed interval and piece rate compensation, a natural extension is a two-part tariff that combines the two. Under a two-part tariff with parameters $\nu_t$ and $\phi_t$, agent $j$ earns $\nu_t + \phi_t X^j_t$ when she works in period $t$ and answers $X^j_t$ calls. $\phi_t$ is again a piece rate while $\nu_t$ is fixed and independent of the volume of calls answered.

As in the discussion above if $(N^*, \eta^*, K^*)$ is an optimal solution to the firm’s optimization problem, then any two-part tariff that has

$$\nu_t + \frac{\phi_t \lambda_i}{C_\alpha(\lambda)} = \eta^*_i, \ i = l, h$$

will generate the same outcomes and is a valid implementation for the firm’s optimal decision. Two-part tariff allows for more degrees of freedom. While the optimal piece rate is uniquely defined from the optimal interval compensation, here there are multiple options.

4.3. A piece rate with a minimum guarantee

A two-part tariff is not the only way to combine fixed per-interval payment with a piece rate. A call center we have worked with pays agents a piece rate but with a guaranteed minimum payment level. Uber has also been reported to guarantee an hourly rate at some time periods (Kirsner 2014).

We denote the parameters of the contract by $\kappa_t$ and $\phi_t$. $\phi_t$ is again piece rate while $\kappa_t$ is a fixed interval payment. Under this scheme, agent $j$ earns $\max \left\{ \kappa_t, \phi_t X^j_t \right\}$ when she works in period $t$ and answers $X^j_t$ calls. The agent then earns the guaranteed minimum when she answers fewer than $\kappa_t/\phi_t$ calls. Relative to a simple piece-rate contract or a two-part tariff, this contract reduces the variability in the agent’s earnings.

Fully analyzing this contract would necessitate calculating the distribution of $X^j_t$ for both high and low demand periods. We take a simpler route. Recall that an agent’s capacity is one call per
time period. We assume that an agent serves exactly one call or no calls in a given period. Given that calls are allocated randomly across agents, the probability that an agent serves a customer (when there are $A_t$ agents working) is $\lambda_t/A_t$. Thus, the agent’s expected compensation is given on such an interval by

$$\kappa_t \left(1 - \frac{\lambda_t}{A_t}\right) + \phi_t \frac{\lambda_t}{A_t}.$$ 

Following the procedure outlined above, one can determine a $(\kappa_i, \phi_i)$ for $i = l, h$ to replicate the firm’s optimal $(N^*, \eta^*, K^*)$. However, a guaranteed minimum payment level has an appealing property: It can replicate the outcome of the firm’s optimal fixed interval contract with a static contract. That is, one can set $\kappa_l = \kappa_h$ and $\phi_l = \phi_l$ while still having agents working under different market conditions earning different expected compensation levels.

To see this, let $(N^*, \eta^*, K^*)$ be an optimal solution to the firm’s optimization problem with static compensation. The optimal $\kappa^*$ and $\phi^*$ are then found from

$$\kappa^* \left(1 - \frac{\lambda_i}{C_\alpha(\lambda_i)}\right) + \phi^* \frac{\lambda_i}{C_\alpha(\lambda_i)} = \eta_i^*, \ i = l, h.$$ 

The system of equations has a unique solution if $\lambda_h/C_\alpha(\lambda_h) \neq \lambda_l/C_\alpha(\lambda_l)$, which holds given our assumptions about economies of scale in the service process.

While this compensation mechanism provides a valid implementation for the firm’s optimal solution, a caveat is in order. The solution $(\kappa^*, \phi^*)$ might have $\kappa^* < 0$, which can be interpreted as charging agents for working on an interval. Although some have argued that Uber would be better off charging drivers a flat fee and letting them keep 100% of their fares (The Economist 2014), imposing a sign-up fee may not be acceptable in many settings. The firm could still choose to implement a minimum guarantee contract but set a different contract for each market condition. If the contract terms vary with the demand rate, the results are similar to the case of a two-part tariff in that we would have a continuum of contracts that all implement the firm’s optimal decisions.

5. **Endogenizing The Service Level: A Newsvendor Model**

To this point we have assumed that the firm was contractually obliged to deliver a specific level of service and consequently commit to a specific staffing level. That obligation obscured whether allowing self scheduling actually costs the firm anything; there was no obvious benchmark against which to compare the outcome of self scheduling to determine whether allowing workers to choose when to work raised or lowered the firm’s profit. We now examine this question by allowing the firm to set agent compensation and the number of agents in a setting in which expected revenue increases with the staffing level. Specifically, we assume that the firm faces a newsvendor problem in each time period and demand unmet in a period is lost. This setting seems also more appropriate to model companies like Uber and Lyft whose customers are individuals (rather than firms).
For this analysis we specialize our general model \( (2) \) as follows. We consider a \( T \)-periods horizon where the number of customers in period \( t \) is a random variable \( \lambda_t \) drawn from a distribution \( G_t \). Demand draws are independent across periods. Serving a customer requires an agent and an agent can only serve customer per period. If the realized number of customer exceeds the available number of agents, excess demand is lost. Each served customer earns the firm \( p \) while firm gets nothing from unserved customers. The firm’s revenue in period \( t \), given a staffing level \( A_t \), is then
\[
R(A_t, \lambda_t) = pS(A_t, \lambda_t) = p \left( \int_{0}^{A_t} x g_t(x) \, dx + A_t \bar{G}_t(A_t) \right),
\]
where \( g_t \) is the (strictly positive) density of \( G_t \) and \( \bar{G}_t = 1 - G_t \).

We assume the firm pays agents active in period \( t \) a fixed compensation \( \eta_t \) whether or not they serve a customer. (We discuss how a piece rate can be implemented at the end of this section.) The firm’s profit over the horizon is
\[
\Pi_T(A, \Lambda) = \sum_{t=1}^{T} (R(A_t, \lambda_t) - \eta_t A_t),
\]
and the firm’s optimization problem is
\[
\max_{\eta, N, K} \Pi_T(A, \Lambda) \quad \text{s.t.} \quad A_t = NF(\eta_t) \wedge K_t, \ t = 1, \ldots, T, \\
\eta_t \geq \beta, \ t = 1, \ldots, T. \tag{2}
\]

To begin the analysis we assume that customer arrivals are identically distributed in each period and hence drop the dependence on the time period from the notation. We first optimize the compensation level assuming that the pool size is ample, that \( \beta = 0 \), and that no access caps are used. The unique optimal compensation level \( \eta^* \) must satisfy
\[
\bar{G} (NF(\eta^*)) = \frac{\eta^* + \frac{F(\eta^*)}{f(\eta^*)}}{p}, \tag{3}
\]
The uniqueness of \( \eta^* \) follows from assuming \( F \) is logconcave which implies that the reversed hazard rate \( \frac{f(\eta^*)}{F(\eta^*)} \) is monotonically decreasing.

In the mechanism design literature, \( \eta^* + \frac{F(\eta^*)}{f(\eta^*)} \) is known as the virtual cost. That is, the decision maker acts as if her marginal cost is \( \eta^* + \frac{F(\eta^*)}{f(\eta^*)} \) even though she pays agents only \( \eta^* \). This becomes clear if one considers a firm facing a similar newsvendor problem but that is able to hire however many agents it wants \( \eta^* \). The firm’s problem is then
\[
\max_{A} \ pS(A, \lambda) - \eta^* A.
\]
The firm’s optimal staffing level given $\eta^* A(\eta^*)$ is found from

$$\bar{G}(A(\eta^*)) = \frac{\eta^*}{p}. \quad (4)$$

We will refer to this as the benchmark newsvendor problem. Comparing (3) and (4) leads to the following proposition.

**Theorem 5** The participation constraints decreases the firm’s staffing level and profit, i.e.,

$$NF(\eta^*) \leq A(\eta^*) \quad \text{and} \quad pS(NF(\eta^*), \lambda) - \eta^* NF(\eta^*) \leq pS(A(\eta^*), \lambda) - \eta^* A(\eta^*), \quad \text{for all} \quad N \geq 0.$$

We thus have a clear statement that self scheduling is costly. The firm ends up with fewer agents (and thus less revenue) than it would want to have at the compensation rate $\eta^*$. Intuitively, the benchmark newsvendor analysis holds the cost of capacity constant so increasing the number of agents does not increase the cost of capacity that the firm already has. Under self scheduling, increasing the compensation rate to draw in more agents means paying more for the agents that were willing to work at a lower rate.

Self scheduling is also costly for consumers. If the firm could hire as many agents as it liked at $\eta^*$, it would provide a service level (in the sense of covering all demand) of $G(A(\eta^*)) = \frac{p-\eta^*}{p}$. Under self scheduling the comparable service level is

$$G(NF(\eta^*)) = \frac{p-\eta^* - \frac{F(\eta^*)}{f(\eta^*)}}{p}.$$

Consequently, a firm that uses self scheduling will be a less reliable service provider.

Of course, the difference between $NF(\eta^*)$ and $A(\eta^*)$ may be negligible so the impact on the firm’s profit and customer service level are trivial. Hence we now consider how the pool size, the retail price and the distribution of the agent’s availability threshold affect the firm’s action. To this end, we write $\eta^*_e$ to capture explicitly the dependence of the optimal agent compensation on the availability threshold distribution.

**Theorem 6** The firm’s profit and service level increases as either $N$ increases, $p$ increases, or the availability threshold distribution decreases in the reversed hazard rate order: i.e., $\Pi(NF_1(\eta^*_{F_1,N})) \geq \Pi(NF_2(\eta^*_{F_2,N}))$ given $F_1$ and $F_2$ if

$$\frac{f_1(\tau)}{F_1(\tau)} \leq \frac{f_2(\tau)}{F_2(\tau)}.$$

The compensation rate for agents falls as either $N$ increases or the availability threshold distribution decreases in the reversed hazard rate order. Agent compensation increases as $p$ increases.
As when the service level was exogenously fixed, the firm prefers a large agent pool. A large pool promises more agents with low availability thresholds, which allows for a lower payment to agents. Holding the pool size fixed, a smaller distribution of availability thresholds means that a greater fraction of the pool is available at any compensation rate, which increases the firm’s profit and the service level. As the retail price increases, agent compensation also increases. However, these gains are not so large that they result in lower firm profit or service level; customers who pay more may reasonably expect better service.

For the case of increasing the pool size, we can go a step further than simply saying that the service level increases. Since \( \frac{F(\eta)}{f(\eta)} \) is increasing in \( \eta \), we also have that the gap in the service level offered by a firm allowing self scheduling and one solving the benchmark newsvendor problem also falls. Consequently, while self scheduling is less profitable, a self-scheduling firm sacrifices little when it has a large pool of agents.

It is not possible to prove such general conclusions for shifts in the availability cost distribution or the retail price. Numerical experiments, however, suggest that a smaller threshold distribution or a higher retail price reduce the gap between self scheduling and benchmark staffing levels. Figure 4 provides an example, comparing self scheduling to the benchmark newsvendor problem as the retail price increases. We see that as \( p \) increases the gap in the service level – and hence staffing levels and staffing costs – falls but remain somewhat large (the ratio between the self scheduling service-level and that of the benchmark mode is at below 90% even when \( p \) is as large as 10). However, the gap in profit is much less significant and the profits are practically identical (approaching the 100% ratio as \( p \) reaches 10).

To this point, we have ignored any constraint on agent earnings. Recall that when the service level was exogenously specified, imposing the condition that \( \eta \geq \beta \) could result in the firm imposing an access cap in some periods while limiting its pool size. We now consider the impact of such a constraint when the firm controls the service level.

First, it is obvious that if \( \eta^* \) for a given \( N \) is greater than \( \beta \) (where \( \eta^* \) is determined by \( \beta \)), the earnings constraint is not binding and the firm can use its previous decision. If \( \eta^* < \beta \) the firm will have to set the compensation per interval at \( \beta \) and will get \( NF(\beta) \) servers interested. The following is a characterization of the firm’s optimal decision.

**Theorem 7** The optimal solution with earnings constraints has \( \eta^* = \beta \) and \( N^* = \frac{G^{-1}(\frac{p}{\beta})}{F(\beta)} \). For all value of \( N \neq N^* \) such that \( \eta^*_N < \beta \) the firm can strictly increase its profit by setting a cap and the optimal cap is set at \( A(\beta) \).

Given a large enough pool, an earnings constraint eliminates any difference between self scheduling and the benchmark problem. If \( N \geq N^* \), the self-scheduling firm is able to attract all the agents.
Figure 4: Sensitivity to retail price $p$: self scheduling relative to benchmark newsvendor. Demand is normally distributed (mean 100 and standard deviation 30) and availability costs following a gamma distribution (shape parameter of 2 and scale parameter of 3), and pool size of 400.

it wants at compensation $\beta$ – just as we assumed in the benchmark newsvendor problem. Hence, its staffing level and profit is the same as under the benchmark problem.

Note that when the firm is allowed to choose the optimal pool size $N^*$, capping the number of agents that work is not necessary. This may seem inconsistent with our results for the call center setting, but it is not. There too, if we take the arrival rate to be the same $\lambda_l = \lambda_h$ in every period, capping access is not needed under the optimal decision. The cap is important when we move to a time-varying environment.

Let us then assume, as in §3, that there are two types of intervals (low and high) with respective demand distributions $G_l$ and $G_h$ such that $G_h$ is stochastically greater than $G_l$. Let us assume that there are $T_l$ interval of low demand and $T_h$ of high demand. The firm thus faces problem (2) with

$$\Pi_T(A, \Lambda) := (\rho_l T_l S(A_l) + \rho_h T_h S(A_h)) - (\eta_l T_l A_l + \eta_h T_h A_h).$$

Let $\eta_{i,N}^*$ be the solution to (3) when the demand distribution is $G_i$, $i = l, h$, the pool size in $N$, and $K = \infty$. Let $A_i(\beta)$ be the solution to (4) with $\eta = \beta$ and the demand distribution is $G_i$, $i = l, h$.

**Theorem 8** (i) Fix $N$ and suppose $\beta = 0$. Then, the optimal compensation is lower on the type-l periods, i.e.,

$$\eta_{l,N}^* \leq \eta_{h,N}^*.$$
and, consequently, the staffing level is lower. The service level is, however, higher in low demand periods.

(ii) Suppose that $\beta > 0$. Then every optimal solution has $\eta^*_l,N = \eta^*_h,N = \beta$ and $N^* \geq A_h(\beta)/F(\beta) = \frac{G^{-1}_{h}\left(\frac{F(\beta)}{\beta}\right)}{\beta}$, and the assigned capacity satisfies $N^* F(\beta) \land K^*_l = A_l(\beta)$ and $N^* F(\beta) \land K^*_h = A_h(\beta)$. There exists an optimal solution with $N^* = \frac{G^{-1}_{h}\left(\frac{F(\beta)}{\beta}\right)}{\beta}$. In this solution the firm sets a cap in the low demand period $K^*_l = A_l(\beta)$, and $K^*_h = \infty$.

Our findings from §3 are thus generalizable to this setting: in the presence of earning constraints, the pool size is determined by the high demand period, the firm places a cap on the low demand period but not necessarily on the high demand period.

Before closing this section, we make two observations. First, we have done this analysis assuming that the firm pays a fixed per-period amount. Alternatively, the firm could employ a piece rate or one of the other contracts examined in §4. As in §4 it can be argued that the two schemes are equivalent: one can determine a piece rate that generates the same outcome as the optimal hourly wage and vice versa.

Second, our results generalize beyond a newsvendor setting. If the firm faces a general revenue function $R(A_t, \lambda_t)$ which is concave and increasing in $A_t$ given some market condition $\lambda_t$, a firm allowing self-scheduling will understaff and earn a lower profit at its optimal compensation rate $\eta^*_t$ relative to a firm that can hire any number of agents at $\eta^*_t$.

6. Concluding Remarks

We have examined a model in which a service provider allows its agents to choose when to work, a scheme used in many growing industries. The firm faces time varying demand over a horizon and must offer compensation that attracts enough workers to provide an adequate service level. We show that allowing self-scheduling is costly; when the firm is free to set any service level, it picks a lower service level than it would in a standard newsvendor setting. When the firm has a priori committed to a service level, self scheduling leads to the loss of the economies of scale one typically expects in a service system. Interestingly, both of these issues are mitigated when the firm imposes a cap on the number of agents active in low volume periods in order to efficiently satisfy a constraint on agents’ earnings.

Our model is an abstraction of reality and there are several ways in which it could be extended. For example, we have assumed that agents all draw their availability thresholds from the same distribution and this distribution does not vary over the horizon. In practice a service provider might recruit different types of agents whose availability varies in different ways. For example, stay-at-home parents are (stereotypically) up early but busy with family chores in the late afternoon. In contrast, college students (again stereotypically) find early mornings a challenge but may be
free in the afternoon. Given two such pools, the questions would be how many of each type the firm should recruit and how this would affect the compensation it needs to offer. Note that such an analysis would be significantly more complicated since the exact sequence of busy and slow periods would now matter.

An additional consideration would be competition between agents for service providers. Someone who is qualified to drive for Uber could also choose to drive for Lyft. This has led to competition between ride-sharing services to attract drivers [Kirsner 2014]. It is reasonable to expect such competition to drive up agent compensation, but the firm might have several ways in which to boost agent earnings – particularly when the firm employs a piece rate (as is the case in the ride-sharing industry). A firm could raise the piece rate outright or cap the number of active agents to increase the utilization of those working. In a monopoly setting, this is a straightforward analysis but with competition different firms might follow different strategies.

References


Appendix

A. Proofs for the call center model

The following lemma is used in the proof of Theorem 1. Given a random variable $X$ with distribution $F$, let $\bar{F}_F(x) = E[X|X \leq x]$, $S_F(x) = x - E[X|X \leq x]$ and $r_F(x) = f(x)/F(x)$.

Lemma A.1 If $f$ is log-concave then $r(x)S_F(x) \leq 1$ for all $x \geq 0$.

Proof. The log concavity of $f$ implies that $G(x) = \int_0^x F(z)\,dz$ is log-concave \cite{An1995}. Let $\tilde{G}(x) = \log |G(x)|$. We then have

\[
\tilde{G}'(x) = \frac{G'(x)}{G(x)} = \frac{F(x)}{\int_0^x F(z)\,dz} \quad \text{and} \quad \tilde{G}''(x) = \frac{f(x)}{\int_0^x F(z)\,dz} - \frac{F(x)^2}{(\int_0^x F(z)\,dz)^2}
\]

Log-concavity of $G(x)$ further implies that

\[
\frac{f(x)}{\int_0^x F(z)\,dz} \leq \frac{F(x)^2}{(\int_0^x F(z)\,dz)^2}
\]

\[
\frac{f(x)}{F(x)} \leq 1, \quad \text{and} \quad r_F(x) S_F(x) \leq 1.
\]

Proof of Theorem 1. We consider first the case that $F$ is concave. Let $G$ be the inverse of $F$ (and, in particular, also convex). The convexity and increasing monotonicity of $G$ guarantee that $G(x)/x$ is increasing and convex. Recall that $\eta$ is the minimal value such that $NF(\eta) = C_\alpha(\lambda)$. With this notation the cost per call is given by

\[
c(\lambda) = \eta \frac{C_\alpha(\lambda)}{\lambda} = G(C_\alpha(\lambda)/N) \frac{C_\alpha(\lambda)}{\lambda} = \frac{G(C_\alpha(\lambda)/N) N}{C_\alpha(\lambda)/N} \frac{N}{\lambda}.
\]

Letting $x(\lambda) = C_\alpha(\lambda)/N$, we have

\[
c(\lambda) = \frac{G(x(\lambda)) N}{x(\lambda)},
\]

Thus, we have expressed $c(\lambda)$ as the product of two increasing functions and, in particular, it is increasing in $\lambda$.

To prove that the result holds for a sufficiently large $a$, recall again that $c(\lambda) = \eta C_\alpha(\lambda)/\lambda$ and, in particular, satisfies the equation

\[
NF \left( \frac{c(\lambda) \lambda}{C_\alpha(\lambda)} \right) = C_\alpha(\lambda).
\]

(5)

Taking derivatives on both sides of this equation, we have

\[
NF(c(\lambda) \lambda/C(\lambda)) \left[ \frac{c'(\lambda) \lambda + c(\lambda) C_\alpha(\lambda) - C_\alpha'(\lambda) \lambda c(\lambda)}{C_\alpha^2(\lambda)} \right] = C'_\alpha(\lambda).
\]

(6)

Replacing $f = rF$ and using \cite{5} we have that

\[
NF \left( \frac{c(\lambda) \lambda}{C_\alpha(\lambda)} \right) = NF \left( \frac{c(\lambda) \lambda}{C_\alpha(\lambda)} \right) r_F \left( \frac{c(\lambda) \lambda}{C_\alpha(\lambda)} \right) = C_\alpha(\lambda) r_F \left( \frac{c(\lambda) \lambda}{C_\alpha(\lambda)} \right),
\]

(7)
so that
\[ c'(\lambda) = \frac{1}{\lambda r_F} \left( \frac{c(\lambda)}{c^\prime(\lambda)} \right) \left[ c^\prime(\lambda) - r \left( \frac{c(\lambda)}{C_a(\lambda)} \right) c(\lambda) \left[ 1 - \frac{C^\prime_a(\lambda)}{C_a(\lambda)} \right] \right]. \]

By Lemma A.1 we have that \( x r_F(x) = r_F(x) S(x) + r_F(x) \tau_F(x) \leq 1 + r_F(x) \tau_F(x) \). The log-concavity of \( r \) guarantees that \( r \) is a non-increasing function of \( x \) (see Remark 2 on Bergstrom and Bagnoli[2005] Page 466) replacing \( f \) with \( F \) there). Thus, fixing \( \epsilon > 0 \), there exists \( \eta \) such that \( \sup_{x \in (\eta, \infty)} r_F(x) \leq \epsilon \) and, in turn, \( r_F(x) \tau_F(x) \leq \epsilon \tau_F(\infty) = \epsilon E[\tau] =: \theta \). In turn, \( x r_F(x) \leq 1 + \theta E[\tau] \) for \( x \in (\eta, \infty) \). We next claim that the function \( x r_F(x) \) is continuous on \([0, \eta]\) (if one defined \( 0 = 0 \)). It is clearly continuous on \((0, \infty)\) and it is easy to verify that \( x r_F(x) \to 0 \). It thus has a maximum on the compact set \([0, \eta]\). Denote this maximum by \( \nu \). And we have that \( x r_F(x) \leq 1 + \theta + \nu =: \varpi \) for all \( x \). Thus,
\[ r_F \left( \frac{c(\lambda)}{C_a(\lambda)} \right) c(\lambda) = r_F \left( \frac{c(\lambda)}{C_a(\lambda)} \right) \frac{c(\lambda)}{C_a(\lambda)} \lambda \leq \varpi \frac{C_a(\lambda)}{\lambda}. \]

Note that by assumption, specifically since \((C_a(\lambda) - \lambda)/\lambda\) is decreasing we have that \( \lambda C^\prime_a(\lambda)/C_a(\lambda) \leq 1 \) and by the above that
\[ 0 \leq r \left( \frac{c(\lambda)}{C_a(\lambda)} \right) c(\lambda) \left[ 1 - \frac{C^\prime_a(\lambda)}{C_a(\lambda)} \right] \leq \varpi \frac{C_a(\lambda)}{\lambda} \left[ 1 - \frac{C^\prime_a(\lambda)}{C_a(\lambda)} \right] \]
by our assumptions that \( C_a(\lambda)/\lambda \to 1 \) and \( C^\prime_a(\lambda)/\lambda \to 1 \) as \( \lambda \to \infty \). We conclude that the right hand side above goes to 0 as \( \lambda \) increases and in turn
\[ c'(\lambda) \geq \frac{1}{2 \lambda r_F} \frac{1}{\frac{c(\lambda)}{C_a(\lambda)}} \geq 0, \]
for all sufficiently large \( \lambda \) and appropriately chosen \( N \) as required.

**Proof of Theorem 2.** This theorem is a direct consequence of the derivation preceding its statement.

**Proof of Theorem 3.** We argue first that *any* optimal solution \((\eta^*, N^*, K^*)\) has \( N^* \geq C_a(\lambda_b)/F(\beta) \). Suppose, towards contradiction, that \( N^* < C_a(\lambda_b)/F(\beta) \) then the firm must use \( \eta_b > \beta \) to guarantee enough capacity in the high demand periods, i.e., that \( N^* F(\eta_b) \geq C_a(\lambda_b) \). Since the firm can increase \( N^* \) without changing \( \eta^*_b \) or \( A^*_1 = K^*_1 \) \* \( N^* F(\eta^*_b) \) it will necessarily do so and use that to decrease \( \eta^*_b \) and its costs so its profits would go up contradicting the optimality of \( N^* \).

We conclude that \( N^* \geq C_a(\lambda_b)/F(\beta) \) and \( \eta_b^* = \beta \) in any optimal solution. Since \( N^* F(\beta) \geq C_a(\lambda_b) \geq C_a(\lambda_1) \) the firm has not reason to use \( \eta^*_b > \beta \). We conclude that under any optimal solution \( \eta^*_b = \eta^*_b = \beta \). Consequently, the objective function value, under any optimal solution, is of the form
\[ p_l \lambda_l T_l + p_h \lambda_h T_h - \beta ((N^* F(\beta)) \land K^*_1 + (N^* F(\beta)) \land K^*_1). \]

For the sake of the service level constraints it must be the case that \( K^*_1 \geq C_a(\lambda_1) \). Now, since \( N^* F(\beta) > C_a(\lambda_b) > C_a(\lambda_1) \) it is optimal, under any solution to set \( K^*_1 = C_a(\lambda_1) \). Finally, since the cost is monotone increasing in \( N^* \) for each value of \( K^*_1 \) we can choose \( N^* = C_a(\lambda_b)/F(\beta) \). In this case, we can set \( K^*_b = \infty \). Thus, we have shown that any optimal solution has \( N^* \geq F(\beta)/C_a(\lambda_b) \) and \( K^*_1 = C_a(\lambda_b) \) and that there exist optimal solutions with \( K^*_b = \infty \). This completes the proof of item (i).

Item (ii) of the theorem follows from the above. Since, under any optimal solution, \( \eta^*_b = \eta^*_b = \beta \), \( A_l = NF(\beta) \land K^*_1 = C_a(\lambda_1) \) \( A_b = N^* F(\beta) \land K^*_b = C_a(\lambda_b) \), the cost per call satisfies
Finally, note that by the assumed continuity of $F$, finite caps we can always construct a solution with the same overall profit but with infinite caps.

Item (iii) is now immediate noting that

$$c(N^*, \lambda_i) - c(N^*, \lambda_h) = \beta \left( \frac{C_a(\lambda_i)}{\lambda_i} - \frac{C_a(\lambda_h)}{\lambda_h} \right)$$

and the economies of scale guarantee that the right hand side converges to 0 by the assumption that $C_a \lambda / \lambda \to 1$ as $\lambda \to \infty$.

**Proof of Theorem 4.** Consider first an optimal solution that has no binding cap. Namely, for each $i = l, h$, if $K_i < \infty$ then $K_i > NF(\eta_i)$. Then, the cost of the firm does not change if we increase $K_i$ indefinitely without affecting the cost (or the constraints) and, in particular, the cap $K_i$ is not needed for optimality.

It thus suffices to consider an optimal solution that has $NF(\eta_i) < K_i$ for some $i$. Without loss of generality let us assume that $i = l$. Since $NF(\eta_l) > K_l$ we have in particular that $\eta_l NF(\eta_l) > \eta_l K_l$. Since the function $x NF(x)$ is continuously decreasing in $x$, we can find $\hat{\eta}_l < \eta_l$ such that $\hat{\eta}_l NF(\hat{\eta}_l) = \eta_l K_l$. Setting $K_i = \infty$ and replacing $\eta_l$ with $\hat{\eta}_l$ we have constructed a new solution with the same cost and the same average earnings since these are given by

$$\bar{\beta} \leq \frac{1}{T} \left( \eta_l \left( NF(\eta_l) \wedge K_l \right) T_l + \eta_h \left( NF(\eta_h) \wedge K_h \right) T_h \right) = \frac{1}{T} \left( \hat{\eta}_l NF(\hat{\eta}_l) T_l + \eta_h \left( NF(\eta_h) \wedge K_l \right) T_l \right).$$

Moreover, this solution is feasible for the service level constraint. Indeed, recall that we started from an optimal solution so that it must be the case that $K_i \geq C_a(\lambda_i)$. Suppose now that $NF(\hat{\eta}_l) < K_l$. Then, $\hat{\eta}_l NF(\hat{\eta}_l) < \hat{\eta}_l K_l < \eta_l K_l$, which is a contradiction to how we chose $\hat{\eta}_l$. We conclude that $NF(\hat{\eta}_l) \geq C_a(\lambda_i)$ and is, in particular, feasible for the service level constraint.

We conclude that given an optimal solution (for the problem with average-earning constraints) that has finite caps we can always construct a solution with the same overall profit but with infinite caps.

**Proof of Lemma 7.** Fix $\lambda, \phi, N > 0$ such that $NF(\phi) > \lambda$. Recall the equilibrium equation

$$NF \left( \frac{\phi \lambda}{A} \right) = A. \quad (7)$$

By the assumed continuity of $F$, the function $g(x) = NF(\phi \lambda / x)$ is continuous on $(0, \infty)$ and trivially decreasing in $x$. In turn, the function $g(x) - x$ is continuous on $(0, \infty)$ and strictly decreasing in this range with $g(x) - x \to N$ as $x \to 0$ and $g(x) - x \to -\infty$ as $x \to \infty$. A unique $x \in (0, \infty)$ such that $g(x) - x = 0$ must exist. Finally, note that $x = 0$ cannot be a solution to (7) since then $0 = NF(\infty) = N$.

**B. Proofs for the newsvendor model.**

**Proof of Theorem 5.** The right hand side of (8) is greater than that of (6) so that, since $\bar{G}$ is decreasing in its argument, we must have that $NF(\eta^*) \leq A(\eta^*)$. Next,

$$\Pi(NF(\eta^*)) - \Pi(M^*(\eta^*)) = p \int_{A(\eta^*)}^{NF(\eta^*)} x g(x) \, dx + p NF(\eta^*) \bar{G}(NF(\eta^*)) - \eta^* NF(\eta^*)$$
Note that
\[ p \int_{A(\eta^*)}^{NF(\eta^*)} xg(x) \, dx \leq pNF(\eta^*) (\bar{G}(A(\eta^*)) - \bar{G}(NF(\eta^*))) \]
Thus,
\[ \Pi(NF(\eta^*)) - \Pi(A(\eta^*)) \leq pNF(\eta^*) (\bar{G}(A(\eta^*)) - \bar{G}(NF(\eta^*))) + pNF(\eta^*) \bar{G}(NF(\eta^*)) - \eta^* NF(\eta^*) = 0 \]
\[ \square \]

**Proof of Theorem 6**: We first show the monotonicity results for the compensation followed by the service level and, finally, the profits.

**Compensation**: The fact that compensation increases with \( p \) and decreases with \( N \) is evident from the optimal fractile formula (3). For instance, suppose that as \( p \) increases the compensation actually decreases. Then, the left-hand-side of (3) decreases by the monotonicity of \( F/f \) (recall that log-concavity of \( f \)), so that the right-hand side \( \bar{G}(NF(\eta)) \) must also decrease which entails (since \( F \) is decreasing and \( \bar{G} \) is decreasing) that \( \eta \) increases which would be a contradiction.

To prove that the compensation increases with the agent cost distribution \( F \) note that, if \( F_1 \) is smaller than \( F_2 \) in the reverse hazard rate order, then it is also smaller in the regular stochastic ordering sense, \( \bar{F}_1(x) \leq \bar{F}_2(x) \) (or \( F_1(x) \geq F_2(x) \)), so that
\[ \bar{G} (NF_1 (\eta_2^*)) \leq \bar{G} (NF_2 (\eta_2^*)) \]
since \( \bar{G} (x) \) is decreasing. Further,
\[ \eta_2^* + \frac{F_2 (\eta_2^*)}{f_2 (\eta_2^*)} \leq \eta_2^* + \frac{F_1 (\eta_2^*)}{f_1 (\eta_2^*)}. \]

Recalling that
\[ \bar{G} (NF_1 (\eta_1^*)) = \frac{\eta_1^* + \frac{F_1 (\eta_1^*)}{f_1 (\eta_1^*)}}{p} \quad \text{and} \quad \bar{G} (NF_2 (\eta_2^*)) = \frac{\eta_2^* + \frac{F_2 (\eta_2^*)}{f_2 (\eta_2^*)}}{p} \]
Thus,
\[ G(NF_1(\eta_2^*)) \leq G(NF_2(\eta_2^*)) = \eta_2^* + \frac{F_2 (\eta_2^*)}{f_2 (\eta_2^*)} \leq \eta_2^* + \frac{F_1 (\eta_2^*)}{f_1 (\eta_2^*)} \]
Since the left hand side increases in the argument and the right hand side decreases, we conclude that \( \eta_1^* \leq \eta_2^* \) and, in particular, that facing agents with a smaller cost distribution, the firm offers a lower wage but uses more capacity and provides a higher service level.

**Service level**: The fact that the service level increases with \( p \) is evident from the optimal fractile formula (3) and from the fact, already proved, that the optimal compensation increases with \( p \). Similarly, the optimal service level increases with \( N \). For the effect of agent costs on service level, recall that compensation increases with agent costs so that, if \( F_1 \) is small than \( F_2 \) in the reverse hazard order cases, then by the monotonicity of \( F/f \), we have that
\[ \eta_1^* + \frac{F_1 (\eta_1^*)}{f_1 (\eta_1^*)} \leq \eta_2^* + \frac{F_2 (\eta_2^*)}{f_2 (\eta_2^*)} \]
which guarantees that the right-hand-side of (3), increases with agent cost and, in turn, the service level decreases (recall that, while (3) uses \( \bar{G} \), the service level is given by one \( 1 - \bar{G}(NF(\eta^*)) \)).
\textbf{Profits:} Recall that the profit of the firm is given by

\[ \Pi(NF(\eta^*)) = p \int_0^{NF(\eta^*)} xg(x) \, dx + NF(\eta^*) \left( \frac{\eta^* + \frac{F(\eta^*)}{f(\eta^*)}}{p} \right) - \eta^* NF(\eta^*) \]

\[ = p \int_0^{NF(\eta^*)} xg(x) \, dx + NF(\eta^*) \frac{F(\eta^*)}{f(\eta^*)}. \]

where we plugged in the first order condition \(3\) for \(\eta^*\). To show that the profits increase in \(N\), we first take derivative on both sides of \(3\) to get (to simplify notation let \(\eta = NF(\eta^*)\))

\[ g(NF(\eta(N))) (F(\eta(N)) + N f(\eta(N)) \eta' (N)) = \frac{1}{p} \left( \eta' (N) + \left( \frac{F}{f} \right) \eta' (N) \right) \]

Recall that the service level is weakly decreasing in \(N\) and the wages decrease in \(N\) and we conclude that \(\eta' (N) \leq 0\) and we conclude that

\[ F(\eta(N)) + N f(\eta(N)) \eta' (N) \geq 0 \quad \text{and} \quad \left( \eta' (N) + \left( \frac{F}{f} \right) \eta' (N) \right) \geq 0 \]

which we use next. Take a derivative of the profit with respect to \(N\) to get

\[ \frac{\partial \Pi(NF(\eta(N)))}{\partial N} = pNF(\eta(N)) g(NF(\eta(N))) (F(\eta(N)) + N f(\eta(N)) \eta' (N)) \]

\[ + F^2(\eta(N))/f(\eta(N)) + NF(\eta(N)) \left( \eta' (N) + \left( \frac{F}{f} \right) \eta' (N) \right). \]

Both lines of the right hand side are positive by \(8\) and we conclude that the profit increases in \(N\). The fact that it increase in \(p\) is proved similarly.

Finally, to prove that the profit decrease as the cost distribution increases in the reverse hazard ordering, we use the (already proved) fact that the service level decrease with the cost distribution. In particular, the left hand side of \(3\) increases with the cost distribution and since \(\tilde{G}\) is decreasing it means that the capacity goes down. Namely, that if \(F_1\) is smaller than \(F_2\) in the reverse hazard order then \(NF_1(\eta_1^*) \geq NF_2(\eta_2^*)\) and

\[ \Pi(NF_1(\eta_1^*)) - \Pi(NF_2(\eta_2^*)) = p \int_{NF_2(\eta_2^*)}^{NF_1(\eta_1^*)} xg(x) \, dx + NF_1(\eta_1^*)F_1(\eta_1^*)/f_1(\eta_1^*) - NF_2(\eta_2^*)F_2(\eta_2^*)/f_2(\eta_2^*) \]

\[ \geq NF_2(\eta_2^*) \left( p\tilde{G}(NF_2(\eta_2^*)) - p\tilde{G}(NF_1(\eta_1^*)) \right) + NF_2(\eta_2^*) \left( F_1(\eta_1^*)/f_1(\eta_1^*) - F_2(\eta_2^*)/f_2(\eta_2^*) \right) \]

\[ = NF_2(\eta_2^*) \left[ p\tilde{G}(NF_2(\eta_2^*)) - F_2(\eta_2^*)/f_2(\eta_2^*) \right] - \left( p\tilde{G}(NF_1(\eta_1^*)) - F_1(\eta_1^*)/f_1(\eta_1^*) \right) \]

\[ = NF_2(\eta_2^*) |\eta_2^* - \eta_1^*| \geq 0. \]

The last equality follows from \(3\) that implies, in particular, that \(p\tilde{G}(NF_1(\eta_1^*)) = \eta_1^* + F_1(\eta_1^*)/f_1(\eta_1^*)\). The final inequality follows from the fact (already proved) that compensation increases with agent cost, i.e., that \(\eta_2^* \geq \eta_1^*\). This concludes the proof that profits increase as agent costs decrease in the reverse hazard ordering.

\[ \square \]
Proof of Theorem

Suppose that \( N \) is such that \( \eta^*(N) > \beta \). Recall that the firm’s profits are increasing in \( N \) and the compensation is decreasing in \( N \). Thus, the firm will optimally increase \( N \) (and decrease \( \eta^*_N \)) until it hits \( \beta \). Thus, any optimal solution must have \( \eta^* = \text{beta} \). The firm’s optimal \( N \), is then given by maximizing (over \( N \)), the profits

\[
\Pi(NF(\beta)) = pS(NF(\beta)) - \beta NF(\beta).
\]

Recall that the problem \( \max_{x \geq 0} \Pi(x) \) is a standard problem and has an optimal solution \( x^* \). characterized by \( \bar{G}(x) = (\beta + F(\beta)/f(beta))/p \). The optimal level of \( N \) is thus given by the (unique) solution to

\[
\bar{G}^1(NF(\beta)) = \frac{\beta}{p}.
\]

or, equivalently,

\[
N^* = \frac{\bar{G}^{-1} \left( \frac{\beta}{p} \right)}{F(\beta)}.
\]

This concludes the proof of the first part of the theorem. For the second part, take \( N \neq N^* \) with \( \eta^*_N < \beta \). The firm, to meet, the earnings constraint must increase the compensation to \( \beta \) in which case \( NF(\beta) \) agents sign up and the firm’s profit is given by

\[
\Pi(NF(\beta)) = p \int_0^{NF(\beta)} xg(x) \, dx + pNF(\beta) \bar{G}(NF(\beta)) - \beta NF(\beta).
\]

Recall that

\[
\Pi(A(\beta)) = p \left( \int_0^{A(\beta)} xg(x) \, dx + A(\beta) \bar{G}(A(\beta)) \right) - \beta A(\beta)
\]

\[
= p \int_0^{A(\beta)} xg(x) \, dx
\]

where we use the fact that, by definition, \( \bar{G}(A(\beta)) = \beta/p \). In turn,

\[
\Pi(NF(\beta)) - \Pi(A(\beta)) = p \int_{A(\beta)}^{NF(\beta)} xg(x) \, dx + pNF(\beta) \bar{G}(NF(\beta)) - \beta NF(\beta)
\]

Note that

\[
p \int_{A(\beta)}^{NF(\beta)} xg(x) \, dx \leq pNF(\beta) \left( \bar{G}(A(\beta)) - \bar{G}(NF(\beta)) \right),
\]

Thus,

\[
\Pi(NF(\beta)) - \Pi(A(\beta)) \leq pNF(\beta) \left( \bar{G}(A(\beta)) - \bar{G}(NF(\beta)) \right) + pNF(\beta) \bar{G}(NF(\beta)) - \beta NF(\beta) = 0
\]

where we used the fact that \( \bar{G}(A(\beta)) = \beta/p \). In fact, since \( NF(\beta) > A(\beta) \) (strictly),

\[
pNF(\beta) \left( \bar{G}(A(\beta)) - \bar{G}(NF(\beta)) \right) > p \int_{A(\beta)}^{NF(\beta)} xg(x) \, dx
\]

we can conclude that

\[
\Pi(NF(\beta)) - \Pi(A(\beta)) < 0,
\]

so that the firm is better off with the cap. Moreover, by the definition of \( A(\beta) \) it is immediate that \( A(\beta) \) is the optimal cap. \( \square \)
Proof of Theorem. For item (i), we fix $N$. Recall that $\eta^*_h(N)$ and $\eta^*_l(N)$ are characterized through the equations

$$G_h(NF(\eta^*_h(N))) = \frac{\eta^*_h(N) + \frac{F(\eta^*_h(N))}{f(\eta^*_h(N))}}{p} \quad \text{and} \quad G_l(NF(\eta^*_l(N))) = \frac{\eta^*_l(N) + \frac{F(\eta^*_l(N))}{f(\eta^*_l(N))}}{p}$$

Suppose, to reach a contradiction, that $\eta^*_h(N) < \eta^*_l(N)$. Then, using the log-concavity of $F$ (which implies, in particular, that $F/f$ is increasing), we have that

$$G_h(NF(\eta^*_h(N))) = \frac{\eta^*_h(N) + \frac{F(\eta^*_h(N))}{f(\eta^*_h(N))}}{p} < \frac{\eta^*_l(N) + \frac{F(\eta^*_l(N))}{f(\eta^*_l(N))}}{p} = G_l(NF(\eta^*_l(N)))$$

(9)

Since $G_h(x) \geq G_l(x)$ for all $x \geq 0$ and since $G$ has a strictly positive density, $\eta^*_h(N) < \eta^*_l(N)$ would also imply that

$$G_h(NF(\eta^*_h(N))) \geq G_l(NF(\eta^*_l(N))) \geq G_l(NF(\eta^*_h(N)))$$

which is a contradiction to (9). It must be, then, the case that $\eta^*_h(N) \geq \eta^*_l(N)$. Consequently, the staffing levels satisfy $NF(\eta^*_h(N)) \geq NF(\eta^*_l(N))$. Also, utilizing again the fact $F/f$ is increasing, we have that

$$1 - G_h(NF(\eta^*_h(N))) = 1 - \frac{\eta^*_h(N) + \frac{F(\eta^*_h(N))}{f(\eta^*_h(N))}}{p} < 1 - \frac{\eta^*_l(N) + \frac{F(\eta^*_l(N))}{f(\eta^*_l(N))}}{p} = 1 - G_h(NF(\eta^*_h(N)))$$

so that the service level is higher on low demand periods.

We turn to item (ii). Consider a pool size $N$ such that $N < A_h(\beta)/F(\beta) = \bar{G}_h^{-1}(\beta/p)/F(\beta)$. We will show that such a level cannot be optimal. There are two cases to consider.

Suppose, first, that $\eta^*_h(N)$ (the uncapped compensation as characterized by the solution to (3)) satisfies $\eta^*_h(N) < \beta$. From item (i) of this theorem we know that, in that case, $\eta^*_l(N) \leq \eta^*_h(N) < \beta$. For a given level $N$, we can treat both periods as independent and, as in the second part of Theorem, the firm will set its compensation levels at $\beta$ and use, in both types of periods, caps at $K_i = A_i(\beta)$ and $\bar{K}_h = A_h(\beta)$ where $A_i(\beta)$ is the solution to (3) for $i = l, h$. The assigned capacity is $NF_h(\beta) \wedge A_h(\beta)$ and $NF_l(\beta) \wedge A_l(\beta)$ in the high and low demand periods, respectively, and the firm’s profits are given by

$$\Pi_h(NF_h(\beta) \wedge A_h) + \Pi_l(NF_l(\beta) \wedge A_l),$$

which is strictly increasing in $N$ for $N < A_h(\beta)/F_h(\beta)$. Since $\eta^*_h(N)$ is decreasing in $N$, the use of the cap continues to be optimal as the firm increases $N$ until it reaches $N = A_h(\beta)/F(\beta)$. Thus, any optimal solution must have $N^* \geq A_h(\beta)/F(\beta)$. If the firm chooses $N = \bar{G}_h^{-1}(\beta/p)/F(\beta)$ no cap is needed at the high demand period because $NF(\beta) = A_h(\beta)$. A cap is needed in the low demand period unless $\eta^*_l(N) = \eta^*_h(N) = \beta$ (by item (i) of this theorem it always holds that $\eta^*_l(N) \leq \eta^*_h(N)$).

Suppose, instead, that there exists $N < A_h(\beta)/F(\beta)$ with $\eta^*_h(N) > \beta$. Since the earnings constraint is not binding the firm will use $\eta^*_h(N)$ as the optimal compensation. For such values of $N$, firm does not need to use a cap in the high demand period. If $\eta^*_l(N) < \beta$ it will use a cap as above in the low demand period as above. If $\eta^*_l(N) > \beta$ then the firm will decrease also its compensation on low demand periods.

In either cases, the firm increases its profits by increasing $N$ as long as $\eta^*_l(N) > \beta$. Let $\tilde{N}$ be the smallest pool size such that $\eta^*_l(\tilde{N}) = \beta$. We claim that $\tilde{N} \geq A_h(\beta)/F(\beta)$. Indeed, if $\tilde{N} < A_h(\beta)/F(\beta) = \bar{G}_h^{-1}(\beta/p)/F(\beta)$ then the firm will decrease also its compensation on low demand periods.
\( \bar{G}_h^{-1}(\beta/p)/F(\beta) \), then, \( \bar{G}_h(\tilde{N}F(\beta)) < \beta/p \) but, at the same time (by (\ref{3})), \( \bar{G}_h(\tilde{N}F(\beta)) = (\beta + F(\beta)/f(\beta))/p \geq \beta/p \) which is a contradiction. We conclude that \( \tilde{N} \geq A_h(\beta)/F(\beta) \). Note that if \( \tilde{N} > A_h(\beta)/F(\beta) \) the firm will optimally set a cap in both demand periods to \( K_l = A_l(\beta) \) and \( K_h = A_h(\beta) \). Thus, the firm can continue decreasing \( N \) until it hits \( A_h(\beta)/F(\beta) \) without decreasing its profits. At this point no cap is needed in the high demand period because \( N = A_h(\beta)/F(\beta) \) and it is needed in the low demand periods if \( \eta^0_l(N) < \eta^0_h(N) = \beta \).

\( \Box \)