Operations in the on-demand economy: Staffing services with self-scheduling capacity

Itai Gurvich, Martin Lariviere and Antonio Moreno-Garcia
Kellogg School of Management, Northwestern University, Evanston, IL 60208,

Motivated by recent innovations in service delivery such as ride-sharing services and work-from-home call centers, we study capacity management when workers self-schedule. We assume that agents determine for themselves whether or not to work in a given period. The service provider thus seeks to maximize its profit (revenue from served customers minus capacity costs) when it controls capacity only indirectly. Agents choose when to work based on the compensation offered and their individual availability. To guarantee adequate capacity, the firm must offer sufficiently high compensation.

These novel service platforms provide a variety of benefits to the firm, the agents and the service's users. However, our analysis shows that self-scheduling can impose costs on the firm and its customers. Relative to the setting in which the firm can dictate when agents work, the firm has lower profits and the customers a higher chance of not being served. Furthermore, in the face of time varying demand, self scheduling results in lower service level in high demand periods. We show that the firm has an incentive to increase its pool of agents in order to drive down the compensation rate it must offer. If the firm must offer a minimum compensation rate, it no longer chooses an arbitrarily large pool but it does limit agent flexibility by restricting the number of agents that can work in some time intervals. Our key results are robust to the agent-compensation mechanism and to the pricing capability of the firm.

Key words: strategic servers, on-demand economy, independent capacity, distributed systems, service operations, Uber.

1. Introduction

Staffing in service environments is a challenging problem. Firms must control costs while assuring adequate capacity to serve demand. 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 day, each worker has been told when she is expected to begin and end her shift. Furthermore, these directives have been backed by implicit (and often explicit) consequences for not adhering to an assigned schedule.

In many novel service settings, however, firms are surrendering this power. Instead of ordering workers to punch in and out at appointed times, firms are allowing agents to create their own schedules, choosing whether and when to work based on personal preferences. 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 industries such as ride-sharing services (e.g., Uber

and Lyft), work-from-home call centers (e.g., Arise Virtual Solutions and LiveOps), or delivery services (e.g., Instacart) which must have capacity available to service demand as it arises.

These service providers have put themselves in a tenuous position. On the one hand, they need to provide their customers with good service. Ride-sharing services, for example, 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[1] flexibility and cannot simply dictate when they should work. Grocery shopping service Instacart, for example, states that its delivery agents can create their own flexible schedule on its recruiting page.[2] Work-from-home call center LiveOps makes a similar pitch:

> "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!"[3]

Flexibility and control of one's schedule are important to agents. In a study of Uber drivers (conducted for Uber), 85% of respondents cited the ability "to have more flexibility in my schedule" as a motivation to drive for the company (Hall and Krueger 2015). Additionally, Lyft and Uber have pointed to the fact that drivers set their own schedules in contesting lawsuits on whether drivers should be deemed employees or independent contractors (Levine and McBride 2015). Consequently, these firms cannot simply renege on allowing agents to self-schedule. They must instead use incentives schemes to induce the right number of agents to be available at the right time.

A service provider must also assure that its agents have adequate earnings over time. Some of these firms aggressively recruit and compete with each other for agents.[4] There are blogs that inform about work conditions in competing services.[5] Firms are consequently rightly concerned when websites ask whether a particular company is a "work-from-home scam"[6] or when former

[1] Describing the people serving customers for these firms requires some finesse. Generally, those answering calls or driving customers are not employees. 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.

[2] www.instacart.com/shoppers. Accessed April 17, 2015.

[3] http://join.liveops.com/ Accessed August 28, 2014

[4] http://www.forbes.com/sites/ellenhuet/2014/05/30/how-uber-and-lyft-are-trying-to-kill-each-other/

[5] http://therideshareguy.com/category/lyft-vs-uber/

[6] See workathomemoms.about.com/od/callcenterdataentry/a/arise.htm accessed on Aug 28, 2014.

agents complain in public forums that a firm is "the worst company ever" offering "below average" pay.[7]

The provider's problem can thus be understood as managing agent participation on two different time scales. On a longer-term basis (measured in weeks or months), the firm must maintain an adequate pool of eligible agents. In order to keep agents in the pool, the firm must ensure that agents earn enough to make collaborating with the firm an attractive opportunity. On a short-term basis (measured in hours or less), it must attract enough – but not too many – agents for each time interval over some horizon to maximize its profit while achieving a desired service level.

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 in each period 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 it instead used alternative compensation schemes, such as a piece rate, that depend on the number of customers an agent serves. Finally, we allow the firm to impose a **cap** on 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 employ a newsvendor setting and find that the optimal decision is an elegant variant of the classical critical fractile solution. Specifically, suppose that the firm offers agents a wage of η and receives revenue p from successfully serving a customer. Under conventional staffing (i.e., assuming that the firm can order any number of agents to work in a given period), the firm would employ enough agents so that the probability of turning customers away is $\frac{\eta}{p}$. The corresponding probability under self-scheduling, however, is $\frac{\eta}{p} + \frac{F(\eta)}{pf(\eta)}$, where F is a distribution governing agent availability and f its density.

We immediately have that self-scheduling is costly to the firm. The firm chooses a lower staffing level than it would if it could dictate when agents work (assuming the same wages) resulting in lower profits. This in turn is costly to customers who face a higher chance of not being served.

[7] See www.glassdoor.com/Reviews/Employee-Review-LiveOps-RVW2743190.htm accessed on Aug 28, 2014.

Poor customer service is exacerbated when demand varies over the horizon. We consider a horizon with both high and low demand periods (in the sense of having a stochastically larger or smaller demand distribution). The service provider offers agents higher pay in high demand periods and thus makes more capacity available. However, the service level customers see falls.

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 demand periods and still induce enough agents to work. Once there is a constraint on agent earnings, however, the firm cannot slash wages. This drives up costs both because it pays more and because that higher pay induces too many agents to work. In particular, low-volume periods will be overstaffed. Capping the number of active agents addresses this problem. Note that this implies that agents must sacrifice some scheduling flexibility in order to guarantee a minimum compensation level.

The necessity of a cap does not go away if one replaces a per-period wage with a piece rate. However, its role changes. Under a per-period wage, a cap keeps the firm from paying for agents it does not want at the prevailing wage. Under a piece rate, the cap keeps excessive competition between agents from diluting agents' earnings.

Our work is related to the literature on principal-agent models (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 customers with homogeneous requests that any available agent can handle. The question is then not how one job gets matched with one agent, but how the firm can assure it has sufficient capacity to meet demand.

To our knowledge there is only one other paper that explicitly deals with self-scheduling agents. Ibrahim and Arifoglu (2015) consider a firm facing a two-period staffing problem. Each agent prefers

one period over the other but each is certain to work in some period. The firm must then decide how many agents to induce to work in each period given a queuing structure with abandonments. We consider an arbitrary number of periods and our model manipulates agent availability as opposed to preferences for specific periods.

2. Model

We consider a service provider selling to customers over a horizon composed of T time intervals. In period t (for $1 \le t \le T$), the firm's revenue is determined by the number of available agents and market conditions. Let A_t denote the number of agents available in period t and $\mathbf{A} = (A_1, \ldots, A_T)$. We assume that each agent can serve one customer per period making the firm's staffing level equivalent to its capacity. We assume that market conditions in period t are captured by a probability distribution G_t. That is, the actual demand in period t, D_t, is drawn from G_t. One expects, for example, G_t to exhibit day of the week or time of day seasonality (e.g., demand for ride sharing services is highest at rush hour). Let $\mathbf{G} = (G_1, \ldots, G_T)$.

Let $R_t(A_t, G_t)$ denote the firm's revenue in a period with A_t available agents and market conditions G_t. Then,

$$R(A_t, G_t) = pS_t(A_t) = p\left(\int_0^{A_t} x g_t(x)\, dx + A_t \bar{G}_t(A_t)\right), \tag{1}$$

where g is the density of G, which we assume to be strictly positive, and $\bar{G} = 1 - G$. Note that $S_t(A_t)$ represents expected unit sales in period t given staffing level A_t. Note that the retail price p is fixed over the horizon. In §4, we allow the firm to choose p. The firm's revenue over the horizon is $R_T(\mathbf{A}, \mathbf{G}) = \sum_{t=1}^{T} R_t(A_t, G_t)$.

We assume that the firm pays agents η_t for being available in period t. For now, we assume that compensation is implemented through a per-interval compensation (e.g., paying \$15 per hour). We discuss alternative compensation schemes in §4. The firms profit at period t is then given by

$$\Pi(A_t, G_t) = R(A_t, G_t) - \eta_t A_t,$$

and its profit over the horizon by $\Pi_T(\mathbf{A}, \mathbf{G}) = \sum_{t=1}^{T} \Pi(A_t, G_t)$

Given η, the firm would like to use staffing levels $\mathbf{A}^*$ that maximize Π_T and schedule A_t^* agents to be available in period t. However, under self scheduling 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. In the case of a ride sharing service such as Uber, the pool would consist of all drivers in a geographic area that have been through the firm's review process. 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 in that period to be 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 τ 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 consider having F depend on the time period in §4. 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 only working in periods in which they expect to earn more than the availability threshold they have drawn for that period. Thus, an agent with realized availability threshold τ in period t makes herself available to work if the firm offers compensation η_t greater than τ. The total number of agents *interested* in working in period t is then $NF(\eta_t)$. 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 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 $\Pi_T(\mathbf{A}, \mathbf{G})$ by manipulating its available control levers. We consider three. The first is the pool size N. Since training agents takes time, this decision must be made up front. The pool size is thus constant over the horizon. The second variable is the 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, $NF(\eta_t)$, exceeds the number the firm wants, it can choose to limit access only to the number it needs. With a cap K_t, the staffing level is $A_t = NF(\eta_t) \wedge K_t$. Allowing the possibility of an access cap 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.[8]

To this basic problem we can add a constraint related to agent welfare. As discussed in the introduction, 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. We consider a *per-period earnings* constraint that requires $\eta_t \geq \beta$ for all $t = 1, \ldots, T$, i.e., that the compensation

[8] Other rationing mechanisms are possible; for example, Netessine and Yakubovich (2012) discuss several settings in which better workers are given priority.

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. If $\beta \equiv 0$ the firm faces no earnings constraints.

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

$$\begin{aligned} &\max_{\eta,N,K} \Pi_T(\mathbf{A},\mathbf{G}) \\ &\text{s.t. } A_t = NF(\eta_t) \wedge K_t,\ t=1,\dots,T, && \text{(Participation)} \\ &\qquad \eta_t \geq \beta,\ t=1,\dots,T. && \text{(Earnings)} \end{aligned}$$

3. Analysis

In this section we establish the following three key results:

(i) Theorem 1: Participation is costly to both firm and customers. Relative to a setting where it can summon the necessary agents for a "industry standard" market wage, the firm is going to have a smaller profit, staff with fewer agents and, in turn, provide a lower service level.

(ii) Theorem 2: Agents are better off in high demand periods (their compensation is higher) but customers are worse off (their service level is lower).

(iii) Theorem 3: In the presence of earnings constraints, the firm must use all of the tools in its toolbox. To maximize its profit, it is necessary for the firm to cap access in low demand periods but not necessarily in high demand period. Thus, an earnings guarantee comes at a cost to agents – their flexibility to self-schedule will be compromised.

We will further show that these findings are robust to whether the firm pays a per-period wage or a piece rate. However, the role of the cap changes between these settings; see §4.

3.1. The cost of self scheduling

To begin, we assume that customer arrivals are identically distributed in each period, i.e., $G_t \equiv G$, 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 the firm does not consider earnings constraints (i.e., $\beta = 0$), and that no access caps are used.

The firm then maximizes $R(A,G) - \eta A$ where

$$R(A,G) = pS(A) = p\left(\int_0^A xg(x)\,dx + A\bar{G}(A)\right), \tag{2}$$

The following lemma is immediately derived from the first-order condition.

Lemma 1 *The unique optimal compensation level η^* satisfies*

$$G(NF(\eta^*)) = 1 - \frac{\eta^* + \frac{F(\eta^*)}{f(\eta^*)}}{p}. \tag{3}$$

The uniqueness of η^* follows from the logconcavity of F, 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 η^*.

It is instructive to contrast how many agents the firm induces to work under self scheduling to the number it would order to work in a natural benchmark problem. Specifically we take as a benchmark a staffing problem in which the firm can order any number of agents to work at given rate η:

$$\Pi^*(\eta) := \max_A \; pS(A) - \eta A,$$

which has the solution $A(\eta)$ given by

$$G(A(\eta)) = 1 - \frac{\eta}{p}. \tag{4}$$

The following theorem shows that self-scheduling (specifically, the participation constraints) reduces the staffing (hence service level) as well as the firm's profit. (All proofs appear in the appendix.)

Theorem 1 *For any given $N \geq 0$, the participation constraint decreases the firm's staffing level and profit , i.e., $NF(\eta^*) \leq A(\eta^*)$ and $\Pi(NF(\eta^*), G) \leq \Pi(A(\eta^*), G)$.*

We thus have a clear statement that self-scheduling is costly to the firm. It ends up with fewer agents (and thus less revenue) than it would want to have at the compensation rate η^*. 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. Customers also pay a price as they have a greater chance of not receiving service if the firm opts for self scheduling.

The extent of the difference in outcomes between the self-scheduling setting and the benchmark newsvendor depends on problem parameters. We next consider how the pool size, the retail price and the distribution of the agent's availability threshold affect the firm's actions. To this end, we write η^*_F to capture explicitly the dependence of the optimal agent compensation on the availability threshold distribution.

Lemma 2 *The firm's profit increases as either N increases, p increases, or the availability threshold distribution decreases in the reversed hazard rate order: i.e., $\Pi(NF_1(\eta^*_{F_1,N}), G) \geq \Pi(NF_2(\eta^*_{F_2,N}), G)$ 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 decreases as either N increases or the availability threshold distribution decreases in the reversed hazard rate order. Agent compensation increases as p increases. Finally, the service level increases as either N or p increase.

A large pool promises more agents with low availability thresholds, which allows for a lower payment to agents. Additionally, since $\frac{F(\eta)}{f(\eta)}$ is increasing in η, (so that it decreases as η decreases) 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.

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. As the retail price increases, agent compensation also increases. However, these gains are not so large that they result in a lower firm profit. Also, customers who pay more may reasonably expect better service.

Notice that Lemma 1 makes no claim about the dependence of the service level on the threshold distribution. This is because a clear statement cannot be made as Figure 1 illustrates. Demand is uniform over $[0, 100]$ while the availability threshold is one of two power function distributions, $F_1(x) = x$ for $0 < x < 1$ or $F_2(x) = x^2$ for $0 < x < 1$. Note that F_2 is larger than F_1 in the reverse hazard rate ordering. The left-hand panel of Figure 1 shows that, as stated in Lemma 1, the offered compensation falls as either the pool size increases or the availability distribution decreases (from F_2 to F_1). The right hand panel shows, however, that the lower compensation rate under a smaller distribution does not necessarily translate to a higher service level. For smaller pool sizes, customers see a higher service level when agents have the larger availability threshold distribution F_2. This relationship is reversed when the pool size is large. Intuitively, two countervailing forces are at

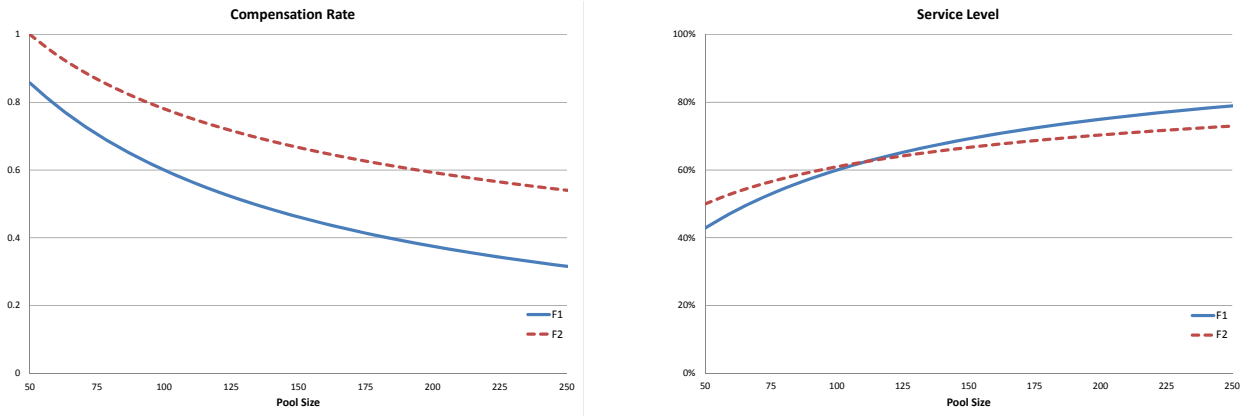

Figure 1: The impact of pool size and availability threshold distribution on agent compensation (left-hand panel) and service level (right-hand panel)

play. On the one hand, a smaller threshold distribution means that more agents will be willing to work for any value of η, which favors a higher service level. On the other, a smaller threshold distribution means that the firm faces a higher virtual cost for any value of η, which argues for a lower service level. Larger pool sizes amplifies the former effect causing it to be the dominating force as the number of agents grows.

3.2. The role of an earnings constraint

To this point, we have ignored any constraint on agent earnings. We now consider the impact of such a constraint. First, it is obvious that if η^*, for a given N, is greater than β (where η^* is determined by (3)), the earnings constraint is not binding and the firm can use η^*. If $\eta^* < \beta$, the firm optimally sets the compensation per interval at β and gets $NF(\beta)$ interested agents. The following is a characterization of the firm's optimal decision.

Lemma 3 *The optimal solution with earnings constraints has $\eta^* = \beta$ and $N^* = \bar{G}^{-1}(\beta/p)/F(\beta)$. For all values 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 it wants at compensation β – just as we assumed in our benchmark problem. Hence, its staffing level and profit are the same as under the benchmark problem.

3.3. Time-varying demand

Note that when the firm is allowed to choose the optimal pool size N^*, capping the number of agents that work is not necessary if demand does not vary. The cap is important when we move to a time-varying environment as we now show. Suppose 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 in the sense of first order stochastic dominance. Let us assume that there are T_l intervals of low demand and T_h of high demand. The firm thus faces the problem of maximizing

$$\Pi_T(\mathbf{A}, \mathbf{G}) := p\left(T_l S_l(A_l) + T_h S_h(A_h)\right) - (\eta_l T_l A_l + \eta_h T_h A_h),$$

where S_i $(i = l, h)$ is as in (2) with G replaced by G_i. Let $\eta_{i,N}^*$ be the solution to (3) when the demand distribution is G_i, $i = l, h$, the pool size is N, and $K = \infty$. Let $A_i(\beta) = \bar{G}_i^{-1}(\beta/p)$ be the solution to (4) with $\eta = \beta$ and the demand distribution G_i, $i = l, h$.

Theorem 2 *Fix N and suppose that $\beta = 0$. Then, the optimal compensation is lower on low demand periods, i.e., $\eta_{l,N}^* \leq \eta_{h,N}^*$, and the staffing level is, consequently, lower. The service level is, however, higher in low demand periods.*

Theorem 3 *Suppose that $\beta > 0$. Then, every optimal solution has $\eta^*_{l,N} = \eta^*_{h,N} = \beta$ and $N^* \geq A_h(\beta)/F(\beta) = \bar{G}_h^{-1}(\beta/p)/F(\beta)$. The assigned capacity satisfies $N^*F(\beta) \wedge K^*_l = A_l(\beta)$ and $N^*F(\beta) \wedge K^*_h = A_h(\beta)$. In particular, in every optimal solution, the firm uses a cap $K^*_l = A_l(\beta)$ in the low demand period.*

These theorems offer two important insights. First, the ability to cap the number of active agents is a crucial to controlling the firm's costs when it must guarantee a minimum earning level. Without it, low demand periods would be overstaffed. Second, even with a cap, customers will see worse service in high demand periods unless the firm has a very large agent pool (i.e., $N \geq N^*$).

Before closing this section, we make a couple of observations. First, we could consider other revenue models than the newsvendor. All that is needed for the insights to persist is that the expected unit sales be increasing and concave in the staffing level. One could, for example, suppose that sales are given by an Erlang loss model in which recruiting more agents results in fewer loss sales.

That said, the newsvendor is applicable in a wide variety of settings. Facing significant uncertainty in call volume, a newsvendor model provides a good approximation for call-center optimization; see e.g. Bassamboo et al. (2010). To relate this specifically to our setup, consider a call center with a single group of servers serving a single type of customers with finite patience. If the number of agents is A and the call volume comes from a distribution G. Then, $S(A)$ provides a good approximation for the number of calls served. The average number of calls that abandon is the expected volume minus those served. If a contract with a client compensates the call center a p for each call served, the call center is optimizing capacity so as to maximize $pS(A) - \eta A$ just as in our study above. In this way, the self-scheduling newsvendor captures, at least in first order, the challenges faced by a call-center provider such as Arise Virtual Solutions or LiveOps.

4. Extensions

We now consider three extensions of our base model: (i) alternative compensation schemes in which agent earnings are tied to the volume of customers served, (ii) a price-dependent newsvendor setting in which the firm sets both agent compensation and the retail price; and (iii) agents that have preferences that vary over the horizon.

4.1. Volume-dependent compensation schemes

Thus far the firm, in our model, implemented its policy by paying agents a fixed per-period amount η^*: an agent that signs-up to work Monday 10:00-10:30 gets η^* regardless of the number of customers served. Ride-sharing service such as Lyft and Uber compensate drivers by splitting fares with them. Similarly, call centers like LiveOps and Arise Virtual Solutions use piece-rate compensation or

some combination of piece rate and a guaranteed per-interval minimum. The firm might reasonably 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.

Here we show that within our model with risk-neutral agents many reasonable compensation mechanisms are equivalent. Namely, that there exists a translation from one scheme to the other that generates the same outcomes in terms of staffing, service level and firm profit.

In particular, we will focus on *piece-rate* compensation in which an agent earns ϕ_t per completed transaction in period t. To determine how much an agent earns under such a scheme, we need to know how many transactions she completes. Let X_t^j denote the number of customers agent j serves in period t. The distribution of X_t^j depends on a several factors including the number of active agents, the demand distribution G_t, and how the firm allocates work among agents. If some agents are given higher priority as demand is allocated among working agents, they will earn more money than those with lower priority. Here we assume that jobs are distributed uniformly among the active agents.[9] Thus, if x_t jobs arrive on interval t and A_t agents are active, each will receive roughly x_t/A_t jobs. Since x_t is random, each agent will receive $S_t(A_t)/A_t$ where $S_t(\cdot)$ is as in (2) with G replaced by G_t. An agent's expected earnings in period t $\mu_t = \phi_t S(A_t)/A_t$ and the number of interested agents is then $NF(\phi_t S(A_t)/A_t)$.

Recall that the number of agents interested in working in period t when the firm pays a fixed amount η_t is $NF(\eta_t)$. In comparing these values, note that under a fixed rate scheme, an agent can determine whether or not to work by considering only her own availability threshold. Under a piece rate scheme, however, an agent must consider both her threshold and what other agents are doing. We must in this case consider an equilibrium among the agents in which the number of interested agents is equal to the number that join, i.e, that $NF\left(\phi_t S_t(A_t)/A_t\right) = A_t$.

Lemma 4 *Fix N, G_t, and ϕ_t. There then exists an equilibrium A_t^e, characterized by the unique solution to the equation*

$$NF\left(\frac{\phi_t S_t(A_t^e)}{A_t^e}\right) = A_t^e. \tag{5}$$

It is a priori conceivable that this equilibrium structure introduces constraints into the firm's optimization problem or, in other words, that an optimal solution (N^*, η^*, K^*) to the firm's optimization problem is not implementable via a piece rate. The following simple argument is a proof

[9] For a setting in which routing is non-uniform see Stouras et al. (2013).

to the contrary: the firm can move from per-interval compensation to piece-rate compensation without compromising its profits or customer service level.

Suppose that the firm is using a feasible solution (N, η, 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

$$NF\left(\frac{\phi'_i S_i(A_i)}{A_i}\right) = A_i,$$

where $A_i = NF(\eta_i)$. Since (N, η, 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 ϕ' follows from the continuity of F. With this choice of ϕ', the number of agents that sign-up in equilibrium is (using Lemma 4) the unique solution to $NF(\phi'_i S_i(A_i)/A_i) = A_i$ which must equal A_i by construction.

If (N^*, η^*, K^*) is an optimal solution to the firm's optimization problem, then the firm can set the piece rate at $\phi^*_i = \eta^*_i A_i / S_i(A_i)$ where $A_i = N^* F(\eta^*_i) \wedge K^*_i$. With this translation, the optimal solution (N^*, η^*, K^*) to the firm's problem with interval compensation is equivalent to the solution (N^*, ϕ^*, 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 staffing level is the same and, hence, (iii) the expected firm profits and customer service level are the same.

There are, however, subtle 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. 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 size 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 ϕ from making themselves available.

Lastly, the cap on the number of active agents plays a different role under a piece rate than under a fixed per-period compensation. Under the latter, the firm's labor costs are fixed with regard to the demand realization. The cap serves to control this fixed cost and prevents the firm from paying for labor on which it would expect an inadequate return. From this perspective, a cap would seem to be unnecessary when the firm moves to a piece rate system. Labor is no longer a fixed cost and unutilized capacity is apparently costless to the firm. That unutilized capacity, however, is costly to the firm. A large number of available agents reduces everyone's expected utilization and expected earnings for a given ϕ. Competition between agents undermines the firm's ability to compensate

agents adequately with a relatively low piece rate. Without a cap, the firm would have to raise the piece rate driving up its cost of serving customers.

Piece rate and fixed compensation are two extremes. Fixed compensation means agents do not face any volume risk while under a piece rate they carry all the risk. A two-part tariff (i.e., $\nu_t + \phi_t X_t^j$) or a piece rate with a minimum guarantee (i.e., $\max\left\{\kappa_t, \phi_t X_t^j\right\}$) offer intermediate mechanisms. A call center we have worked with pays agents a piece rate with a guaranteed minimum payment level. Uber has also been reported to guarantee an hourly rate at some time periods (Kirsner 2014). Given our analysis for piecerate it is not surprise (and, indeed, can be easily shown) that these mechanisms are also equivalent, within our model, to fixed compensation.

4.2. Price-dependent newsvendor

Our assumption thus far, that retail price is fixed regardless of whether demand is high or low, is appropriate in some settings (e.g., a work-from-home call centers). Yet, other services (notably ride-sharing firms) raise their prices when demand increases. This calls into question one of our earlier results that customer experience a lower service in high-demand periods. In Theorem 2 we showed that if the retail price is fixed, an increase in staffing costs results in the firm picking a lower service level. If now the retail price also increases, it is not clear that it is still optimal to let customer service level fall.

To examine these issues, we suppose that demand in a low-volume interval given a retail price p is a random variable ξ_p with distribution $G_l(x|p)$ and that ξ_p becomes smaller in the sense of first order stochastic dominance as p increases. That is, $G_l(x|p) \leq G_l(x|\hat{p})$ for all x for all $p \leq \hat{p}$. Next we assume that demand in a high-volume interval for a given price is $\theta\xi_p$ for some $\theta > 1$. The corresponding demand distribution is then $G_h(x|p) = G_l\left(\frac{x}{\theta}|p\right)$. We assume that there is sufficient structure on $G_l(x|p)$ that the firm has unique, pricing and staffing decisions for both low and high volume periods (see Petruzzi and Dada 1999).

Letting $S_i(A,p)$ be expected unit sales in a type $i \in \{l,h\}$ period given staffing level A and retail price p, one can show that

$$S_h(A,p) = \theta S_l\left(\frac{A}{\theta}, p\right).$$

Now consider the benchmark problem in which the firm can hire as many agents as it wants at wage η in either period. Let $\hat{A}_i(\eta)$ and $\hat{p}_i(\eta)$ be the optimal staffing level and price, respectively, for a type $i = l,h$ interval. It is straightforward to show that $\hat{p}_h(\eta) = \hat{p}_l(\eta)$ and $\hat{A}_h(\eta) = \theta\hat{A}_l(\eta)$. Thus a firm which manages its staff in a conventional fashion does not employ period-dependent pricing; it sticks with the same retail price and adjusts its staffing level to achieve the same service level in both high and low volume periods.

This will no longer hold if the firm allows agents to self schedule. Increasing staff above $\hat{A}_l(\eta)$ requires dipping further into the pool of agents which, in turn, drives up staffing cost. Higher costs then lead to a higher retail price. That is, prices surge higher in this framework not because of the market structure but because of higher costs.

The fact that the firm uses surge pricing does not yet tell us how service-level behaves and whether (or not) making the retail price endogenous leads to a departure from Theorem 2. To examine how the service level varies with demand under self scheduling, we work with a specific demand distribution. Suppose that $G_l(x|p) = \frac{x}{D(p)}$ where $D(p)$ is a non-negative strictly decreasing function. Since demand is uniformly distributed, the expected demand (given a price) p is $\frac{D(p)}{2}$. Let $\varepsilon(p) = -\frac{pD'(p)}{D(p)}$ be the elasticity of expected demand. We assume that $\varepsilon(p)$ is increasing, i.e., that demand becomes less elastic as the price falls.

We first consider the benchmark problem, which for high-volume periods is written as

$$\max_{p,A} pS_h(A,p) - \eta A.$$

With the uniformly distributed demand, we have

$$S_h(A,p) = A - \frac{A^2}{2\theta D(p)},$$

which results in the following first order conditions:

$$\frac{A}{\theta D(p)} = 1 - \frac{\eta}{p}, \tag{6}$$

$$\varepsilon(p) = \frac{2}{\frac{A}{\theta D(p)}} - 1. \tag{7}$$

Equation (6) is the classical critical fractile solution that ties the capacity A to a targeted service level. Equation (7) relates the service level to the elasticity of demand: the higher the service level, the lower the elasticity. To go the other way, a higher elasticity corresponds to a lower service level.

Substituting (6) into (7) yields an implicit expression for the optimal price $\hat{p}$

$$\hat{p} = \eta \frac{1+\varepsilon(\hat{p})}{\varepsilon(\hat{p})-1}. \tag{8}$$

Equation (8) does not depend on θ. This is a specific instance of our more general argument above that, in the benchmark problem, the retail price is not sensitive to the scale of demand so the same retail price is optimal in high and low demand periods. Further, the right-hand side of (8) is decreasing in p if $\varepsilon(p)$ is strictly increasing. Consequently, $\hat{p}$ is increasing in η so that higher agent wages move the firm to a higher level of elasticity on the demand curve. Going back to (7),

a higher elasticity corresponds to a lower service level. Thus when faced with higher staffing costs, the firm charges more but offers worst service.[10]

Turning to the self-scheduling setting, the firm's problem for given a pool-size of N is

$$\max_{p,\eta} pS_h\left(NF\left(\eta\right),p\right)-\eta NF\left(\eta\right),$$

which yields the following first order conditions (the analogues to (6) and (7))

$$\begin{aligned}\frac{NF\left(\eta\right)}{\theta D\left(p\right)} &= 1-\frac{\eta+\frac{F(\eta)}{f(\eta)}}{p}\ ,\\ \varepsilon\left(p\right) &= \frac{2}{\frac{NF(\eta)}{\theta D(p)}}-1.\end{aligned}$$

Comparing the problem of the self-scheduling firm with the benchmark setting, we again see (recall (3)) that the self-scheduling firm works with an inflated marginal cost of capacity. That higher cost leads to higher price than one would have in the benchmark problem with wage rate η and, if $\varepsilon\left(p\right)$ is strictly increasing, to a lower service level. Further, the chosen compensation rate is now increasing in θ. Thus a firm that lets its agents self-schedule will charge more but, as in Theorem 2, offer worse service in high-demand periods.

4.3. Period-dependent threshold distributions

In our base model, the distribution of agent threshold values is independent of the period. This is obviously unrealistic. Many people opt to work for a self-scheduling firm in part because existing obligations (e.g., having young children) make working a conventional schedule difficult. However, many of these obligations have a known schedule (e.g., the preschool gets out at the same time every weekday); this should be reflected in the distribution of threshold values.

Here, we suppose that the distribution of threshold values depends on the time interval of the horizon. Let Ω_T be the set of all time intervals. Let Ω_d and Ω_u be subsets of Ω_T such that $\Omega_d\cup\Omega_u=\Omega_T$ and $\Omega_d\cap\Omega_u=\emptyset$. An agent draws her threshold value for period t from F_d $[F_u]$ if $t\in\Omega_d$ $[t\in\Omega_u]$. Further,

$$F_u\left(t\right)\leq F_d\left(t\right)$$

for all t. Ω_d is then the set of *desirable* time intervals in the sense that each agent has a higher probability of drawing a low threshold in these intervals than in the *undesirable* intervals of Ω_u. Stated another way, a given compensation rate will induce more agents to work in a desirable interval than in undesirable interval.

[10] This conclusion depends on $\varepsilon\left(p\right)$ being strictly increasing. If $D\left(p\right)=p^{-\tilde{\varepsilon}}$, the elasticity of demand is constant at $\tilde{\varepsilon}$, making $\hat{p}$ proportional to η and the optimal service level independent of η.

When there is no earnings constraint, agent preferences over periods only modestly complicate the firm's problem. Let Ω_h and Ω_l be the collection of time periods in which demand is high or low respectively. When the threshold distribution that does not depend on the time period, the compensation offered in period t depends only on whether t falls in Ω_h or Ω_l. If the distribution varies with the time interval, compensation in period t depends on whether t lies in $\Omega_d \cap \Omega_h$ or $\Omega_d \cap \Omega_l$ and so on. The firm must then calculate four different compensation rates using the appropriate version of (3).

Things get more interesting when the firm must satisfy a non-trivial per-period earning. Now the firm offers β in every period and the question becomes how large a pool to recruit. Regardless of whether a high demand period falls in Ω_d and Ω_u, the firm would want $A_h(\beta) = \bar{G}_h^{-1}(\beta/p)$ working. The pool size necessary to achieve this staffing level is higher if the period in question is undesirable. Thus, if $\Omega_u \cap \Omega_h \neq \emptyset$, the firm sets $N = A_h(\beta)/F_u(\beta) > A_h(\beta)/F_d(\beta)$. However, assuming $\Omega_d \cap \Omega_h \neq \emptyset$, then preferred, high-volume periods will be overstaffed. Consequently, we conclude that the firm may cap access to high-volume periods if it must satisfy a non-trivial earnings constraint and the distribution of agent thresholds varies with the time period.

It is conceptually straightforward (albeit notationally cumbersome) to extend these results to having two types (say, A and B) of agents each of which has its own set of desirable and undesirable time periods. Assume that the firm offers the same compensation to both types of agents.[11] If there is no earnings constraint, the firm would now need to have eight different compensation rates that depend on the demand and whether or not the period is desirable for Type A agents, Type B agents or both. With a non-trivial earnings constraint, the action turns on how many agents of each type to recruit. The number of agents will be determined by a subset of the possible kinds of high-demand periods (e.g., periods that both types undesirable and those that Type A agents find undesirables but Type B desires but not those periods that both types desire). Caps will not be necessary in the periods that determine the staffing levels but will in other high-demand periods.

5. Concluding Remarks

We studied a model in which a service provider allows its agents to choose when to work, a scheme being adopted in many service markets. 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 to both the firm and its customers. The firm picks a lower service level than it would in a standard newsvendor setting, lowering its profits and making it harder for customers to get served. These issues are mitigated if the firm can recruit a large

[11] This would be appropriate if the agents types (say, stay-at-home parents and college students) affects their availability but not their productivity.

pool of agents. As the pool size grows, the firm pays agents less and less and the gap between self-scheduling and the benchmark newsvendor problem decreases. Of course, agents are worse off.

Interestingly, the gap between self-scheduling and the benchmark problem is also closed if the firm must guarantee that agents earn some minimum amount per period. Indeed, in this setting the gap completely disappears. The firm chooses its pool large enough that it can get, on a period-by-period basis, exactly the number of agents it needs at the guaranteed wage. However, to control its cost, the firm must cap the number of agents working in some period. That is, if the firm must satisfy an earnings constraint, it limits the ability of agents to work when they want. This result is robust to whether firm offers a fixed wage per period or a volume-dependent piece rate.

Our model is an abstraction of reality and there are ways in which it could be extended. For example, one could consider competition between service providers for agents. 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

Allon, G., A. Bassamboo, E. B. Çil. 2012. Large-scale service marketplaces: The role of the moderating firm. *Management Science* **58**(10) 1854–1872.

Bassamboo, A., R. S. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bergstrom, T., M. Bagnoli. 2005. Log-concave probability and its applications. *Economic theory* **26** 445–469.

Hall, J. V., A. B. Krueger. 2015. An analysis of the labor market for ubers driver-partners in the united states. Working paper, Princeton.

Ibrahim, R., K. Arifoglu. 2015. Managing large service systems with self-scheduling agents. Working paper, UCL.

Kalanick, T. 2012. Surge pricing follow up. URL `http://blog.uber.com/2012/01/03/surge-pricing-followup/`.

Kirsner, S. 2014. What happened when Boloco founder John Pepper became an Uber driver. URL `http://www.boston.com/business/technology/innoeco/2014/02/what_happened_when_a_boston_en.html`.

Laffont, JJ, D. Martimort. 2009. *The theory of incentives: the principal-agent model*. Princeton Univ. Press.

Levine, D., S. McBride. 2015. Uber, Lyft face crucial courtroom test over driver benefits. *Routers* .

Moreno, A., C. Terwiesch. 2014. Reputation in online service marketplaces: Empirical evidence from a transactional dataset. *Information Systems Research.* Forthcoming.

Netessine, S., V. Yakubovich. 2012. The darwinian workplace. *Harvard Business Review* **90**(5) 25.

Petruzzi, N. C., M. Dada. 1999. Pricing and the newsvendor problem: A review with extensions. *Operations Research* **47**(2) 183–194.

Salanie, B. 1997. *The Economics of Contracts: A Primer*. The MIT Press.

Stouras, K., K. Girotra, S. Netessine. 2013. First ranked first to serve: A tournament approach to call centers. Working paper, INSEAD.

Appendix

Proof of Theorem 1: The right hand side of (3) is smaller than that of (4) so that, since G is increasing in its argument, we must have that $NF(\eta^*) \leq A(\eta^*)$. For the profit comparison, notice that $NF(\eta^*)$, would generate in the benchmark problem the profit $\Pi(NF(\eta^*), G)$. Since $A(\eta^*)$ is, by definition, the optimal solution for the fixed wages η^*, it must be the case that $\Pi(A(\eta^*), G) \geq \Pi(NF(\eta^*), G)$. □

Proof of Lemma 2: We first show the monotonicity results for the compensation followed by the service level and, finally, the profits. It is useful to re-write (3) as

$$\bar{G}(NF(\eta^*)) = \frac{\eta^* + \frac{F(\eta^*)}{f(\eta^*)}}{p}. \tag{9}$$

Compensation: That compensation strictly increases with p and strictly decreases with N is evident from (9). Consider for instance p. Suppose to reach a contradiction that, as p increases, the compensation η^* actually decreases. Then, the right-hand-side of (9) decreases by the monotonicity of F/f so that the left-hand side $\bar{G}(NF(\eta))$ must also decrease. This would entail (since F is increase and $\bar{G}$ is decreasing) that η^* increases with p which is a contradiction.

To prove that the compensation increases with the agent availability distribution F notice that, assuming F_2 dominates F_1 in the reverse hazard rate ordering,

$$\eta_2^* + \frac{F_2(\eta_2^*)}{f_2(\eta_2^*)} \leq \eta_2^* + \frac{F_1(\eta_2^*)}{f_1(\eta_2^*)}. \tag{10}$$

Further, 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, since $\bar{G}$ is decreasing,

$$\bar{G}(NF_1(\eta_2^*)) \leq \bar{G}(NF_2(\eta_2^*)) \tag{11}$$

By (9),

$$\bar{G}(NF_1(\eta_1^*)) = \frac{\eta_1^* + \frac{F_1(\eta_1^*)}{f_1(\eta_1^*)}}{p} \text{ and } \bar{G}(NF_2(\eta_2^*)) = \frac{\eta_2^* + \frac{F_2(\eta_2^*)}{f_2(\eta_2^*)}}{p},$$

so that combining (10) and (11) we have

$$\bar{G}(NF_1(\eta_2^*)) \le \bar{G}(NF_2(\eta_2^*)) = \eta_2^* + \frac{F_2(\eta_2^*)}{f_2(\eta_2^*)} \le \eta_2^* + \frac{F_1(\eta_2^*)}{f_1(\eta_2^*)}. \tag{12}$$

Assume now, to reach a contradiction, that $\eta_2^* < \eta_1^*$. Then, since the right hand side of (12) increases strictly in the compensation and the left hand side strictly decreases, we would get that

$$\bar{G}(NF_1(\eta_1^*)) < \bar{G}(NF_1(\eta_2^*)) \le \eta_2^* + \frac{F_1(\eta_2^*)}{f_1(\eta_2^*)} < \eta_1^* + \frac{F_1(\eta_1^*)}{f_1(\eta_1^*)},$$

which contradicts (9). We conclude that $\eta_1^* \le \eta_2^*$.

Service level: The fact that the service level increases with p is evident from the optimal fractile formula (9) and from the fact, already proved, that the optimal compensation increases with p. Similar is the observation that the optimal service level increases with N.

Profits: To show that profits increase with the pool size N, take $N_2 > N_1$. Let $\eta_{N_1}^*$ be the optimal compensation level at N_1. Since $N_2 > N_1$, $N_2F(\eta_{N_1}^*) > N_1F(\eta_{N_1}^*)$. In particular, we can find $\bar{\eta} < \eta_{N_1}^*$ such that $N_2F(\bar{\eta}) = N_1F(\eta_{N_1}^*)$. With this $\bar{\eta}$, then, the firm gets the same staffing level under $(N_2, \bar{\eta})$ as under $(N_1, \eta_{N_1}^*)$ and, in turn, the same revenue. The staffing costs are smaller under $(N_2, \bar{\eta})$ since $\bar{\eta}N_2F(\bar{\eta}) = \bar{\eta}N_1F(\eta_{N_1}^*) < \eta_{N_1}^* N_1F(\eta_{N_1}^*)$. Thus, the pair $(N_2, \bar{\eta})$ generates a higher profit for the firm than the pair $(N_1, \eta_{N_1}^*)$. In particular, $(N_2, \eta_{N_2}^*)$ generates higher profits that $(N_1, \eta_{N_1}^*)$.

An identical argument is used to study the effect of an increase (in the sense of reverse hazard ordering) in the availability distribution starting with the observation that, since reverse hazard rate ordering implies stochastic ordering, $NF_1(\eta_{F_2}^*) \ge NF_2(\eta_{F_2}^*)$ where N is fixed and $\eta_{F_2}^*$ is the optimal compensation under F_2. If $NF_1(\eta_{F_2}^*) = NF_1(\eta_{F_2}^*)$, then under F_1, $\eta_{F_2}^*$ generates the same profit as the optimal solution for F_1 and, in particular, the optimal profit under F_1 is higher. If the inequality is strict, i.e., $NF_1(\eta_{F_2}^*) > NF_2(\eta_{F_2}^*)$ we can proceed, as before, by finding $\bar{\eta}$ that generates the same staffing and revenue but lower staffing cost. □

Proof of Lemma 3: Let η_N^* be the optimal compensation in (3) when the pool size is N. Suppose that N is such that $\eta_N^* > \beta$. By Lemma 2, the firm's profits are strictly increasing in N and the compensation is decreasing in N. Thus, the firm will optimally increase N (and decrease η_N^*) until it hits β and we conclude that any optimal solution must have $\eta_N^* = \beta$. The firm's optimal N, is then given by maximizing (over N), the profits

$$\Pi(NF(\beta), G) = pS(NF(\beta)) - \beta NF(\beta).$$

This is a standard newsvendor problem so that the optimal level of N is given by the (unique) solution to $\bar{G}(NF(\beta)) = \beta/p$, or, equivalently, $N^* = \bar{G}^{-1}\left(\frac{\beta}{p}\right)/F(\beta)$, as claimed.

For the second part of the theorem, take $N \ne N^*$ with $\eta_N^* < \beta$. The firm, to meet, the earnings constraint must increase the compensation to β in which case $NF(\beta)$ agents sign up and the firm's profit is given by

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

Recall that

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

where we use the fact that, by definition, $\bar{G}(A(\beta))=\beta/p$. There are two cases to consider depending on whether $NF(\beta)>A(\beta)$ or $NF(\beta)<A(\beta)$. The case that $NF(\beta)=A(\beta)$ is ruled out by the assumption that $N\neq N^*$. Suppose that $NF(\beta)>A(\beta)$ (the other case is argued identically).

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

Notice that

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

Thus,

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

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

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

we can conclude that

$$\Pi(NF(\beta),G)-\Pi(A(\beta),G)<0,$$

so that the firm is better off with the cap. By the definition of $A(\beta)$ it is immediate that $A(\beta)$ is the optimal cap. □

Proof of Theorem 2: Here we fix N and omit it from the subscript. Recall that η_l^* and η_h^* are characterized through the equations

$$\bar{G}_h(NF(\eta_h^*))=\frac{\eta_h^*+\frac{F(\eta_h^*)}{f(\eta_h^*)}}{p}\text{ and }\bar{G}_l(NF(\eta_l^*))=\frac{\eta_l^*+\frac{F(\eta_l^*)}{f(\eta_l^*)}}{p}$$

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

$$\bar{G}_h(NF(\eta_h^*))=\frac{\eta_h^*+\frac{F(\eta_h^*)}{f(\eta_h^*)}}{p}<\frac{\eta_l^*+\frac{F(\eta_l^*)}{f(\eta_l^*)}}{p}=\bar{G}_l(NF(\eta_l^*)).\tag{13}$$

Since F and G have strictly positive densities $F(\eta_h^*)<F(\eta_l^*)$ so that (since $\bar{G}$ is strictly decreasing) $\bar{G}_l(NF(\eta_h^*))>\bar{G}_l(NF(\eta_l^*))$. Using the assumed stochastic ordering we then have that

$$\bar{G}_h(NF(\eta_h^*))\geq\bar{G}_l(NF(\eta_h^*))>\bar{G}_l(NF(\eta_l^*)),$$

which is a contradiction to (13). It must be then that that $\eta_h^*\geq\eta_l^*$. Consequently, the staffing levels satisfy$NF(\eta_h^*)\geq NF(\eta_l^*)$. Finally, since F/f is increasing, $\eta_h^*+F(\eta_h^*)/f(\eta_h^*)\geq\eta_l^*+F(\eta_l^*)/f(\eta_l^*)$ and

$$G_h(NF(\eta_h^*))=1-\frac{\eta_h^*+\frac{F(\eta_h^*)}{f(\eta_h^*)}}{p}<1-\frac{\eta_l^*+\frac{F(\eta_l^*)}{f(\eta_l^*)}}{p}=G_h(NF(\eta_h^*))$$

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

Proof of Theorem 3: Consider a pool size $N < A_h(\beta)/F(\beta) = \bar{G}_h^{-1}(\beta/p)/F(\beta)$. We will show that such a level cannot be optimal. There are two cases to consider depending on how $\eta^*_{h,N}$ in (3) relates to β.

Case I: $\eta^*_{h,N} < \beta$. In this case, in the absence of the earnings constraint the firm would optimally choose a compensation level below β. For a given level N, we can treat both types of periods (high and low) independently and, by Lemma 3, the firm sets its compensation levels at β and utilizes a cap $K_l = A_l(\beta)$ in the low demand periods where $A_l(\beta)$ is the solution to (4) with demand distribution G_l. In this case the cap in high demand periods is unncecessary because $NF(\beta) < A_h(\beta)$.

The active capacity is then $NF_h(\beta) \wedge A_h(\beta)$ and $NF_l(\beta) \wedge A_l(\beta)$ in the high and low demand periods. Thus, the firm's profits for values of $N < A_h(\beta)/F(\beta)$ with $\eta^*_{h,N} < \beta$ is given by

$$\bar{\Pi}(N) := T_h \Pi(NF_h(\beta) \wedge A_h(\beta), G_h) + T_l \Pi(NF_l(\beta) \wedge A_l(\beta), G_l).$$

Notice that $\bar{\Pi}(N)$ is increasing in N. Since $\eta^*_{h,N}$ is decreasing in N, it continues to hold that $\eta^*_{h,N} < \beta$ as we increase N. Thus, the firm's profit follows $\bar{\Pi}(N)$ and is increasing in N and, in particular, 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 $\eta^*_{l,N} = \eta^*_{h,N} = \beta$ (notice that by Theorem 2 it is always the case that $\eta^*_{l,N} \leq \eta^*_{h,N}$).

Case II: $\eta^*_{h,N} > \beta$. Since the earnings constraint is not binding the firm will use $\eta^*_{h,N}$ as the optimal compensation in the high demand period. By Lemma 2, the firm could increase its profit on high demand period by increasing N. If also, $\eta^*_{l,N} > \beta$, the same applies to low demand periods so that strictly increasing N is optimal. If $\eta^*_{l,N} < \beta$ the firm uses a cap in the low demand period and, as before, the firm's profit are increasing in N.

Thus, as long as N is such that $\eta^*_{h,N} > \beta$, the firm can increase its profits by increasing N. Let $\widetilde{N}$ be the smallest pool size such that $\eta^*_{h,N} = \beta$. (Recall that we treat N as continuous variable so that such a $\tilde{N}$ exists). It must be the case that $\widetilde{N} \geq A_h(\beta)/F(\beta) = \bar{G}_h^{-1}(\beta/p)/F(\beta)$. Otherwise, $\bar{G}_h(\widetilde{N}F(\beta)) < \beta/p$ but, at the same time, being a solution to (3), $\eta^*_{h,N} = \beta$ satisfies $\bar{G}_h(\tilde{N}F(\beta)) = (\beta + F(\beta)/f(\beta))/p \geq \beta/p$ which is a contradiction. We conclude that $\widetilde{N} \geq A_h(\beta)/F(\beta)$.

Finally, by Lemma 2 if $\widetilde{N}$ is strictly greater than $A_h(\beta)/F(\beta)$, the firm will set a cap to $K_l = A_l(\beta)$ and $K_h = A_h(\beta)$. As above, the firm can then decrease 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 $N = A_h(\beta)/F(\beta)$ but it is needed in the low demand periods if $\eta^*_{l,N} < \eta^*_{h,N} = \beta$. □

Proof of Lemma 4: Consider the function $h(x) := NF(\phi S(x)/x) - x$. It is easily verified that $S(x)/x \to 1$ as $x \to 0$, so that $h(x) = NF(\phi S(x)/x) - x \to NF(\phi)$ as $x \to 0$. Since F is bounded by 1 we have, as $x \to \infty$, that $h(x) \to -\infty$. Combined, we just established that both $h(x) \to -\infty$ as $x \to \infty$ and $h(x) \to NF(\phi)$ as $x \to 0$. Since F and G have densities the function $h(x)$ is continuous on $(0, \infty)$ so that there must exist x_0 such that $h(x_0) = 0$. The fact that this point is unique then follows from the fact that h is monotone decreasing. Indeed, since $S'(x) = \bar{G}(x)$, $h'(x) = Nf\left(\phi \frac{S(x)}{x}\right) \phi \frac{S'(x)-1}{x^2} - 1 = -Nf\left(\phi \frac{S(x)}{x}\right) \phi \frac{1-\bar{G}(x)}{x^2} - 1 < 0$. □