Task Allocation and On-the-job Training

Mariagiovanna Baccara* SangMok Lee† Leeat Yariv‡§

27th April 2022

Abstract

We study dynamic task allocation when providers' expertise evolves endogenously through training. We characterize optimal assignment protocols and compare them to discretionary procedures, where it is the clients who select their service providers. Our results indicate that welfare gains from centralization are greater when tasks arrive more rapidly, and when training technologies improve. Monitoring seniors' backlog of clients always increases welfare but may decrease training. Methodologically, we explore a matching setting with endogenous types, and illustrate useful adaptations of queueing theory techniques for such environments.

Keywords: Dynamic Matching, Training-by-Doing, Market Design.

*Olin School of Business, Washington University in St. Louis. Email: mbaccara@wustl.edu

†Department of Economics, Washington University in St. Louis. Email: sangmoklee@wustl.edu

‡Department of Economics, Princeton University. Email: lyariv@princeton.edu

§We thank Ricardo Alonso, Heski Bar-Isaac, Francisco Buera, Nima Haghpanah, and Chiara Margaria for helpful comments. We gratefully acknowledge financial support from the National Science Foundation, grant SES 1629613.

1 Introduction

1.1 Overview

Employees in a wide variety of careers and industries receive on-the-job training. Surgical residency programs revolve around a progressive increase in residents' responsibilities in the operating room.[1] Law firms routinely assign clients to associates on track to become partners. In such settings, employers face an ongoing allocation problem: which tasks or clients to allocate to more experienced and senior employees, and which to junior ones. Experienced providers typically offer more reliable service, but junior employees are often more available and require hands-on practice to become seniors later on. Existing (one-sided) matching protocols presume agents' attributes—here, providers' quality—are fixed over time; see, e.g., Echenique, Immorlica, and Vazirani (2021). However, in the presence of on-the-job training, the allocation of tasks within an organization affects future providers' expertise.

We propose a new and tractable framework to address the task-allocation problem when agents' types are endogenously determined through training. At the heart of the allocation problems we study is a trade-off between clients' immediate and future service quality. We characterize optimal protocols by which a social planner would assign clients to service providers. We also identify how organizations perform in discretionary settings, where it is the clients who select their service providers. This allows us to analyze conditions under which organizations would especially benefit from centralizing the task-allocation process.

Our results indicate that welfare gains from centralization are greater when tasks arrive more rapidly, and for improved training technologies. Monitoring seniors' backlog of clients always increases welfare. Nonetheless, we illustrate that monitoring may decrease the scope of training in the organization.

As common in many applications, we focus on settings in which the organization cannot price services based on providers' seniority. The allocation protocol is therefore the only instrument by which the organization can impact both the quality of immediate service and the expertise of providers in the future. If an organization insists on providing only senior service, the quality of service is high, but there is no training. Natural attrition, through retirement or job changes, would then result in a scarcity of experienced personnel, leading to prolonged wait times.[2]

[1] The American Board of Surgery specifies precise training requirements measured both through the number of surgeries and the trainee's role in them (see http://www.absurgery.org)

[2] Wait times are critical in determining organizations' performances. Elit, O'Leary et al. (2014), Wijeysundera et al. (2014), and Kaltenmeier et al. (2019) are all studies that quantify increased risks of complications and mortality associated with a longer wait between diagnosis and various surgical procedures. In judiciary

This trade-off between immediate and future service quality cannot be analyzed using standard tools offered by the matching literature. Existing models generally consider agents' "types" as exogenously given, rather than as the result of past assignments.

In our model, clients seeking service arrive over time at a Poisson rate. The organization is comprised of junior and senior providers. Service by juniors is immediate. Senior service quality is higher, but entails a costly wait. We assume clients who seek senior service form a queue. For example, medical patients may have to wait for a consultation with an experienced specialist and appointments for legal counsel are often provided on a first-come-first-served basis. The processing speed of clients by seniors depends on the seniors' volume in the organization. The more seniors available, the speedier the rate at which clients in the queue are served.

The organization's composition evolves over time. Specifically, juniors who perform service become seniors via a training technology. In steady state, the fraction of clients directed at seniors affects wait times through two distinct channels. The more clients join the senior queue, the longer the wait times. That is the direct channel. But, there is also an indirect channel through training. Fewer clients served by juniors leads to less training and slower senior processing speeds. The queueing literature offers a rich analysis of models in which servers varying in speed cater to clients that arrive over time (see, for instance, Leon-Garcia, 2008). However, in that literature, service quality or speed are fixed and independent of prior experience—there is no indirect channel. We introduce techniques for incorporating the link between evolving expertise and service quality, allowing for endogenous client arrival speeds.

We consider two alternative information environments. First, we study the case in which decision makers—the social planner in the centralized setting or individual clients in the discretionary environment—observe the current state of the queue when selecting service providers. This constitutes what we refer to as the *perfect-monitoring* case. For instance, academic department chairs could link the assignment of faculty members to various committees based on their existing workloads. Likewise, individuals seeking help from an attorney may be informed of the length of the wait time they will experience. While the perfect-monitoring case serves as a natural benchmark, in many environments queues cannot be consistently tracked. Therefore, the second information environment we consider is one in which decision makers cannot monitor the queue over time and cannot condition their choices on its current state. This is the *limited-monitoring* case. For example, when drafting a curriculum for all surgical residents

systems, trial delays often result in detainees waiting for a decision in prison, causing higher costs, overcrowding and worse living conditions. As of 2019, 18.9% of the prison population in Europe consisted of detainees waiting for a final decision on their case, see the annual reports available at http://www.prisonobservatory.org/ .

in the U.S., policy makers need to establish a required level of involvement in the operating room, which does not depend on the current logistical needs of any specific hospital. Similarly, patients with an urgent condition may not know the volume of others currently waiting in line when choosing which local emergency room to drive to.

In the perfect-monitoring case, we focus on symmetric threshold-based allocation policies: clients are served by juniors if and only if the queue for senior service has reached a certain threshold. Our first set of results fully characterize the optimal as well as the equilibrium thresholds clients utilize. The social planner accounts for the full distribution of wait times and the training constraint. Modifications of the classical queueing model allow us to transform the social planner's objective into a static constrained optimization problem. As we show, the objective is continuously differentiable and single-peaked. Therefore, we can use a first-order condition approach to establish a characterization of the optimal policy. In contrast, the equilibrium discretionary threshold is set so that the *last* client willing to wait for senior service is roughly indifferent between the two types of service.

Our characterizations allow for natural comparative statics. Certainly, a greater relative value for senior service, or lower wait costs, lead to less training but higher quality of service. The impacts of changes in clients' arrival rate or improvements in training are more subtle. More clients arriving, as a consequence of, say, closure of nearby hospitals, can potentially cause more congestion, but at the same time offers additional training opportunities. Improved technology, due to new online training options, simulated activities such as surgeries for doctors or trials for lawyers, can help generate more seniors, but potentially cause fewer clients to turn to juniors. We show that both increased clients' arrival rate and improved training technologies lead to higher expected service quality and a greater mass of seniors, both in equilibrium and under the optimal policy. The impacts on expected wait times are non-monotonic, with maximal expected wait times occurring for intermediate values of arrival rates and training efficacy.

Under limited-monitoring, decision makers, the social planner or the clients themselves, choose the *probability* with which they enter the seniors' queue, without seeing its status. Our second set of results characterizes the optimal and equilibrium probabilities of seeking senior service with limited monitoring. As in the perfect-monitoring case, the optimal policy maximizes the expected welfare subject to the training constraint, and we utilize the classical queueing model to identify a unique solution. In the discretionary equilibrium, the fraction of clients joining the queue for senior service makes *any* client indifferent between senior and junior service.

Some of the comparative statics corresponding to the limited-monitoring case resemble those of the perfect-monitoring case. Both increased clients' arrival rate and improved training technologies yield increases in average service quality. However, their effects on expected wait times are different. An increase in either leads to monotonically decreased wait times for senior service in the centralized setting, and no change in wait times in the discretionary setting. With limited monitoring, the welfare gap between the optimal and discretionary settings increases as clients' arrival rate grows or the training technology improves—such changes in the environment make centralized interventions unequivocally more impactful on clients' welfare.

Centralization naturally increases overall welfare, regardless of monitoring precision. Discretionary settings feature higher average service quality at the cost of longer wait times. Indeed, in discretionary settings, each client waiting in line for senior service imposes two types of externalities. First, she imposes a longer wait on those that follow her in the queue. Second, she deprives the organization from potential training opportunities, resulting in longer future wait times for senior service.

Our last set of results evaluates the effects of monitoring precision on centralized and discretionary processes. Perfect monitoring allows decision makers to utilize the senior queue only when it is sufficiently short. In contrast, with limited monitoring, agents can only rely on expectations of queue length.

In the discretionary setting, the impacts of monitoring are nuanced. If the training technology is relatively inefficient, we show that a higher fraction of clients seek senior service under perfect monitoring than under limited monitoring. Improved monitoring then increases the average service quality but decreases training. The reverse holds if the training technology is highly efficient. Nonetheless, regardless of the training technology's efficacy, the equilibrium welfare is always higher under perfect monitoring.

In the centralized setting, monitoring is always beneficial. With perfect monitoring, the social planner can emulate the fraction of clients sent to seniors in the limited-monitoring case, but do so more efficiently, directing clients to the senior queue only when it is short enough. Such a policy maintains service quality and reduces wait times. As we show, in the optimal policy, the social planner chooses a higher threshold than that. Namely, in the centralized setting, monitoring leads to higher quality and less training.

Taken together, our results suggest the value of centralization, particularly when clients' arrival rate is high and training is efficient. They also indicate the value of monitoring in such allocation problems. Nonetheless, the benefits of monitoring come at a cost organizations

should be aware of. While monitoring always increases clients' overall welfare, it can result in less training. This is the case in centralized settings and, when the training technology is relatively inefficient, also in discretionary environments.[3] We hope the new methodology we introduce opens the door for future work that allows agents' characteristics, in our case service providers' expertise, to evolve with their market experiences.

1.2 Related Literature

To our knowledge, there is no work inspecting allocation protocols when agent characteristics are endogenous to the protocol itself. Specifically, we are not aware of theoretical or empirical work that discusses the link between task assignments and the scope of training in organizations.

Different aspects of our model are reminiscent of work in several areas. The problem of how an organization should optimally juggle tasks arriving over time has been studied in the context of judicial systems by Coviello, Ichino, and Persico (2014) and Bray, Coviello, Ichino, and Persico (2016). Gavazza and Lizzeri (2007) consider a model of queueing for services and study service providers who maximize their free time and can increase their service speed at a cost. Increasing transparency, by revealing wait times to clients, is then detrimental to efficient servers and reduces servers' incentives to invest in service speed. Nonetheless, the training component, and the heterogeneity of service providers it generates, is absent from these papers.

Settings related to supervised training, absent a task-allocation decision or considerations of service delays, are explored in several papers. For example, Lizzeri and Siniscalchi (2008) consider parents who decide how much to shelter their children from mistakes, which are risky but provide useful learning opportunities. Garicano and Rayo (2017) study the optimal training speed from the perspective of an employer who, at every period, can determine how much knowledge to transfer to an apprentice. Larger knowledge transfers increase both the apprentice's productivity as well as her outside option if she decides to leave the employer. They show that the trade-off results in inefficiently long apprenticeships.[4]

Work in organization theory has studied how to allocate opportunities to heterogeneous

[3] Certainly, one could consider centralized solutions that account for the amount of training per se in their objective. We return to this point in our conclusions.

[4] There is also a vast literature on workers' training in general equilibrium models of human capital accumulation (see, for example, Acemoglu, 1997, and Acemoglu and Pischke, 1999). This work abstracts from the task-allocation problem with on-the-job training. It typically focuses on how market frictions can explain why firms are willing to invest in workers' training despite the fact that market competition and labor mobility prevents them from reaping its full returns. Chari and Hopenhayn (1991) consider a dynamic model of technological innovation, where investment in new technologies depends on prior investments in older technologies—for instance, through the training of employees.

individuals who may have comparative advantages in exploiting them. This question has inspired insights on the optimal way to design knowledge-based organizations by, among others, Garicano (2000) and Garicano and Rossi-Hansberg (2012). While agents' expertise evolves in some of these models, the operating mechanism is quite different than ours.[5] The idea that task assignment may change randomly over time depending on the organization's needs is explored in Bird and Frug (2021). None of these papers examines the interaction between task allocation and employees' training.

Our paper is also related to a recent and growing literature on dynamic allocation and matching, starting from the seminal work of Ünver (2010). For a review of that literature, see Baccara and Yariv (2021). To our knowledge, heterogeneity of types in this literature is always assumed to be exogenously determined.[6]

Our analysis expands on techniques from the queueing literature. For a review of the relevant models see, for example, Hassin and Haviv (2003) or Leon-Garcia (2008).

2 A Simple Model of On-the-job Training

2.1 The Setup

Our model focuses on client or task allocation with on-the-job training. Clients seeking service arrive at the system over time $t \in [0, \infty)$ following a Poisson process with arrival rate λ. There are two types of service providers: juniors and seniors. We assume that seniors are better equipped to handle clients. Formally, we assume the value corresponding to a senior handling a client is h, whereas the value derived from a junior handling a client is l, where $h > l > 0$. The difference $h - l$ can stand for the literal difference in the service quality provided, for the relative risk of critical mistakes during service, and so on.

For simplicity, we assume there is an infinite pool of juniors. Therefore, clients directed at juniors experience no wait. Clients directed at seniors form a queue and are served on a first-in-first-out (FIFO) basis. Let $\mu \in \mathbb{R}_+$ denote the mass of seniors. As the mass of senior providers increases, senior service becomes more rapid. Formally, the completion of service provided by

[5] For example, Garicano and Rossi-Hansberg (2012) explore a dynamic setting in which individuals acquire skills by experiencing exceptional problems related to new technologies. Some acquire more problem-solving expertise than others and, over time, these "experts" can use their skills to solve problems experienced by others by becoming managers in hierarchical organizations, or external consultants.

[6] In particular, several papers in that literature have utilized queueing models with exogenously-fixed agent types: e.g., Bloch and Cantala (2017), Leshno (2021), and, Ashlagi, Burq, Jaillet, and Manshadi (2019). Monitoring quality is rarely considered as a market-design instrument, although Arnosti, Johari, and Kanoria (2015) illustrate the impact of transparency in employment markets.

seniors follows a Poisson process with parameter μ. The mass μ need not be interpreted literally as the volume of seniors. Rather, it can be any service speed proxy that responds to the mass of seniors.[7] In principle, one could consider various tasks the organization handles, ones in which seniors have an advantage and others in which they do not. In our setting, easy tasks that both provider types can handle equally well would naturally be directed to juniors. We focus on the allocation of more difficult tasks, where a meaningful trade-off exists.

Since clients directed at seniors might experience a wait in the queue, the overall payoff from directing a client to seniors is $h - cW$, where $c > 0$ is the waiting cost and W is the client's wait time in the queue. We assume that a client does not experience waiting costs while receiving service. For the applications we consider, one's own expected service time may affect payoffs differentially—for example, when waiting for medical service, a patient's appointment duration depends on the particular doctor she meets, but her overall wait time depends on the set of available doctors. Any additional costs entailed by the time spent receiving service are incorporated in h and l.

We consider centralized and discretionary allocation problems. In a *centralized allocation*, a planner allocates clients to seniors or juniors. We consider a benevolent social planner whose objective is to maximize the average client payoff. While normatively appealing, such a benevolent planner can reflect umbrella organizations that issue guidelines for professional training across organizations.

In a *discretionary allocation*, upon their arrival, clients *choose* whether to join the queue for senior service, or seek immediate service by juniors. We also consider varying degrees of monitoring. With *perfect monitoring*, decision makers—the social planner or the clients—observe the queue for senior service and allocation decisions can be contingent on its current length. With *limited monitoring*, decision makers do not observe the evolving status of the queue for senior service. Therefore, allocation decisions are independent of the current queue status or past client allocations.[8]

Last, we consider a stylized model of training. In either the discretionary or centralized setting, the volume of clients served by juniors determines the mass of seniors in the organization. There is a training constraint that requires a certain time-average number of clients to be served by juniors in order to maintain any mass of seniors. Formally, we posit a training

[7] One could consider a discrete version in which the (integer) *number* of seniors evolves over time and providers can serve clients simultaneously, corresponding to multi-server models in the queueing literature. We discuss this alternative setting at the end of the section.

[8] In each of the settings we analyze, clients are guaranteed an expected payoff of at least $l > 0$. Thus, participation constraints are satisfied ex-ante.

function $f : \mathbb{R}_+ \to \mathbb{R}_+$. When the time-average number of clients sent to juniors is x, the mass of seniors is given by $\mu = f(x)$. We assume that f is differentiable, (weakly) increasing, and (weakly) concave in x. We also assume that $f(0) \leq \lambda$ so that it is not feasible for all clients to join the queue for senior service without that queue imploding and generating arbitrarily long wait times.[9] Since μ represents the speed at which seniors process clients, any improvement in training technology corresponds to an upward shift of the function f.

The training function we consider subsumes various features of promotion and hiring procedures in organizations. For example, $\mu = f(x)$ can be implicitly defined as the solution of $p\mu = g(x)$, where g is a production function and senior providers' mass decays every time unit by a fraction of p because of exit (e.g., retirement) or skill deterioration. The training technology can also reflect a screening process of job applicants or junior employees. In such a scenario, applicants, new employees, or interns are assigned some clients and l represents the average skill level in the general intern population. The function f captures the fraction of high-quality interns that are then selected for permanent positions. The linear technology case, where $f(x) = ax$, with $a > 0$, is a particularly useful one to consider. While it is in many ways special, it fits well with numerous applications. For instance, in the screening interpretation, it reflects cases in which only a fixed fraction of juniors justify promotion. In general, the parameter a can be a proxy for the training efficacy.

There are several special cases that serve as important benchmarks:

- **Infinitely costly delay** ($c \to \infty$). This case captures settings in which allocations are urgent, such as surgeries for trauma patients, emergency fire or police calls, etc.

- **Costless delay** ($c \to 0$). Routine tasks and services, such as dental check-ups, low-stakes legal proceedings, and so on, are often characterized by very low costs of delay.

- **No on-the-job training** ($f(x) = \overline{\mu}$ for any x). This is the case when skill acquisition on the job is limited in scope or duration, as is arguably the case for journal editorial teams, higher court judges, grant-allocation panels, etc. It is also the case for low-skilled labor-intensive jobs, including, for example, many jobs in the food and construction industries. A constant production function additionally captures settings in which training practices are separate from hiring practices or environments in which turn-over is high. For instance,

[9] By a well-known queueing theory result (see details in the Appendix), if all clients join the queue, namely $x = 0$, the average waiting time in the queue with a processing rate $\mu(> \lambda)$ is $\frac{1}{\mu-\lambda} - \frac{1}{\mu}$, which becomes arbitrarily large as μ approaches λ. The other potential "corner" allocation, $x = \lambda$, never occurs since $h > l$.

many academic departments employ post-docs that are not destined to receive a tenure-track position.

2.2 Discussion of Assumptions

Several assumptions in our setting merit discussion.

FIFO queueing protocol We assume that the queue for senior service is governed by a FIFO protocol. While this assumption has no impact on the characterization of the optimal centralized mechanisms, it is important for our results pertaining to equilibrium outcomes in discretionary settings.[10] The order of arrivals is tied to the order of service in many organizations, and FIFO is commonly used. For example, when scheduling a medical visit with a specialist, patients often have the option to select the first available appointment on the calendar. Indeed, queues for an assortment of, if not most, services—construction jobs, home and car improvements, etc.—operate on a FIFO basis. Other priority protocols such as last-in-first-out (LIFO) are well-known to reduce negative externalities in discretionary settings, but at the same time involve significant implementation challenges.[11]

Flow costs of waiting In our model, clients incur a fixed flow cost for the duration of their wait. An alternative way to model waiting costs would be through discounting. We use flow costs in our setting for two reasons. First, we believe flow costs might be more important for design objectives in the applications we speak to. Indeed, with discounting, once clients have waited for a very long time, their marginal contribution to welfare becomes negligible, and a social planner can all but ignore them in allocation decisions.[12] This is hardly the case when clients are medical patients seeking treatment, students awaiting their grades, etc. As noted, FIFO protocols are commonly used in practice for these sorts of applications. The underlying premise of FIFO is that those who waited longer should not be punished, and is therefore antithetical to discounting in settings such as ours. The second reason for considering flow costs is technical in nature, as it allows for far greater tractability. Indeed, there are several difficulties discounting presents. With discounting, the benefits of serving a client depend on the time that client already spent in the system. As a result, the relevant state space for a

[10] In a centralized setting, the social planner's objective function incorporates the average wait experienced by the clients. Therefore, because of waiting costs' linearity, the planner's optimal policy is unaffected by the priority protocol.

[11] In particular, they are subject to manipulation as they introduce incentives to leave and re-enter queues. They are also sometimes considered "unfair" in that individuals who just entered the system are served first, while others, who have been waiting, remain in the queue. See Margaria (2019) and references therein.

[12] Ortoleva, Safonov, and Yariv (2021) discuss how discounting impacts optimal allocations of items.

social planner is vast: each state specifies not only the number of clients waiting in the queue, but also their precise arrival times. In addition, the randomness present in our environment suggests that the timing of service is in itself a random variable. Keeping track of expected exponentially discounted values then introduces non-trivial complications in itself.

Single-server setting Our model speaks to tasks that are dealt with by one server. These can be thought of as small tasks, e.g. suture application in the medical context, or drafting an estate plan in the legal context.[13] For tractability, we conside a single-server queue for senior service. Our setup approximates a multi-server queue when the steady-state probability that all senior servers are idle in the multi-server queue is close to 0. This occurs in our setting if $\frac{h-l}{c} \to \infty$. When no server is ever idle—when waiting costs are low, or when the benefits of senior service are high—the rate at which a client waiting first in line is served corresponds to the *minimal time* at which one of the servers completes her current task. The minimum of Poisson-distributed variables follows a Poisson distribution. For example, if n servers with service rate $\frac{\mu}{n}$ are never idle, the overall service time follows a Poisson distribution with parameter μ.[14]

Unlimited supply of juniors Our model assumes an infinite supply of juniors available at any time. Consequently, clients seeking junior service experience no wait. This assumption is designed to capture the idea that, in many settings, unskilled labor is more readily available than more experienced labor. Certainly, one could assume that juniors are available in limited supply as well and that clients seeking their service wait in a separate queue. The model would then be less tractable. We view such an extension as an interesting direction for future analysis.

3 Perfect Monitoring

We start by analyzing the case in which the volume of clients queueing to be served by seniors is observed: by the social planner in the centralized setting, and by entering clients in the

[13] One could certainly think of more elaborate tasks, where juniors may be trained by a senior they work with. Such mentorship-based training can be captured by an explicit dependence of the production technology on the mass of seniors available. Such an extension is beyond the scope of the current paper.

[14] A formal argument for the limited monitoring case is as follows. Take the arrival rate $q\lambda$ and (aggregate) process rate μ. Let $\mathbf{E}[W]$ be the clients' average wait time in the single-server setup, which we study. Let $\mathbf{E}[W_n]$ be the average wait time in the multi-server setup, where n servers each have a processing rate of $\frac{\mu}{n}$. If the server utilization $\rho = \frac{q\lambda}{\mu}$ is close to 1,

$$\frac{\mathbf{E}[W_n]}{\mathbf{E}[W]} = \frac{\frac{C(n,n\rho)}{\mu - q\lambda}}{\frac{q\lambda}{\mu(\mu-q\lambda)}} = \left(\rho + \rho(1-\rho)\frac{n!}{(n\rho)^n \sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!}}\right)^{-1} \to 1,$$

where $C(.,.)$ is the Erlang C formula. A similar argument holds for the perfect monitoring case.

discretionary setting. This analysis serves as a useful benchmark for quantifying the impact of monitoring technology improvements (see Section 6). In addition, this information environment fits numerous applications well. For example, in many organizations, administrators are assigned duties based on their current workload. Analogously, patients may call clinics to learn about their expected wait times in advance of deciding which provider to seek service from. Moreover, in some settings in which clients have discretion over service choice, new technologies allow monitoring of the line prior to a decision—e.g., apps such as No Wait or Yelp allow patrons to observe the line at restaurants before arriving on the premises. In what follows, we analyze the outcomes such monitoring yields.

3.1 Threshold-Based Allocations

Both in the discretionary and in the centralized settings, we focus on symmetric threshold-based allocation policies: clients are served by readily available juniors if and only if the queue for senior service, including those waiting or being served, has reached a threshold k.[15]

Formally, we consider a continuous-time Markov chain for the number of clients waiting in the queue or currently being served. With fixed arrival and service rates, the processes we analyze correspond to those referred to as M/M/1/k queues in the queueing literature (see our primer in the Online Appendix). In our setting, however, both the seniors' service rate and the arrival rate into their queue are determined endogenously by the chosen threshold.

Without loss of generality, we assume that $k \geq 1$ throughout our analysis. Whenever the seniors queue is empty, it is optimal for any individual client and the planner to seek senior service, which comes at a higher quality. The queue size is bounded above by k, and the unique steady-state distribution across the relevant states of the queue, $\{0, 1, \ldots, k\}$, is as follows.[16]

Lemma 1 (Steady-state Distribution under a Threshold Policy) *Given $\mu > 0$ and $k \geq 1$, the probability p_j of having j clients in the queue for senior service in the unique steady state is*

$$p_j = \begin{cases} \frac{1}{k+1} & \text{if } \mu = \lambda \\ \frac{(\lambda/\mu)^j(1-(\lambda/\mu))}{1-(\lambda/\mu)^{k+1}} & \text{if } \mu \neq \lambda. \end{cases} \qquad \forall j = 0, 1, \ldots, k. \tag{1}$$

[15] Since the length of service is distributed exponentially, the expected time of service completion is independent of the time at which service has begun. It follows that the relevant statistic for a newly-arrived client is the *number* of clients in the senior queue, including any client currently being served.

[16] The state $j = 0$ corresponds to no clients present, $j = 1$ corresponds to one client currently being served, $j = 2$ corresponds to one client being served and one waiting for service, and so on.

Let Q denote the number of clients waiting in the senior queue, excluding those being served.

$$\mathbf{E}[Q] = \sum_{j=1}^{k}(j-1)p_j = \sum_{j=1}^{k}(j-1)(\lambda/\mu)^j p_0.$$

After algebraic manipulation (see details in the Appendix), we can write:

$$\mathbf{E}[Q] = \begin{cases} \frac{k(k-1)}{2(k+1)} & \text{if } \lambda = \mu \\ \frac{1}{1-(\lambda/\mu)^{k+1}}\left(\frac{(\lambda/\mu)^2-(\lambda/\mu)^{k+1}}{1-(\lambda/\mu)} - (k-1)(\lambda/\mu)^{k+1}\right) & \text{if } \lambda \neq \mu. \end{cases} \tag{2}$$

Clients seek junior service when the state is $j = k$. Denote the average fraction of clients served by seniors by $q \equiv 1 - p_k$. In steady state, the time-average number of clients seeking senior and junior service are $q\lambda$ and $(1-q)\lambda$, respectively.[17]

Let $\mathbf{E}[W]$ denote these clients' average waiting time before being served by seniors, conditional on joining the (possibly empty) queue. As we show in the Appendix, there is a close link between the expected wait time and the expected length of the queue, denoted by $\mathbf{E}[Q]$. Namely, by Little's formula,

$$\mathbf{E}[Q] = \lambda q \mathbf{E}[W].$$

The mass of seniors is governed by the training technology and given by $\mu = f((1-q)\lambda)$.

For exposition sake, it is convenient to relax the integer constraints on k.[18] In Lemma 2, we establish the one-to-one correspondence between any real-valued threshold k and the associated fraction of clients served by seniors q (i.e., the service quality) for any given μ.

Lemma 2 (Service Quality under a Threshold Policy) *For all μ,*

$$q(k;\mu,\lambda) \equiv 1 - p_k = \begin{cases} \frac{k}{k+1} & \textit{if } \lambda = \mu, \\ \frac{1-(\lambda/\mu)^k}{1-(\lambda/\mu)^{k+1}} & \textit{if } \lambda \neq \mu \end{cases} \tag{3}$$

is strictly increasing in $k \in [1,\infty)$, with values in $[\frac{\mu}{\mu+\lambda}, \frac{\mu}{\lambda})$.

Lemma 2 allows us to describe any outcome in terms of either (k,μ) or (q,μ). In what follows, we characterize solutions in terms of (q,μ) as it facilitates a direct comparison of

[17] The arrival of clients assigned to juniors or seniors does not follow a Poisson process. Indeed, a client assigned to juniors suggests the senior queue is long. Hence, a client served by juniors is likely to be closely followed by another.

[18] For any given λ and μ, the formulas for p_k, $\mathbf{E}[Q]$, and $\mathbf{E}[W]$ are defined for any real-valued $k \geq 1$. One can readily derive the corresponding formulations that take the integer constraint into account. In the next footnote we discuss the implications of relaxing the integer constraint on the resulting policies.

solutions under perfect and limited monitoring, which we consider later. For a given pair (q, μ), the corresponding threshold k is identified by the inverse of (3):

$$k(q;\mu,\lambda) \equiv \begin{cases} \frac{q}{1-q} & \text{if } \lambda = \mu \\ \frac{\log(1-q)-\log(1-(q\lambda/\mu))}{\log(\lambda/\mu)} & \text{if } \lambda \neq \mu. \end{cases} \tag{4}$$

3.2 Discretionary and Centralized Allocations

Consider first the case in which clients have discretion over which service to seek upon entering the market. A symmetric equilibrium (k, μ) is defined through two constraints. First, each client optimizes her expected payoff, which we soon spell out, given the size of the queue she observes upon entry and the mass μ of seniors. In particular, each client prefers to join the seniors' queue as k-th in line, but not as $(k+1)$-th in line. Second, the mass μ of seniors is consistent with the training opportunities governed by the threshold k. Namely, for the induced fraction q of clients seeking senior service and characterized in Lemma 2, the remaining fraction $1-q$ of clients is served by juniors. Therefore, it must be that $\mu = f((1-q)\lambda)$. We call this last equality the *training constraint*.

When all clients use the threshold k, any client who arrives when there are $m \geq k$ clients in the senior queue approaches juniors, who are immediately available. Since the senior queue follows a FIFO protocol, the position of any client waiting can only improve over time. In particular, a client who decides to wait for senior service has no reason to leave the queue and get served by juniors at a later point. As service times are distributed exponentially, a client who joins as m-th in the queue for senior service experiences an expected wait time of $\frac{m-1}{\mu}$. Her expected payoff from joining the senior queue is then $h - c\frac{m-1}{\mu}$. Ignoring integer constraints, at the threshold k, the agent is indifferent between receiving that expected payoff, or receiving service immediately from juniors. That is, an equilibrium is defined by two restrictions: the indifference condition $h - c\frac{k-1}{\mu} = l$ and the training constraint $\mu = f((1-q)\lambda)$.

In what follows, we assume an *integer-threshold environment*. That is, we assume there exists a solution that is consistent with an integer threshold. This assumption greatly simplifies our exposition, but is not crucial qualitatively.[19]

From the social planner's perspective, threshold-based policies are optimal within the set

[19] Without this assumption, an equilibrium could be defined similarly. Let $\overline{k} \equiv \max\{k : l \leq h - c\frac{k-1}{\mu_k}\}$. Hence, $l > h - c\frac{\overline{k}}{\mu_{\overline{k}+1}}$. If $l < h - c\frac{\overline{k}}{\mu_{\overline{k}}}$, when all other clients use threshold $\overline{k}$, each one wants to use threshold $\overline{k}+1$ instead. An equilibrium would be defined by $\overline{k}$ together with a randomization, such that some fraction of clients, when finding $\overline{k}$ others in the queue, still join the queue as $\overline{k}+1$-th in line. While all our analysis' qualitative features remain, such randomization requires a custom modification to the steady state of $M/M/1/k$ queues.

of stationary policies that depend only on the number of clients waiting or being served in the senior queue.[20] We now characterize the optimal centralized threshold mechanism. The planner maximizes the clients' average payoff:

$$\max_{k,\mu} p_k l + (1-p_k)(h - c\mathbf{E}[W]).$$

subject to the training constraint $\mu = f(\lambda p_k)$. Let $\theta \equiv \frac{h-l}{c}$ denote the quality differential per unit cost. This problem is equivalent to:

$$\max_{q,\mu} q\lambda\theta - \mathbf{E}[Q] \text{ s.t. } \mu = f((1-q)\lambda).$$

In the linear-production case, $f(x) = ax$, the training constraint $\mu = a(1-q)\lambda$ implies that $\mathbf{E}[Q]$ can be directly described as a function of q.

Lemma 3 *Suppose $f(x) = ax$, for some $a > 0$. The expected number of clients waiting in the queue, $\mathbf{E}[Q]$, is described as follows:*

$$\mathbf{E}[Q] = \begin{cases} \frac{q(2q-1)}{2(1-q)}, & \text{if } q = 1 - \frac{1}{a}, \\ \frac{1}{a(1-q)-1}\left[\frac{q}{a(1-q)} - (1-q)\frac{\log(1-\frac{q}{a(1-q)}) - \log(1-q)}{\log a + \log(1-q)}\right] & \text{otherwise.} \end{cases}$$

It follows that the planner's problem in the linear-production case can be written as

$$\max_{q \in [\underline{q}, \frac{a}{1+a})} q\lambda\theta - \mathbf{E}[Q],$$

where $\underline{q}$ corresponds to the case $k = 1$. Specifically, if $k = 1$, then $q = 1 - p_k = \frac{\mu}{\mu+\lambda}$ and $\mu = a\lambda(1-q)$, so $\underline{q}$ is the solution of $q = \frac{a(1-q)}{a(1-q)+1}$. The upper bound $q < \frac{a}{1+a}$ is required to ensure that $q\lambda < \mu = a(1-q)\lambda$.[21] In the Online Appendix, we show that the objective is continuously differentiable and single-peaked. Therefore, we can use a first-order condition approach.

We have the following characterization of allocations under perfect monitoring.

[20] Consider the set of all, both deterministic and random, stationary policies. No optimal policy would require holding an indefinitely large number of clients in the queue. Therefore, it is without loss of generaility to assume that the maximum number of clients in the queue must be finite, implying a finite state space. Let μ be the mass of seniors that an optimal stationary policy yields. By Theorem 7.1.9 of Puterman (2005), holding μ fixed and ignoring the training constraint, an optimal policy is identified by a threshold k, where $k+1$ is the smallest queue length under which the policy directs an arriving client to juniors. See the Appendix of Baccara, Lee, and Yariv (2020) for details of a similar derivation of an optimal threshold policy.

[21] This is effectively a budget constraint that guarantees there are sufficient seniors to serve all those seeking their service in steady state.

Proposition 1 (Perfect Monitoring)

1. *In the discretionary setting, the unique equilibrium is governed by* (q_P^e, μ_P^e) *that solves:*

$$k(q, \mu; \lambda) = \mu\theta + 1 \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{5}$$

2. *In the centralized setting, when* $f(x) = ax$, *for some* $a > 0$, *any interior optimal policy* (q_P^*, μ_P^*) *solves*:

$$\lambda\theta = \frac{d\mathbf{E}[Q]}{dq} \quad \text{and} \quad \mu = a(1-q)\lambda. \tag{6}$$

3.3 Comparative Statics

The characterizations of the equilibrium and the optimal policies in Proposition 1 can be used to derive some comparative statics with respect to θ, λ, and the training technology. Naturally, as θ increases, either through an increase in the relative benefit $h-l$ of senior service, or through a decrease in waiting costs c, queueing for senior service becomes relatively more attractive, translating into higher average quality and lower training both in discretionary and centralized settings. In particular, when $c \to \infty$, there is no delay. In this case, in both the centralized and discretionary settings, waiting is minimized and clients choose junior service. This yields $q_P^e, q_P^* \to 0$ and $\mu_P^e, \mu_P^* \to f(\lambda)$. In contrast, as c approaches 0, more clients naturally seek senior service. Consequently, if k_P^e and k_P^* denote the equilibrium and optimal thresholds respectively, $k_P^e, k_P^* \to \infty$ and wait times can be arbitrarily large.

We now turn to the impacts of changes in arrival rates and the training technology on outcomes in our perfect-monitoring settings. Changes in arrival rates can reflect market shifts for the demand of particular services. For instance, the introduction and dissemination of electric cars could increase the demand for electricians installing home-charging units. Changing arrival rates can also reflect a redirection of clients following closure of certain service units: hospitals, law firms, etc. The resulting overall arrival rate of clients in the surviving organization would presumably be higher than the arrival rate at each of the original organizations. Higher arrival rates can potentially cause more congestion, but they also represent more training opportunities.

Improved training corresponds to technological advances. For example, the introduction of the Internet offers a multitude of opportunities for training in various tasks, from carpentry, to professional conduct. Similarly, technological advances in the medical world—e.g., the introduction of patient simulation dummies—improve training efficacy of young nurses and doctors.

The impacts of the training efficacy can also be relevant for the comparison of industries that differ in their training features or their training expenditures.[22] The following proposition provides comparative statics in the perfect-monitoring environment.[23]

Proposition 2 (Perfect Monitoring – Comparative Statics) *When $f(x) = ax$, for some $a > 0$, the following comparative statics hold:*

1. *As λ increases, k_P^e, μ_P^e, q_P^e, q_P^*, and μ_P^* increase. Furthermore, as λ approaches 0, both $\mathbf{E}[W_P^e]$ and $\mathbf{E}[W_P^*]$ tend to 0, while when λ grows indefinitely, $\mathbf{E}[W_P^e]$ tends to θ and $\mathbf{E}[W_P^*]$ tends to 0.*
2. *As a increases, k_P^e, μ_P^e, q_P^e, and μ_P^* increase. Furthermore, as a approaches 0 or grows indefinitely, both $\mathbf{E}[W_P^e]$ and $\mathbf{E}[W_P^*]$ tend to 0.*

Consider first the discretionary setting with linear training technology, $f(x) = ax$, with $a > 0$. When λ or a increase, in the space of (q, μ), the graph corresponding to the indifference condition for each λ, $G_1 = \{(q, \mu) : k(q, \mu; \lambda) = \mu\theta + 1\}$ shifts up as λ increases and does not respond to changes in a, see Figure 1. The graph $G_2 = \{(q, \mu) : \mu = a\lambda(1 - q)\}$, corresponding to the training constraint, shifts up with increases in both λ and a, also depicted in Figure 1. Consequently, q_P^e and μ_P^e increase in a. The indifference condition then implies that k_P^e increases. Furthermore, with increases in λ, the figure demonstrates that μ_P^e, and therefore k_P^e, must increase. As we show in the proof, q_P^e increases as well. We also demonstrate that similar conclusions hold for the centralized setting.

Turning to expected wait times, when either the arrival rate or the technology's efficacy are very low, senior service is slow. Thus, whether decisions are discretionary or centralized, no client will be placed in the seniors' queue if a wait is required. Consequently, conditional on seeking senior service, clients face no wait. When arrival rates are very high, or the training technology is very efficient, a positive fraction of clients sent to juniors yields extremely rapid senior service. Thus, when decisions are centralized, expected wait times would be vanishing. The impacts are more nuanced in the discretionary setting. Greater arrival rates or superior training technology increases the efficiency of senior service. As a consequence, however, more clients seek senior service. As we show in the proof, this trade-off is resolved differently for high arrival rates and effective training technology.

[22] See https://trainingmag.com/ for annual reports on training expenditures across industries in the U.S.
[23] We use $\mathbf{E}[W_P^e]$ and $\mathbf{E}[W_P^*]$ for equilibrium and optimal expected wait times, respectively.

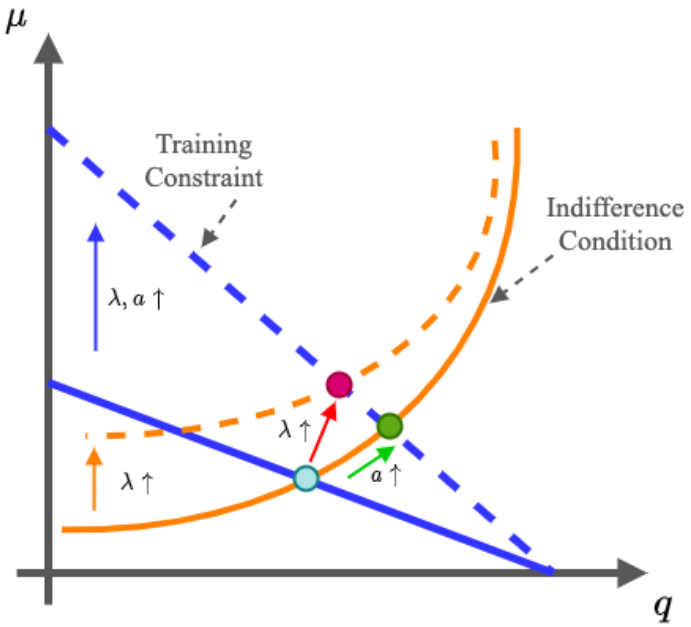

Figure 1: Comparative statics for discretionary settings with perfect monitoring

For intermediate levels of arrival rates or training efficacy, expected wait times in all our settings are strictly positive. Thus, the proposition implies a non-monotonicity of the expected wait times with respect to both clients' arrival rate and the technology's efficacy.

Proposition 2 illustrates the role of training in our environment. Without training, the mass of seniors is exogenously fixed at $\overline{\mu}$—that is, $f(x) = \overline{\mu}$ for any x. It is easy to verify that any increase in λ causes q to decrease and $\mathbf{E}[W]$ to increase in both the discretionary and centralized settings. The effects are different when non-trivial training takes place.

4 Limited Monitoring

In this section we assume that the length of the queue is not observed by decision makers: the clients in the discretionary setting or the planner in the centralized setting. In the discretionary setting, this corresponds to environments in which clients are not informed of the queue's length—namely, the number of clients ahead of them—when deciding which service to seek. For instance, patients needing urgent care may select a clinic to drive to without knowing its current load, graduate students selecting an advisor may have limited information on how busy various professors are, etc. In the centralized setting, limited monitoring corresponds to organizations in which general allocation rules are established without detailed monitoring. For example, medical associations and hospitals need to set policies on the involvement of trainees

in procedures, academic departments may set rules on the number of undergraduate theses each faculty advises, and so on.

4.1 Discretionary and Centralized Allocations

We focus on stationary and symmetric strategies by both the clients (in the discretionary setting) and the planner (in the centralized setting). The characterization in both settings with limited monitoring boils down to the fraction $q \in [0,1]$ of clients that are served by seniors. Therefore, seniors serve clients at a rate μ that is determined, as in the perfect-monitoring case, by the training constraint $\mu = f((1-q)\lambda)$.

Discretionary or centralized allocation policies in the limited-monitoring case are characterized by a pair (q, μ). Specifically, a discretionary equilibrium is defined through two constraints. First, each client optimizes her expected payoff, given all other clients' strategy of seeking senior service with probability q and the mass μ of available seniors. In particular, whenever the equilibrium is interior, $q \in (0,1)$, each client is indifferent between junior and senior service. Second, the induced (q, μ) pair satisfies the training constraint. The centralized solution is identified by a constrained optimization: the social planner selects the probability q with which each client independently joins the seniors' queue to maximize clients' expected payoff, subject to the training constraint.

In both discretionary and centralized allocations, when each client is directed to seniors with probability q, the arrival of clients at the senior queue follows a Poisson process with arrival rate $q\lambda$. In the limited-monitoring case, the *utilization* of senior workers, given by $\frac{q\lambda}{\mu}$, is always in $(0,1)$. Otherwise, with 0 utilization, allocating clients to senior workers would be strictly superior due to no wait; with utilization weakly greater than 1, allocating clients to seniors would be strictly inferior due to excessively long waits.

Since both arrival and service at the seniors' queue follow a Poisson process, the setup corresponds to what is often termed an M/M/1 queue in the queueing literature. As we detail in the Online Appendix' primer, the average waiting time in the queue, conditional on entry, is

$$\mathbf{E}[W] = \frac{1}{\mu - q\lambda} - \frac{1}{\mu} = \frac{q\lambda}{\mu(\mu - q\lambda)}. \tag{7}$$

Intuitively, as the mass of seniors grows, their service becomes more rapid and expected wait time declines. On the other hand, as the arrival rate $q\lambda$ of clients in the seniors' queue grows,

the expected wait time increases.[24] This, together with the training constraint $\mu = f((1-q)\lambda)$, determine an implicit trade-off between quality provided, determined by q, and the average wait, $\mathbf{E}[W]$. Intuitively, as q increases, there are two effects on wait times: more clients are sent to seniors, which tends to increase the wait for senior service, and fewer providers are trained, which reduces μ and therefore increases the wait further. The marginal rate of substitution between quality and expected wait depends on the training technology: the flatter the technology, the more sacrifices in terms of quality are needed to decrease wait by a small amount.

The training constraint yields the feasible set of (q, μ) pairs:

$$C \equiv \{(q, \mu) \mid q \in (0, \mu/\lambda) \text{ and } \mu = f((1-q)\lambda)\}.$$

The planner seeks to maximize the average client's utility, with the objective:

$$\max_{(q,\mu)\in C} q(h - c\mathbf{E}[W]) + (1-q)l.$$

Proposition 3 (Limited Monitoring)

1. *In the discretionary setting, the unique equilibrium is governed by* (q_L^e, μ_L^e) *that solves:*
$$\theta = \frac{\lambda q}{\mu(\mu - q\lambda)} \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{8}$$
2. *In the centralized setting, the planner has a unique optimal policy governed by* (q_L^*, μ_L^*) *that solves:*
$$\theta = \frac{q\lambda}{\mu} \frac{2\mu - q\lambda}{(\mu - q\lambda)^2} \left(\frac{q\lambda}{\mu} f' + 1\right) \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{9}$$

For the discretionary setting, in equilibrium, q must be set so that each client is indifferent between the two service options. Thus, we have:

$$\theta = \frac{h-l}{c} = \mathbf{E}[W].$$

The proposition's claim then follows directly from the formula for $\mathbf{E}[W]$.

To see the intuition for the centralized solution, notice that the objective of the planner can equivalently be written as $q\theta - q\mathbf{E}[W]$. In the proof of Proposition 3, we show that the first-order

[24] The first term in the expression for $\mathbf{E}[W]$ captures a geometric wait time when service occurs at a rate of μ, while clients arrive at a rate of $q\lambda$. This term includes a client's own service time. Since in our setting clients' wait times do not incorporate their own service, which occurs at a rate of μ, we deduct $1/\mu$ in this formulation.

approach is valid for optimizing this objective. At the optimum, we then have $\theta = \frac{d(q\mathbf{E}[W])}{dq}$, which translates into:

$$\lambda(h-l) = c\left(\lambda\mathbf{E}[W] + (q\lambda)\left(\frac{\partial\mathbf{E}[W]}{\partial q} + \left|\frac{\partial\mathbf{E}[W]}{\partial\mu}\right|\left|\frac{d\mu}{dq}\right|\right)\right). \tag{10}$$

Indeed, consider an infinitesimal increase in q. The benefit in terms of service quality is $\lambda(h-l)$, the left-hand side of this condition. The right-hand side corresponds to the overall costs of waiting. The first term, $\lambda\mathbf{E}[W]$, captures the additional wait experienced by clients diverted from juniors to seniors. The remaining terms capture the negative externality on other clients directed at seniors. There is $q\lambda$ inflow of such clients. Additional waiting results from (i) more clients occupying seniors, corresponding to $\frac{\partial\mathbf{E}[W]}{\partial q}$; and (ii) fewer trained providers corresponding to $\left|\frac{\partial\mathbf{E}[W]}{\partial\mu}\right|\left|\frac{d\mu}{dq}\right| = \left|\frac{\partial\mathbf{E}[W]}{\partial\mu}\right|(\lambda f')$. Since $\mathbf{E}[W]$ can be expressed analytically as a function of μ, q, and λ, simple calculus generates the characterization appearing in the proposition.

By Little's formula, $\mathbf{E}[Q] = \lambda q\mathbf{E}[W]$. We can therefore write the first-order condition as $\lambda\theta = \frac{d(\mathbf{E}[Q])}{dq}$, which is the formulation we used in Proposition 1 for the perfect-monitoring case.

4.2 Comparative Statics and Welfare Comparisons

We now turn to some comparative statics resulting from our characterization of the limited-monitoring case. As in the perfect-monitoring case, it is immediate to see that, as θ grows, either through an increase in the relative benefit $h-l$ of service by seniors, or through a decrease in waiting costs c, queueing for senior service becomes relatively more attractive. Consequently, under both the centralized and discretionary settings, the fraction q of clients seeking senior service increases, the mass of seniors decreases, and expected wait times increase.

As in the perfect-monitoring setting, when waiting costs are arbitrarily large, there is no delay. In this case, in both the centralized and discretionary settings, waiting is minimized and clients choose junior service. This yields $q_L^e, q_L^* \to 0$ and $\mu_L^e, \mu_L^* \to f(\lambda)$. In contrast, as c approaches 0, clients naturally seek senior service more. Note, however, that as q increases, the rate of arrivals to the queue, $q\lambda$, increases, the mass of seniors $\mu = f((1-q)\lambda)$ decreases, and $\mathbf{E}[W]$ grows arbitrarily large. As such, the solution $\overline{q}$ of $q\lambda = f((1-q)\lambda)$ is an upper bound on q. In fact, $q_L^e, q_L^* \to \overline{q}$ and $\mu_L^e, \mu_L^* \to f((1-\overline{q})\lambda)$.

4.2.1 Clients' Arrival Rate and Training Technology

We now turn to the impacts of changes in arrival rates and the training technology on outcomes in our limited-monitoring settings. In general, an increase in clients' arrival rate λ always leads to an increase in the mass of seniors and could lead to either an increase or a decrease in the fraction of clients served by seniors in both discretionary and centralized settings. Similarly, changes in training technology have an ambiguous impact when considered generally. Nonetheless, the case of linear training technology yields clear comparative statics with respect to both the arrival rate λ and the efficacy of training.[25]

Proposition 4 (Limited Monitoring – Comparative Statics) *Suppose $f(x) = ax$, for some $a > 0$. Increases in both λ and a are associated with increases in q_L^e,q_L^*,μ_L^e, and μ_L^*. Furthermore, $E[W_L^e]$ is constant in both λ and a, while $E[W_L^*]$ decreases in both λ and a.*

The consequences of changes in arrival rates or training efficacy can be intuitively understood as follows. Suppose clients' arrival rate is doubled, while the same fraction q of clients is served by seniors. For a linear training technology, the fraction of seniors exactly doubles. Therefore, using our expression (7) for the average wait time in the linear case, the expected waiting time is half the original waiting time. This reduction in wait time makes senior service more desirable, and leads to an increase in the optimal fraction of clients served by seniors in both the discretionary and centralized settings.

In terms of training efficacy, consider a small improvement, namely a small increase in a. For a fixed fraction q of clients directed at senior service, the mass of seniors grows due to improved training. Thus, senior service is quicker and the marginal benefit from serving clients by seniors increases. Consequently, more clients are directed at seniors, in both the discretionary and centralized settings.

As for waiting times, in the discretionary setting, since agents' indifference condition $\mathbf{E}[W] = \theta$ does not depend on arrival rates or the training technology, as long as the effective value of being served by seniors relative to juniors is fixed, wait times remain fixed. This is in stark contrast with the perfect-monitoring case, where the indifference condition is based on the utility from entering the queue *at the equilibrium threshold*. Indeed, as Proposition 2 illustrated, with perfect monitoring, expected wait times are non-monotonic in both λ and a.

[25] We use $\mathbf{E}[W_L^e]$ and $\mathbf{E}[W_L^*]$ for equilibrium and optimal expected wait times, respectively, in the limited-monitoring case.

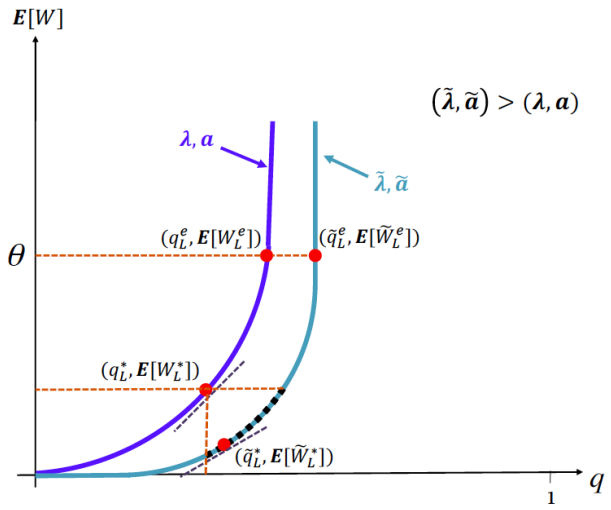

Figure 2: Impacts of changes in arrival rates or training efficacy on quality and wait times with limited monitoring

In the centralized setting, with linear training technology, we have:

$$\mathbf{E}[W] = \left(\frac{1}{a(1-q)-q} - \frac{1}{a(1-q)}\right) \cdot \frac{1}{\lambda} \equiv \frac{z(q;a)}{\lambda}.$$

The planner's optimal choice of q satisfies (9), tantamount to $\theta = \frac{d(q\mathbf{E}[W])}{dq}$, which becomes

$$\theta = \frac{z(q;a)}{\lambda} + \frac{qz'(q;a)}{\lambda}. \tag{11}$$

It is easy to verify that each term on the right-hand side of (11) increases in q (due to the convexity of $\mathbf{E}[W]$) and strictly decreases in a. Thus, if either λ or a increase, q_L^* has to increase. Moreover, since (9) can be also written as

$$\theta = \mathbf{E}[W]\left(a\lambda\mathbf{E}[W] + \frac{2}{1-q}\right), \tag{12}$$

and we established that q increases in a and λ, $\mathbf{E}[W_L^*]$ must decrease in a and λ.

Figure 2 summarizes our discussion, where $(\tilde{\lambda}, \tilde{a}) > (\lambda, a)$ implies that $\tilde{\lambda} \geq \lambda$, $\tilde{a} \geq a$, and at least one of the inequalities is strict.[26] It depicts wait times as functions of q for two values of arrival rates and training efficacies, the unique discretionary equilibrium choices—points at which expected times coincide with θ, and the resulting optimal solutions—points at which the slope of $q\mathbf{E}[W]$ is fixed at θ.

[26] The graph of $\mathbf{E}[W]$ as a function of q asymptotes at $a/(1+a)$. In particular, the asymptote changes with changes in a, but not with changes in λ.

To conclude, Proposition 4 implies that, despite the potential for additional congestion, an increase in clients' arrival rate yields unambiguously positive consequences to the organization's average performance, as quality always increases and wait times (weakly) decrease. The presence of training plays a crucial role in this result. To see this, consider an organization in which training is absent and the mass of seniors is exogenously fixed at $\overline{\mu}$—that is, $f(x) = \overline{\mu}$ for any x. It is easy to verify that any increase in λ would cause q to decrease and $\mathbf{E}[W]$ to remain unchanged in both the discretionary and centralized settings.

4.2.2 Welfare Comparison

We now turn to the impact of some parameters of our limited-monitoring case on clients' expected welfare. The average welfare per client can be written as:

$$V = q\left(h - c\mathbf{E}[W]\right) + (1-q)l = l + q(h-l) - qc\mathbf{E}[W].$$

Denote by V_L^e and V_L^* the average utility per client under the discretionary equilibrium and under the optimal policy, respectively. In the discretionary setting, since in equilibrium clients are indifferent between junior and senior service, $V_L^e = l$. In particular, the welfare gap, $V_L^* - V_L^e$, exhibits the same comparative statics as those of the welfare generated by the socially optimal protocol, V_L^*.

Suppose the value for senior service increases from h_1 to h_2, $h_2 > h_1$, while all other parameters stay fixed. The planner can certainly emulate whatever optimal policy she was following when the value from senior service was h_1. This would yield the same expected waiting costs but increase service quality. Thus, V_L^*, and thereby $V_L^* - V_L^e$, increase in h. A similar argument holds for an increase in the waiting cost c.

The impacts of an increase in arrival rates is more subtle. More rapid arrivals yield more opportunities for training, but also generate more congestion. In general, the effects of increases in λ could go either way. However, for linear training technologies, Proposition 4 indicates that wait times decrease, implying that all clients served by seniors are better off, and that the optimal fraction of clients served by seniors increases. Consequently, V_L^*, and thus $V_L^* - V_L^e$, increase in λ. Similar comparative statics follow for the training efficacy. We therefore have the following corollary.

Corollary 1 (Welfare Gap Comparative Statics) *Suppose $f(x) = ax$, for some $a > 0$. The relative welfare gain from centralization, $V_L^* - V_L^e$, is increasing in both λ and a.*

Our discussion above considers the average welfare. One may also wish to consider the volume of clients served, thereby focusing on $\lambda\left(V_L^* - V_L^e\right)$. The comparative statics of Corollary 1 would continue to hold. However, as arrival rates increase, the benefits of centralization would become even more pronounced as more clients are impacted.

This discussion implies that, with limited monitoring, organizations obtain greater advantages from centralization when the quality of senior service improves, when waiting costs decrease, or when either the arrival rate or the training technology efficacy increase.

5 The Impacts of Centralization

We can now compare the impacts of centralization on outcomes for each of our monitoring scenarios. Intuitively, for both the limited- and the perfect-monitoring settings, since fewer clients are served by seniors in the centralized setting, the average quality of service each client faces is lower. This implies shorter wait times that generate higher overall welfare.

Corollary 2 (Impacts of Centralization) $q_X^* \leq q_X^e$ and $\mu_X^* \geq \mu_X^e$ for $X = L, P$. *These inequalities are strict whenever* $X = L$ *or* $k_P^* > 1$. *In particular, there is more training, a greater mass of seniors, lower average quality, and a lower wait in centralized relative to discretionary settings.*

Technically, regardless of the monitoring level, the feasibility constraint takes the same form for the centralized and discretionary settings. The corollary's proof then stems from a comparison of the optimization constraints that govern each of the solutions.

The inverse link between service quality and wait times is the consequence of two externalities at play. One pertains to training—clients who select the queue for senior service forgo the training opportunities for juniors. The second pertains to the added wait times imposed on others selecting the seniors' queue. Both these externalities push clients to seek senior service more than is optimal, thereby generating fewer seniors and longer wait times than ideal.

6 The Impacts of Monitoring

We now turn to a comparison of outcomes with and without monitoring. Intuitively, monitoring allows clients, or the planner, to condition the decision to seek senior service on the length of the queue. In this section, we show that this increases welfare and yields greater welfare in both

the discretionary and centralized settings. We also identify how monitoring affects outcomes in terms of quality, training, and wait times.

Perfect monitoring enables either clients or the planner to condition entry to the queue on its current length, which allows for lower expected wait times even when the *same* fraction of clients receives senior service. This is captured by the following lemma.

Lemma 4 (Impacts of Monitoring on Wait Times) *For any choice of* (q, μ),

$$\mathbf{E}[W_P] < \mathbf{E}[W_L].$$

Lemma 4 suggests that, when expected wait times are similar with limited or with perfect monitoring, the quality of service afforded by perfect monitoring is higher.

6.1 Discretionary Settings

As it turns out, the indifference conditions governing the discretionary equilibria under both limited and perfect monitoring exhibit a single-crossing property. As we demonstrate, this feature implies that training can go up or down with improved monitoring, depending on training efficacy. Nonetheless, we show that allowing agents to monitor the state of the queue before deciding what service to seek always increases clients' expected welfare in equilibrium.

Formally, consider the indifference graphs, representing the mass of seniors as a function of the share of clients seeking senior service, for limited and perfect monitoring:

$$G_L \equiv \left\{ (q, \mu) : \theta = \frac{q\lambda}{\mu(\mu - q\lambda)} \right\}, \text{ and}$$
$$G_P \equiv \left\{ (q, \mu) : \theta = \frac{k(q, \mu; \lambda) - 1}{\mu} \right\},$$

respectively. Since both graphs are upward-sloping, we say that G_P strictly single crosses G_L from below if there exists a unique $(q', \mu') \in G_L \cap G_P$ such that, if $(q''_P, \mu'') \in G_P$ and $(q''_L, \mu'') \in G_L$ with $\mu'' \neq \mu'$, then either $\mu'' < \mu'$ and $q''_L < q''_P$, or $\mu' < \mu''$ and $q''_P < q''_L$.

Proposition 5A (Impacts of Monitoring in Discretionary Settings) *G_P strictly single crosses G_L from below. Furthermore, welfare is greater when monitoring is perfect.*

Proposition 5A, the intuition for which we soon describe, implies that the comparison of q^e_L and q^e_P depends on the training technology. Consider the linear training characterized by $f(x) =$

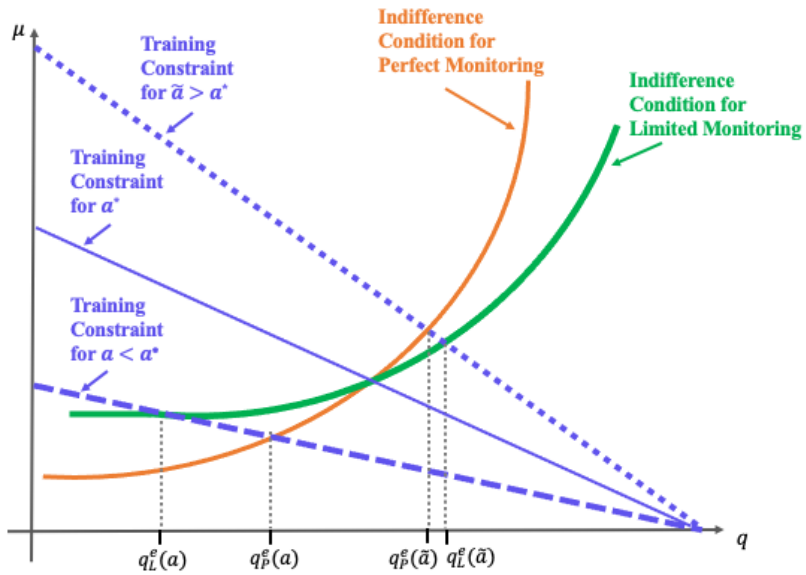

Figure 3: Impacts of monitoring on service quality and training

ax for $a > 0$. Single crossing of the indifference curves under limited and perfect monitoring indicates that the ranking of equilibrium training under limited and perfect monitoring depends on the training efficacy, as depicted in Figure 3. For sufficiently low parameters a, more clients seek senior service under perfect monitoring, which therefore yields fewer trained seniors. This pattern is reversed when the efficacy parameter a is sufficiently high. Figure 3 illustrates the threshold a^* at which the impact of monitoring reverses. We therefore have the following.

Corollary 3 (Training Efficacy and Monitoring in Discretionary Settings) *There exists $a^* > 0$ such that if $f(x) = ax$, then for $0 < a < a^*$, $q_L^e < q_P^e$ and $\mu_L^e > \mu_P^e$, while for $a > a^*$, $q_L^e > q_P^e$ and $\mu_L^e < \mu_P^e$.*

A similar conclusion to that of Corollary 3 can be derived for other classes of training technologies that dominate one another. If $f(\cdot)$ and $g(\cdot)$ are two training technologies such that $f(x) > g(x)$ for all x, then whenever perfect monitoring generates higher service quality under f, it also does so under g. Similarly, whenever perfect monitoring generates lower service quality under g, it also does so under f.[27]

The corollary suggests that the design of monitoring policies should be sensitive to the efficacy of the training technology in place—the impact on the fraction of clients receiving senior service crucially depends on training effectiveness.

[27] A similar corollary would hold for more general training technologies of the form $f(x) = ag(x)$, with a threshold parameter a^* determining the impacts of monitoring on service quality.

To gain some intuition for Proposition 5A and Corollary 3, suppose that the number of clients waiting in the queue is exogenous. For simplicity, suppose that at any point in time, with probability p, no one is waiting, and with the remaining probability $1-p$, one client is waiting. Consider first the case in which the training technology is very inefficient so that senior service is sufficiently slow that any waiting, even if for one other client, is not worthwhile. Under limited monitoring, if p is low enough, joining the senior queue is excessively risky. Consequently, all clients would seek junior service. In contrast, under perfect monitoring, regardless of p, clients who arrive when no one is in the seniors' queue would join it. In particular, more clients seek senior service under perfect monitoring. In contrast, suppose technology is fairly efficient, so that waiting for one person to be served by seniors yields a payoff that is lower than that generated by immediate junior service, but by a very small margin. For sufficiently high p, the expected value of senior service, accounting for the potential waiting costs, would still be higher than that generated by junior service. Thus, under limited monitoring, all clients would approach the senior queue. In contrast, with perfect monitoring, clients who arrive when someone is waiting for senior service would select junior service instead. The comparison then reverses and perfect monitoring generates *fewer* clients being served by seniors.

Why is welfare greater when monitoring is perfect? In the limited-monitoring environment, the *expected* wait time is such that clients are indifferent between the two service types. As already mentioned, this implies that the expected welfare is l. In contrast, with perfect monitoring, it is the clients who wait *maximally* in the queue that are indifferent. Those clients achieve an expected payoff of l. So do clients who arrive when the senior queue is so long that junior service is sought. However, clients who arrive at a shorter senior queue accomplish a higher expected payoff. For instance, clients who arrive at an empty senior queue enjoy immediate senior service and receive a payoff of h. Overall, then, the resulting expected welfare must be strictly higher than l.

6.2 Centralized Settings

For centralized settings, we continue to assume a linear production technology. As we show, monitoring impacts are conclusive in this case and independent of the training efficacy.

Proposition 5B (Impacts of Monitoring in Centralized Settings) *Welfare is greater when monitoring is perfect. Furthermore, suppose $f(x) = ax$ for some $a > 0$. Then, $q_L^* < q_P^*$ and $\mu_L^* > \mu_P^*$.*

Intuitively, when monitoring is perfect, the planner can always implement a threshold that emulates the same fraction of clients directed at seniors as in the limited-monitoring case. While the expected quality of service would then coincide with that achieved under limited monitoring, from Lemma 4, the expected wait time would be lower. In particular, the resulting welfare, even under this potentially sub-optimal policy, is greater when monitoring is perfect. With limited monitoring, the planner equates the marginal benefit of directing more clients to senior service with the marginal costs in terms of wait times—due to both reduced training and increased volume of clients waiting in queue. In the perfect-monitoring case, when using that same policy, the marginal benefit of directing more clients to senior service coincides with that in the limited-monitoring case. However, the marginal cost is lower. That last point requires proof and follows from the fact that perfect monitoring allows for "more efficient" addition of clients to the senior queue. Specifically, there is a bound on how long each client is allowed to wait, which leads to a smaller expected cost of adding clients to the senior queue.

The analysis here suggests that improving monitoring of the senior queue is always beneficial to clients. However, in centralized settings, the increase in welfare comes at the cost of training, as fewer seniors are available under perfect monitoring. If a planner's objective balances concerns about clients' welfare and the distribution of expertise among service providers—due to a concern about wage inequality, fragility of the system, etc.— the design problem could naturally become more nuanced.

7 Conclusions

We study a dynamic task-allocation setting and explore the trade-off between service quality and wait times in organizations. In our environment, junior providers need experience to improve their future service, so service providers' characteristics are endogenous. We characterize the equilibrium outcomes in discretionary settings and the social planner's optimal policy, with perfect and limited monitoring of the seniors' queue.

Externalities imply that when clients choose whom to seek service from, the average service quality is inefficiently high and queues are too long. Several insights follow from our analysis. First, as clients' arrival rate or the training technology's efficacy increase, both service quality and the scope of training increase. Second, when designing general allocation guidelines, the welfare gains from centralization are greater for larger institutions, better training technologies, and lower waiting costs. Finally, we evaluate the impacts of monitoring, giving decision-makers

the ability to condition decisions on the state of the seniors' queue. We show that improved monitoring always increases welfare, but can decrease training. Methodologically, our framework provides a tractable dynamic matching model in which agents' types are endogenous. It also illustrates a set of techniques, grounded in tools developed within queueing theory, which can be employed to study the link between quality and wait times in organizations.

There are several directions in which our study can be extended. Throughout the paper, we focus on settings in which the planner's objective pertains only to clients' welfare. However, many organizations may be concerned with training per se and aim at different objectives, which would be useful to analyze.

Our analysis pertains to only two levels of seniority: our providers are either juniors or seniors. While this simplification allows us to identify a rich set of comparisons, it would be interesting to explore the impacts of finer gradations of evolving "status" in organizations.

We focus on environments in which pricing cannot be utilized: hospitals cannot discriminate patients, administrative tasks assigned to internal staff are rarely priced. We stress that simple pricing mechanisms may not yield the centralized solutions. To see that, consider the limited-monitoring setting and suppose that access to juniors comes at a cost of p_j while access to seniors comes at a cost of p_s. Such a setting would be equivalent to our baseline setting with a value from junior service of $l' = l - p_j$ and a value from senior service of $h' = h - p_s$. Equilibrium indifference conditions would then imply an expected welfare of $l' \leq l$ regardless of prices. A more elaborate analysis of dynamic pricing mechanisms would be an interesting angle for future research. We hope the tools we introduce are useful for inspecting such environments.

8 Appendix

8.1 Proofs for Perfect Monitoring

Throughout, we let $\phi \equiv \frac{\lambda}{\mu}$. When λ is fixed, ϕ and μ uniquely determine one another. For linear training technology, where $f(x) = ax$ with $a > 0$, we have $\mu = a(1-q)\lambda$ and $\phi = \frac{1}{a(1-q)}$. Lemma 1 follows from fundamental queueing theory results described in our Online Appendix.

Proof of Lemma 2: We omit a trivial proof for the case of $\phi = 1$. If $\phi \neq 1$, then $1 - p_k = 1 - \frac{1-\phi}{\phi^{-k}-\phi}$. If either $\phi > 1$ or $\phi < 1$, the expression $1 - p_k$ is strictly increasing in k. When $k = 1$, $1 - p_k = 1 - \frac{\phi}{1+\phi} = \frac{1}{1+\phi}$, and $\lim_{k\to\infty} \frac{1-\phi^k}{1-\phi^{k+1}} = \lim_{k\to\infty} \frac{k\phi^{k-1}}{(k+1)\phi^k} = \frac{1}{\phi}$. ■

Proof of Lemma 3: We take the expression for $\mathbf{E}[Q]$ in (2) and apply (3) and (4). If $\phi = 1$

(i.e., $q = 1 - \frac{1}{a}$), then

$$\mathbf{E}[Q] = \frac{k(k-1)}{2(k+1)} = \frac{q(2q-1)}{2(1-q)}.$$

If $\phi \neq 1$, then $q = 1 - p_k = \frac{1-\phi^k}{1-\phi^{k+1}}$ and $\frac{(1-q)\phi}{1-\phi} = \frac{p_k\phi}{1-\phi} = \frac{\phi^{k+1}}{1-\phi^{k+1}}$. Thus,

$$\begin{aligned}\mathbf{E}[Q] &= \frac{1}{1-\phi^{k+1}}\left(\frac{\phi^2 - \phi^{k+1}}{1-\phi} - (k-1)\phi^{k+1}\right) \\ &= \frac{q\phi^2}{1-\phi} - \frac{(1-q)\phi}{1-\phi}\left(\frac{\log(1-q) - \log(1-q\phi)}{\log\phi}\right).\end{aligned}$$

■

Proof of Proposition 1:

1. We ignore the integer constraint on k and find a solution $(q, \mu) \in G_1 \cap G_2$, where

$$G_1 \equiv \{(q,\mu) : k(q,\mu;\lambda) = \mu\theta + 1\} \quad \text{and} \quad G_2 \equiv \{(q,\mu) : \mu = f((1-q)\lambda)\}.$$

Consider the graph G_1. The function $k(q,\mu;\lambda)$ is continuous in q and μ, and strictly increasing in q, see Lemma 2 and (4). Also,

$$sgn\left(\frac{\partial k(q,\mu;\lambda)}{\partial \mu}\right) = -sgn\left[\frac{q}{1-q\phi}\log\phi - \frac{1}{\phi}\log\left(\frac{1-q}{1-q\phi}\right)\right].$$

Since $\frac{x-1}{x} < \log x < x - 1$, for any $x \neq 1$,

$$\frac{q}{1-q\phi}\log\phi - \frac{1}{\phi}\log\left(\frac{1-q}{1-q\phi}\right) > \frac{q}{1-q\phi}\frac{\phi-1}{\phi} - \frac{1}{\phi}\left(\frac{1-q}{1-q\phi} - 1\right) = 0.$$

Thus, $\frac{\partial k(q,\mu;\lambda)}{\partial \mu} < 0$ for every $\phi \neq 1$. That is, $k(q,\mu;\lambda)$ is strictly decreasing in μ. Therefore, the graph G_1 is continuous and upward sloping: as q increases from 0 to 1, μ increases from $-\frac{1}{\theta}$ to ∞.

The graph G_2 is continuous and downward sloping: as q increases from 0 to 1, μ decreases from $f(\lambda)$ to $f(0)$.

It follows that G_1 and G_2 cross each other once, at $q_P^e \in (0,1)$ and $\mu_P^e \in (f(0), f(\lambda))$.

2. The planner's problem is

$$[P] \quad \max_{q \in [\underline{q}, \frac{a}{1+a})} q\lambda\theta - \mathbf{E}[Q]$$

where

$$\mathbf{E}[Q] = \begin{cases} \frac{q(2q-1)}{2(1-q)}, & \text{if } q = 1 - \frac{1}{a}, \\ \frac{q\phi^2}{1-\phi} - \frac{(1-q)\phi}{1-\phi}\left(\frac{\log(1-q)-\log(1-\phi q)}{\log\phi}\right) & \text{otherwise,} \end{cases}$$

and $\phi = \frac{1}{a(1-q)}$. Given the one-to-one relation between q and ϕ while λ is held fixed, the planner's problem $[P]$ is equivalent to

$$[P'] \qquad \max_{\phi \in [\underline{\phi}, 1+1/a)} \left(1 - \frac{1}{a\phi}\right) \lambda\theta - \mathbf{E}[Q]$$

where

$$\mathbf{E}[Q] = \begin{cases} \frac{(a-1)(a-2)}{2a}, & \text{if } \phi = 1, \\ \frac{1}{a} + \frac{1}{1-\phi}\left(\phi^2 + \frac{\log(a(1-\phi)+1)}{a\log\phi}\right) & \text{otherwise.} \end{cases}$$

The lower bound on ϕ corresponds to the choice of $k = 1$. Then, from $q = 1 - p_k = \frac{1}{1+\phi}$ and the linear constraint $\phi = \frac{1}{a(1-q)}$, we can obtain $\underline{\phi}$ as the unique (positive) solution of $\phi = \frac{1+\phi}{a\phi}$, or equivalently of $a\phi^2 - \phi - 1 = 0$. Namely, $\underline{\phi} = \frac{1+\sqrt{1+4a}}{2a}$.

The following technical Lemma, whose proof appears in the Online Appendix, ensures that the objective in $[P']$ is continuously differentiable and strictly concave in the single choice variable ϕ.

Lemma A1 *$E[Q]$ is strictly convex and continuously differentiable.*

Lemma A1 guarantees that the objective of the original problem $[P]$ is continuously differentiable and single-peaked in the single choice variable q, which concludes the proof.■

Proof of Proposition 2:

In Part I of the proof, we address the comparative statics of the threshold, k_P, the training, μ_P, and the quality, q_P, in both the discretionary and centralized settings with respect to changes in λ and a. In Part II, we address the expected waiting times, $\mathbf{E}[W_P]$, in both the discretionary and centralized settings with respect to changes in λ and a.

Part I: Impact of changes in λ and a on k_P, μ_P, and q_P.

First, consider the discretionary setting. When $f(x) = ax$, $a > 0$, the equilibrium (q_P^e, μ_P^e) is identified as the intersection of two graphs:

$$G_1 = \{(q, \mu) : k(q, \mu; \lambda) = \mu\theta + 1\} \quad \text{and} \quad G_2 = \{(q, \mu) : \mu = a(1-q)\lambda\}.$$

In the proof of Proposition 1, we showed that G_1 is upward sloped and G_2 is downward sloped. Since the threshold $k(q, \mu, \lambda)$ is strictly decreasing in μ and strictly increasing in q and λ, the graph G_1 shifts to the left when λ increases, and remains unchanged when a increases. The graph G_2 shifts to the right if either λ or a increase. Therefore, μ_P^e and q_p^e are increasing in a. Thus, from the indifference condition, k_P^e increases as well. Similarly, k_P^e and μ_P^e increase in λ. To show that q_p^e increases in λ, observe that $\phi \equiv \frac{\lambda}{\mu} = \frac{1}{a(1-q)}$ and Lemma 2 implies

$$k = \frac{\log(1-q) - \log(1-q\phi)}{\log \phi} = -\frac{\log(a(1-\phi)+1)}{\log \phi} - 1, \tag{13}$$

for $\phi \neq 1$, and $k = a - 1$ for $\phi = 1$, which is the limit of (13) as ϕ approaches 1. Then, by the indifference condition,

$$\begin{aligned} \frac{k-1}{\mu} = \theta &\iff \phi(k-1) = \lambda\theta \\ &\iff \phi\left(-\frac{\log(a(1-\phi)+1)}{\log \phi} - 1\right) = \lambda\theta. \end{aligned} \tag{14}$$

The left-hand side of the last equality strictly increases in ϕ. Hence, the solution ϕ_P^e increases in λ. Last, $\phi_P^e = \frac{1}{a(1-q_P^e)}$ implies that q_P^e also increases in λ.

Next, consider the centralized setting. We focus on the space of (q, ϕ), where $\phi \equiv \frac{\lambda}{\mu} = \frac{1}{a(1-q)}$ does not depend on λ. From the proof of Proposition 1, recall that $\mathbf{E}[Q]$ in $[P']$ is continuously differentiable at $\phi = 1$, and ϕ is restricted to be in $[\underline{\phi}, 1 + 1/a)$. It is straightforward to show that the first-order condition corresponding to an interior optimal solution is $\frac{d\mathbf{E}[Q]}{d\phi} = \frac{\lambda\theta}{a\phi^2}$.

Since $\mathbf{E}[Q]$ is a convex function of ϕ, the derivative $\frac{d\mathbf{E}[Q]}{d\phi}$ increases in ϕ. Thus, if λ increases, ϕ_P^* increases. Similarly, if θ increases, $\mu_P^* = \lambda/\phi_P^*$ decreases. To see the impact of increases in λ on the mass of seniors and fraction of clients served by them, rewrite the first-order condition above as $\left(\frac{d\mathbf{E}[Q]}{d\phi}\right)\phi = \frac{\lambda\theta}{a\phi} = \frac{\mu\theta}{a}$. Since ϕ_P^* increases, the left-hand side of the equality increases, which implies that μ_P^* increases. The fraction q_P^* increases as well because of the training constraint $\phi_P^* = \frac{1}{a(1-q_P^*)}$.

Last, we show that ϕ_P^* decreases in a, implying that μ_P^* increases in a. Consider any a such that $\phi_P^* \neq 1$ is an interior solution. The optimal ϕ_P^* satisfies the first-order condition:

$$\begin{aligned} &\frac{\lambda\theta}{a\phi^2} - \frac{\phi(2-\phi)}{(1-\phi)^2} - \frac{\log(a(1-\phi)+1))(-\log\phi + (1/\phi) - 1)}{a(1-\phi)^2(\log\phi)^2} = 0 \\ &\iff w(\phi; a) \equiv \frac{\lambda\theta}{\phi^2} - \frac{a\phi(2-\phi)}{(1-\phi)^2} - \frac{\log(a(1-\phi)+1))(-\log\phi + (1/\phi) - 1)}{(1-\phi)^2(\log\phi)^2} = 0. \end{aligned} \tag{15}$$

By the Implicit Function Theorem, $\frac{d\phi}{da} = -\frac{dw/da}{dw/d\phi}$. Also, we showed in the proof of Proposition 1 that the objective function of $[P']$ is strictly concave in ϕ. That is, $\frac{dw(\phi,a)}{d\phi} < 0$ at every $\phi \in (\underline{\phi}, 1) \cup (1, 1 + 1/a)$. Hence, to complete the proof of Part I of Proposition 2, it suffices to show that, for any (ϕ, a) such that $\phi \in (\underline{\phi}, 1) \cup (1, 1 + 1/a)$, we have $\frac{dw(\phi,a)}{da} < 0$.

Observe that

$$\phi \geq \underline{\phi} = \frac{1 + \sqrt{1 + 4a}}{2a} \iff (2a\phi - 1)^2 \geq 1 + 4a \iff a \geq \frac{1 + \phi}{\phi^2}.$$

From (15), we get

$$\begin{aligned}\frac{dw(\phi, a)}{da} =& -\frac{\phi(2-\phi)}{(1-\phi)^2} + \frac{1}{(1-\phi)(\log\phi)(a(1-\phi)+1)} \\ & - \frac{a}{(\log\phi)(a(1-\phi)+1)^2} + \frac{-(\log\phi) + (1/\phi) - 1}{(1-\phi)^2(\log\phi)^2}\frac{1-\phi}{a(1-\phi)+1} \\ =& -\frac{\phi(2-\phi)}{(1-\phi)^2} - \frac{a}{(\log\phi)(a(1-\phi)+1)^2} + \frac{1}{\phi(\log\phi)^2(a(1-\phi)+1)}.\end{aligned}$$

To show that $\frac{dw(\phi,a)}{da} < 0$, we distinguish between three cases. First, if $\phi \geq 2$, we multiply $\frac{dw(\phi,a)}{da}$ by $-(1-\phi)(\log\phi)(a(1-\phi)+1) > 0$, and obtain

$$\begin{aligned}& -(1-\phi)(\log\phi)(a(1-\phi)+1)\frac{dw}{da} \\ &= a\phi(2-\phi)(\log\phi) + \frac{\phi(2-\phi)(\log\phi)}{1-\phi} + \frac{a(1-\phi)}{a(1-\phi)+1} - \frac{1-\phi}{\phi(\log\phi)},\end{aligned}$$

which is strictly decreasing in a. Hence, we obtain an upper bound of the above expression by substituting a with its lower bound $\frac{1+\phi}{\phi^2}$. The upper bound, which is a function of ϕ only, is less than -2.278 for every $\phi \geq 2$.

If $1 < \phi < 2$, we have

$$\begin{aligned}(a(1-\phi)+1)^2\frac{dw}{da} &= -\frac{\phi(2-\phi)(a(1-\phi)+1)^2}{(1-\phi)^2} - \frac{a}{\log\phi} + \frac{a(1-\phi)+1}{\phi(\log\phi)^2} \\ &= -a^2\phi(2-\phi) - a\left(\frac{2\phi(1-\phi)}{1-\phi} + \frac{1}{\log\phi} - \frac{1-\phi}{\phi(\log\phi)^2}\right) - \frac{\phi(2-\phi)}{(1-\phi)^2} + \frac{1}{\phi(\log\phi)^2}.\end{aligned}$$

For any $1 < \phi < 2$, $\frac{2\phi(1-\phi)}{1-\phi} + \frac{1}{\log\phi} - \frac{1-\phi}{\phi(\log\phi)^2} > 2.435$, so the above expression is strictly decreasing in a. Substituting a with its lower bound $\frac{1+\phi}{\phi^2}$ results in an upper bound that is a function of ϕ only, and is lower than -0.82.

Finally, if $\phi < 1$, we have

$$\frac{(a(1-\phi)+1)^2}{a}\frac{dw}{da} = -\frac{\phi(2-\phi)}{(1-\phi)^2}\left(a(1-\phi)^2+2(1-\phi)+\frac{1}{a}\right) - \frac{1}{\log\phi} + \frac{1}{\phi(\log\phi)^2}\left(1-\phi+\frac{1}{a}\right)$$
$$= -a\phi(2-\phi) + \frac{1}{a}\left(\frac{1}{\phi(\log\phi)^2} - \frac{\phi(2-\phi)}{(1-\phi)^2}\right) - \frac{2\phi(2-\phi)}{1-\phi} - \frac{1}{\log\phi} + \frac{1-\phi}{\phi(\log\phi)^2}.$$

For any $\phi < 1$, $\frac{1}{\phi(\log\phi)^2} - \frac{\phi(2-\phi)}{(1-\phi)^2} > \frac{13}{12}$, so the above expression is strictly decreasing in a. Substituting a with its lower bound $\frac{1+\phi}{\phi^2}$ results in an upper bound that is a function of ϕ, and less than $-\frac{47}{24}$. This concludes the proof of Part I.

Part II: Impact of changes in λ and a on expected wait times.

Consider the discretionary setting. If a client joins the senior queue in state j (when j other clients are waiting or being served), her expected wait time is $\frac{j}{\mu}$. The probability for a client to join the senior queue in state j conditional on joining is $\frac{p_j}{1-p_k}$. Thus, the expected wait is $\mathbf{E}[W] = \sum_{j=0}^{k-1} \frac{j}{\mu}\frac{p_j}{1-p_k}$. Our formulation implies that $p_j = \frac{(1-\phi)\phi^j}{1-\phi^{k+1}}$ and $\frac{1}{1-p_k} = \frac{1-\phi^k}{1-\phi^{k+1}}$. Thus,

$$\mathbf{E}[W] = \frac{1-\phi}{\mu(1-\phi^k)}\sum_{j=0}^{k-1} j\phi^j = \frac{1-\phi}{\mu(1-\phi^k)}\frac{\phi}{1-\phi}\left(\frac{1-\phi^{k-1}}{1-\phi} - (k-1)\phi^{k-1}\right).$$

In equilibrium, $\frac{k_P^e - 1}{\mu_P^e} = \theta$. Hence,

$$\mathbf{E}[W_P^e] = \frac{1}{\mu_P^e(1-\phi_P^e)}\frac{\phi_P^e - (\phi_P^e)^{k_P^e}}{(1-(\phi_P^e)^k)} - \theta\frac{(\phi_P^e)^{k_P^e}}{1-(\phi_P^e)^{k_P^e}}.$$

We first show that $\lim_{a\to\infty}\mathbf{E}[W_P^e] = 0$. If a increases, the right hand side of (13) diverges. From (14), it follows that ϕ_P^e must decrease to 0. Since $\phi_P^e = \frac{\lambda}{\mu_P^e}$, we have $\lim_{a\to\infty}\mu_P^e = \infty$, and $\lim_{a\to\infty}\mathbf{E}[W_P^e] = 0$ from the above equality.

We now show that $\lim_{\lambda\to\infty}\mathbf{E}[W_P^e] = \theta$. Note that $\phi_P^e = \frac{\lambda}{\mu_P^e}$ is bounded by $\underline{\phi} = \frac{1+\sqrt{1+4a}}{2a}$ and $\overline{\phi} = 1 + \frac{1}{a}$. Thus, $\lim_{\lambda\to\infty}\mu_P^e = \infty$. The boundedness of ϕ_P^e also implies from (14) that $a(1-\phi_P^e)+1$ approaches 0 as λ increases indefinitely. Hence, by (14), $\lim_{\lambda\to\infty}(\phi_P^e)^{k_P^e} = \lim_{\lambda\to\infty}\frac{1}{\phi_P^e(a(1-\phi_P^e)+1)} = \infty$, which implies that $\lim_{\lambda\to\infty}\mathbf{E}[W_P^e] = \theta$.

To show that $\lim_{a\to 0}\mathbf{E}[W_P^e] = \lim_{\lambda\to 0}\mathbf{E}[W_P^e] = 0$, we observe that as a or λ approach 0, μ tends to 0 for any $q \in [0,1]$. Hence, the indifference condition yields a threshold of 1, implying that agents seek senior service only if there is no other agent in line (agents experience no waiting cost while being served). The result follows.

We focus on the centralized setting next. For $\lim_{a\to\infty}\mathbf{E}[W_P^*]$, we compare the welfare under

perfect and limited monitoring. The first-order condition with limited monitoring implies (12). Then, as a grows indefinitely, $\mathbf{E}[W_L^*]$ approaches 0, implying that q_L^* tends to the upper bound $\frac{a}{a+1} \to 1$. Thus, as a grows indefinitely, the planner can achieve the maximum average welfare of h. Since the planner can achieve higher welfare with perfect monitoring than limited monitoring (Proposition 5B, proved below), $\mathbf{E}[W_P^*]$ must vanish as well when a increases indefinitely.

For $\lim_{\lambda\to\infty} \mathbf{E}[W_P^*]$, we consider the planner's problem presented after Lemma 3:

$$[P] \max_{q\in[\underline{q},\frac{a}{1+a})} q\lambda\theta - \mathbf{E}[Q] \iff \frac{\lambda}{c}\left[\max_{q\in[\underline{q},\frac{a}{1+a})} q\left(h - \frac{\mathbf{E}[Q]}{q\lambda}c\right) + (1-q)l\right].$$

Note that $\frac{\mathbf{E}[Q]}{q\lambda} = \mathbf{E}[W]$. Under linear training, the expected number of clients waiting, $\mathbf{E}[Q]$, is independents of λ (Lemma 3). Then, for any fixed $q \in (0, \frac{a}{1+a})$, as λ increases indefinitely, the average welfare approaches $qh + (1-q)l$. Thus, the planner can achieve welfare arbitrarily close to $\frac{a}{1+a}h + \frac{1}{1+a}l$, which implies that $\lim_{\lambda\to\infty} \mathbf{E}[W_P^*] = 0$.

Finally, to see that $\lim_{a\to 0} \mathbf{E}[W_P^*] = \lim_{\lambda\to 0} \mathbf{E}[W_P^*] = 0$, observe that μ tends to 0 as a or λ approach 0, regardless of $q \in [0,1]$. When μ_P^* is sufficiently small so that $h - \frac{c}{\mu_P^*} > l$, a planner sends a client to the senior queue only when no waiting is necessary ($k_P^* = 1$). Hence, the expected wait vanishes to zero. ■

8.2 Proofs for Limited Monitoring

Proof of Proposition 3:

For part 1, for any given mass of seniors μ and $q \in (0,1)$, each client's expected payoff from entering the seniors' queue is $h - c\mathbf{E}[W]$, while it is l if served by a junior. A client is indifferent between the two options when

$$l = h - c\mathbf{E}[W] \iff \theta = \mathbf{E}[W] = \frac{\lambda q}{\mu(\mu - \lambda q)}.$$

As $\mu = f((1-q)\lambda)$ and f is differentiable, strictly increasing, and $f(0) = 0$, the right-hand side of the indifference condition is continuous and strictly increasing in q from 0 to ∞. A unique solution q_L^e of the indifference condition exists, and $\mu_L^e = f((1-q_L^e)\lambda)$ follows.

For part 2, we write the planner's problem as follows:

$$\begin{aligned} &\max_{q\in(0,\mu/\lambda)} q\,(h - c\mathbf{E}[W]) + (1-q)l \iff \max_{q\in(0,\mu/\lambda)} q\theta - \tfrac{\lambda q^2}{\mu(\mu-\lambda q)}, \\ &\text{subject to} \quad \mu = f((1-q)\lambda). \end{aligned}$$

The first term of the objective ($q\theta$) is linear in q. The second term ($-\frac{\lambda q^2}{\mu(\mu-\lambda q)}$) is a strictly quasi-concave function of (q,μ) for $q\lambda<\mu$. To see this, let $g(q,\mu)\equiv-\frac{\lambda q^2}{\mu(\mu-\lambda q)}=-\frac{\lambda}{\frac{\mu}{q}(\frac{\mu}{q}-\lambda)}$ and observe that the function $-\frac{\lambda}{x(x-\lambda)}$ is increasing in x if $x>\lambda$. Consider (q_1,μ_1) and (q_2,μ_2) such that $\lambda<\frac{\mu_1}{q_1}\leq\frac{\mu_2}{q_2}$. For any $\gamma\in[0,1]$, we have $\frac{\mu_1}{q_1}\leq\frac{\overline{\mu}}{\overline{q}}\equiv\frac{\gamma\mu_1+(1-\gamma)\mu_2}{\gamma q_1+(1-\gamma)q_2}\leq\frac{\mu_2}{q_2}$. Then, $g(\overline{q},\overline{\mu})\geq\min\{g(q_1,\mu_1),g(q_2,\mu_2)\}$, and the strict quasi-concavity of $g(q,\mu)$ follows.

The feasible set of (q,μ) is convex because f is (weakly) concave. Also, there is no corner solution. Indeed, if $q\to0$, a client joining the queue gains $\theta=\frac{h-l}{c}>0$, while the waiting time $\mathbf{E}[W]$ converges to zero; on the other hand, as q increases to $\overline{q}$, defined as the solution of $q\lambda=f((1-q)\lambda)$, the average waiting time explodes. Hence, the first-order condition is necessary and sufficient for the solution of the planner's problem:

$$\frac{\lambda\theta-\frac{q(\lambda/\mu)^2(2-q(\lambda/\mu))}{(1-q(\lambda/\mu))^2}}{\frac{(\lambda/\mu)q^2(2-q(\lambda/\mu))}{(1-q(\lambda/\mu))^2}}=\left(\frac{\lambda}{\mu}\right)^2f'\iff\theta=\frac{q\lambda}{\mu}\frac{2\mu-q\lambda}{(\mu-q\lambda)^2}\left(\frac{q\lambda}{\mu}f'+1\right).$$

Taking the training constraint into account, the right-hand side is strictly increasing in q. A unique solution q_L^* exists, and $\mu_L^*=f((1-q_L^*)\lambda)$ follows. ■

Proof of Proposition 4: Given $f(x)=ax$, the expected wait time becomes

$$\mathbf{E}[W]=\left(\frac{1}{a(1-q)-q}-\frac{1}{a(1-q)}\right)\cdot\frac{1}{\lambda}\equiv\frac{z(q;a)}{\lambda},$$

and the feasibility condition $q\lambda<\mu=f((1-q)\lambda)$ becomes $q<a(1-q)$. Observe that $z(q;a)$ is strictly increasing in q, and strictly decreasing in a because, for any $a_l<a_h$,

$$\begin{aligned}z(q;a_l)>z(q;a_h)&\iff\frac{1}{a_l(1-q)-q}-\frac{1}{a_h(1-q)-q}>\frac{1}{a_l(1-q)}-\frac{1}{a_h(1-q)}\\&\iff a_la_h(1-q)^2>(a_l(1-q)-q)(a_h(1-q)-q),\end{aligned}$$

which clearly holds, and

$$\begin{aligned}z'(q;a)&=\frac{a+1}{(a(1-q)-q)^2}-\frac{a}{(a(1-q))^2}\\&=\frac{1}{(a(1-q)-q)^2}+a\left(\frac{1}{a(1-q)-q}+\frac{1}{a(1-q)}\right)z(q;a)>0.\end{aligned}\tag{16}$$

In the discretionary setting, the equilibrium (q_L^e,μ_L^e) solves $\theta=\mathbf{E}[W]$. It is immediate to see that q_L^e and μ_L^e are increasing in λ and in a, while $\mathbf{E}[W]$ is unaffected by λ and a.

In the centralized setting, the planner's optimal choice q_L^* solves (10), which becomes

$$\theta = \frac{z(q;a)}{\lambda} + \frac{qz'(q;a)}{\lambda}. \tag{17}$$

Each term of (16) is strictly increasing in $q < a(1-q)$ and strictly decreasing in a. Thus, if either λ or a increase, q has to increase to satisfy (17).

In fact, the first-order condition (17) can be written as

$$\lambda\theta = \frac{1}{a(1-q)-q} - \frac{1}{a(1-q)} + \frac{(a+1)q}{(a(1-q)-q)^2} - \frac{aq}{(a(1-q))^2}.$$

After algebraic manipulation,

$$\begin{aligned} \lambda\theta &= \left(\frac{1}{a(1-q)-q} - \frac{1}{a(1-q)}\right)\left(\frac{a}{a(1-q)-q} + \frac{a}{a(1-q)}\right) \\ &\iff \theta = \mathbf{E}[W]\left(a\lambda\mathbf{E}[W] + \frac{2}{1-q}\right). \end{aligned}$$

We showed that q increases in a and λ. Since the left-hand side of the first-order condition is held fixed at θ, $\mathbf{E}[W]$ must decrease in a and λ. ■

8.3 Proofs for Impacts of Centralization

Proof of Corollary 2: Consider (8) and (9) in Proposition 2 for the limited-monitoring setting. Since $f' \geq 0$,

$$\frac{q\lambda}{\mu}\frac{2\mu - q\lambda}{(\mu - q\lambda)^2}\left(\frac{q\lambda}{\mu}f' + 1\right) \geq \frac{q\lambda}{\mu}\frac{2\mu - q\lambda}{(\mu - q\lambda)^2} > \frac{\lambda q}{\mu(\mu - q\lambda)}.$$

The result $q_L^* < q_L^e$ (hence $\mu_L^* > \mu_L^e$) follows.

Next, consider the perfect-monitoring setting, where the solution (q_P^*, μ_P^*) determines $\phi_P^* = \lambda/\mu_P^*$ and $k_P^* = k(q_P^*; \mu_P^*, \lambda)$ by (4). The proof of Corollary 2 is trivial if $k_P^* = 1$, which corresponds to a corner solution, i.e., the lower bound of any feasible q. Hence, assume an interior solution $k_P^* > 1$. Then, $q_P^* > \frac{1}{\phi_P^*+1}$ by (3) and the first-order condition of Proposition 1 holds. It is convenient to consider

$$\mathbf{E}[Q] = \begin{cases} \frac{q(2q-1)}{2(1-q)}, & \text{if } q = 1 - \frac{1}{a}, \\ \frac{q\phi^2}{1-\phi} - \frac{(1-q)\phi}{1-\phi}\left(\frac{\log(1-q)-\log(1-\phi q)}{\log\phi}\right) & \text{otherwise.} \end{cases}$$

as a function of (q, ϕ), with $\phi = \frac{1}{a(1-q)}$, a function of q. Then, the first-order condition (6)

corresponds to

$$\lambda\theta = \frac{\partial \mathbf{E}[Q]}{\partial q} + \frac{\partial \mathbf{E}[Q]}{\partial \phi}\frac{d\phi}{dq}.$$

An increase of $\phi \equiv \frac{\lambda}{\mu}$, while holding q, a, and λ fixed, corresponds to a decrease in μ, which increases $\mathbf{E}[Q]$. Thus, $\frac{d\phi}{dq} > 0$ and $\frac{\partial \mathbf{E}[Q]}{\partial \phi} > 0$. At the optimal solution, $\lambda\theta \geq \frac{\partial \mathbf{E}[Q]}{\partial q}$.

If $\mu_P^* = \lambda$, implying, $\phi_P^* = 1$,

$$\mu_P^*\theta = \lambda\theta \geq \frac{\partial \mathbf{E}[Q]}{\partial q} = -1 + \frac{1}{2(1-q_P^*)^2} > \frac{q_P^*}{1-q_P^*} = k_P^*,$$

where the last equality follows from (4).

Suppose $\mu_P^* \neq \lambda$, implying $\phi_P^* \neq 1$. If $\phi \neq 1$,

$$\frac{\partial \mathbf{E}[Q]}{\partial q} = \frac{\phi^2}{1-\phi} + \frac{\phi}{(1-\phi)\log\phi}\left(\log\left(\frac{1-q}{1-q\phi}\right) + \frac{1-\phi}{1-q\phi}\right),$$

and, by exploiting (4), we have

$$\begin{aligned}\frac{\partial \mathbf{E}[Q]}{\partial q} - k(q,\phi)\phi =& \frac{\phi^2}{1-\phi} + \left(\frac{1}{1-\phi} - 1\right)\frac{\phi}{\log\phi}\log\left(\frac{1-q}{1-q\phi}\right) + \frac{\phi}{(1-\phi)\log\phi}\frac{1-\phi}{1-q\phi} \\ =& \frac{\phi^2}{(1-\phi)\log\phi}\left(\log\left(\frac{\phi(1-q)}{1-q\phi}\right) + \frac{1-\phi}{\phi(1-q\phi)}\right).\end{aligned}$$

Since $\log x \leq x - 1$ for every $x \in \mathbb{R}$, for $\phi < 1$,

$$\log\left(\frac{\phi(1-q)}{1-q\phi}\right) + \frac{1-\phi}{\phi(1-q\phi)} \leq \frac{\phi(1-q)}{1-q\phi} - 1 + \frac{1-\phi}{\phi(1-q\phi)} \leq \frac{1-\phi}{1-q\phi}\left(\frac{1}{\phi} - 1\right) < 0.$$

If $\phi > 1$, we have $\lim_{\phi\to 1}\log\left(\frac{\phi(1-q)}{1-q\phi}\right) + \frac{1-\phi}{\phi(1-q\phi)} = 0$ and

$$\frac{\partial\left(\log\left(\frac{\phi(1-q)}{1-q\phi}\right) + \frac{1-\phi}{\phi(1-q\phi)}\right)}{\partial\phi} = \frac{(2q\phi-1)(1-\phi)}{\phi^2(1-q\phi)^2} < 0,$$

where the inequality is guaranteed by the fact that $q > \frac{1}{\phi+1}$ implies $2q\phi - 1 > \frac{2\phi}{\phi+1} - 1 = \frac{\phi-1}{\phi+1} < 0$. Thus, $\log\left(\frac{\phi(1-q)}{1-q\phi}\right) + \frac{1-\phi}{\phi(1-q\phi)} < 0$. Therefore,

$$\mu_P^*\theta = \frac{\lambda\theta}{\phi_P^*} \geq \left(\frac{\partial \mathbf{E}[Q]}{\partial q}\right)\frac{1}{\phi_P^*} > k_P^*.$$

Overall, we conclude that $k_P^* < \mu_P^*\theta + 1 = k_P^e$. It follows that $q_P^* < q_P^e$ and $\mu_P^* > \mu_P^e$. ■

8.4 Proofs for Impacts of Monitoring

Proof of Lemma 4: For any choice of (q, μ), the average wait times in the limited- and perfect-monitoring settings are

$$\mathbf{E}[W_L] = \frac{1}{\mu - q\lambda} - \frac{1}{\mu} = \frac{q\lambda}{\mu(\mu - q\lambda)}, \text{ and}$$

$$\mathbf{E}[W_P] = \frac{\mathbf{E}[Q_P]}{(1-p_k)\lambda} = \begin{cases} \frac{k-1}{2\lambda} & \text{if } \lambda = \mu \\ \frac{1}{\mu}\left(\frac{\lambda}{\mu-\lambda} - k\frac{\lambda^k}{\mu^k - \lambda^k}\right) & \text{if } \lambda \neq \mu \end{cases},$$

respectively, where k is given by (4).

If $\lambda = \mu$, then $\frac{\mathbf{E}[W_P]}{\mathbf{E}[W_L]} = \frac{2q-1}{2q} < 1$. If $\lambda \neq \mu$ (i.e., $\phi \neq 1$), then $\mu\mathbf{E}[W_L] = \frac{q\phi}{1-q\phi}$ and we have $\phi^k = \frac{1-q}{1-q\phi}$ from (4), which implies

$$\begin{aligned} \mu\mathbf{E}[W_P] &= \frac{\phi}{1-\phi} - \frac{1-q}{q(1-\phi)\log\phi}\log\left(\frac{1-q}{1-q\phi}\right) \\ &< \frac{\phi}{1-\phi} - \frac{1-q}{q(1-\phi)\log\phi}\left(\frac{1-q}{1-q\phi} - 1\right) = \frac{\phi}{1-\phi} + \frac{1-q}{\log\phi(1-q\phi)}, \end{aligned}$$

where the inequality follows from $(1-\phi)\log\phi < 0$ and $\log x < x - 1$ for $x \neq 1$. We have

$$\begin{aligned} \mathbf{E}[W_L] < \mathbf{E}[W_P] &\impliedby \frac{\phi}{1-\phi} + \frac{1-q}{\log\phi(1-q\phi)} < \frac{q\phi}{1-q\phi} \\ &\iff \frac{1-q}{\log\phi} < q\phi - \frac{\phi(1-q\phi)}{1-\phi} = \frac{-\phi(1-q)}{1-\phi} \\ &\iff \phi\log\phi + 1 - \phi > 0, \end{aligned}$$

and the last inequality holds for every $\phi \neq 1$. Hence, $\mathbf{E}[W_L] < \mathbf{E}[W_P]$. ■

Proof of Proposition 5A: By Propositions 1 and 3, (q_L^e, μ_L^e) solves $\theta = \frac{q\lambda}{\mu(\mu-q\lambda)}$ and $\mu = f((1-q)\lambda)$, while (q_P^e, μ_P^e) solves $k(q, \mu; \lambda) = \mu\theta + 1$ and $\mu = f((1-q)\lambda)$. The graphs

$$G_L \equiv \left\{(q,\mu) : \theta = \frac{q\lambda}{\mu(\mu - q\lambda)}\right\} \quad \text{and} \quad G_P \equiv \left\{(q,\mu) : \theta = \frac{k(q,\mu;\lambda) - 1}{\mu}\right\}$$

represent the indifference conditions under limited and perfect monitoring, respectively. It is easy to verify that both graphs are upward sloping. We show that G_P strictly single-crosses G_L from below. Formally, if $(q_L', \mu') \in G_L$, $(q_P', \mu') \in G_P$, and $q_P' \leq q_L'$, then for any $\mu'' > \mu'$ with $(q_P'', \mu'') \in G_P$ and $(q_L'', \mu'') \in G_L$, we have $q_P'' < q_L''$.

Step 1: If $\mu \leq \lambda$, G_P lies below G_L—if $(q_L, \mu) \in G_L$, and $(q_P, \mu) \in G_P$, then $q_L < q_P$.

Take any $\mu \leq \lambda$ and $q \in [0,1]$ such that $q\lambda < \mu$. If $\mu = \lambda$, then

$$\frac{q\lambda}{\mu(\mu - q\lambda)} = \frac{1}{\lambda}\frac{q}{1-q} > \frac{1}{\lambda}\left(\frac{q}{1-q} - 1\right) = \frac{k(q,\mu;\lambda) - 1}{\mu}.$$

If $\mu < \lambda$, then

$$\frac{q\lambda}{\mu(\mu - q\lambda)} > \frac{k(q,\mu;\lambda) - 1}{\mu} = \frac{1}{\mu}\left(\frac{1}{\log(\lambda/\mu)}\log\left(\frac{1-q}{1-(q\lambda/\mu)}\right) - 1\right)$$

$$\iff \quad \frac{q\lambda}{\mu - q\lambda} > \frac{1}{\log(\lambda/\mu)}\left(\frac{1-q}{1-(q\lambda/\mu)} - 1\right) - 1 \quad (\forall x \neq 1,\, \log x < x - 1)$$
$$\iff \quad \frac{(\lambda/\mu)\log(\lambda/\mu)}{(\lambda/\mu) - 1} > \frac{q\lambda}{\mu},$$

which holds since, for any $x > 1$, $\frac{x \log x}{x-1} > 1 > \frac{q\lambda}{\mu}$. Therefore, if $\mu \leq \lambda$, $(q_L, \mu) \in G_L$, and $(q_P, \mu) \in G_P$, then $q_L < q_P$, which concludes the proof of Step 1.

Step 2: If $\mu > \lambda$, G_P strictly single-crosses G_L from below.

Take any $\mu' > \lambda$ such that $(q'_L, \mu') \in G_L$ and $(q'_P, \mu') \in G_P$ for some $q'_P \leq q'_L$. Then,

$$\theta = \frac{q'_L\lambda}{\mu'(\mu' - q'_L\lambda)} \quad \text{and} \quad \theta = \frac{1}{\mu'}\left(\frac{1}{\log(\lambda/\mu')}\log\left(\frac{1-q'_P}{1-(q'_P\lambda/\mu')}\right) - 1\right).$$

Take any $\mu'' > \mu'$ and let q''_L, q''_P, be such that $(q''_P, \mu'') \in G_P$ and $(q''_L, \mu'') \in G_L$. Also, define $\overline{q}$ such that $\frac{\overline{q}}{\mu''} = \frac{q'_L}{\mu'} \equiv \delta$. We first compare $\overline{q}$ and q''_L.

$$\frac{\overline{q}\lambda}{\mu''(\mu'' - \overline{q}\lambda)} = \frac{\delta\lambda}{\mu''(1 - \delta\lambda)} < \frac{\delta\lambda}{\mu'(1 - \delta\lambda)} = \frac{q'_L\lambda}{\mu'(\mu' - q'_L\lambda)} = \theta.$$

Thus, $\overline{q} < q''_L$. We compare $\overline{q}$ and q''_P using the following auxiliary result:

Claim 4: For any $0 < y < x < 1$, the ratio $\frac{\log(1-y) - \log(1-x)}{x(\log x - \log y)}$ is strictly increasing in x.

Proof of Claim 4: For any $0 < y < x < 1$, the derivative of $\frac{\log(1-y) - \log(1-x)}{x(\log x - \log y)}$ is strictly positive if and only if

$$\frac{x}{1-x}\log\left(\frac{x}{y}\right) > \log\left(\frac{1-y}{1-x}\right)\left(\log\left(\frac{x}{y}\right) + 1\right).$$

If $x = y$, both sides of the above inequality are equal to 0. If $x \neq y$, the left-hand side is strictly increasing in x. Hence, it is sufficient to show that the derivative of the right-hand side

is strictly negative, or equivalently,

$$-\log(x/y) - 1 + \left(\frac{1}{x} - 1\right)\log\left(\frac{1-y}{1-x}\right) < 0.$$

The last inequality holds at $x = y$, and the derivative of the left-hand side is

$$-\frac{1}{x} - \frac{1}{x^2}\log\left(\frac{1-y}{1-x}\right) + \left(\frac{1}{x} - 1\right)\frac{1}{1-x} = -\frac{1}{x^2}\log\left(\frac{1-y}{1-x}\right) < 0,$$

which completes the proof of Claim 4.

To conclude the proof of Step 2, observe that $\lambda < \mu'$. The definition of $\overline{q}$, and $\overline{q} < q_L'' \leq 1$ imply that $\lambda\delta < q_L' < \overline{q} (\leq 1)$. Claim 4 implies

$$\begin{aligned}\frac{1}{\mu''\log(\lambda/\mu'')}\log\left(\frac{1-\overline{q}}{1-(\overline{q}\lambda/\mu'')}\right) &= \frac{\delta(\log(1-\lambda\delta) - \log(1-\overline{q}))}{\overline{q}(\log\overline{q} - \log(\lambda\delta))} \\ > \frac{\delta(\log(1-\lambda\delta) - \log(1-q_L'))}{q_L'(\log q' - \log(\lambda\delta))} &= \frac{1}{\mu'\log(\lambda/\mu')}\log\left(\frac{1-q_L'}{1-(q_L'\lambda/\mu')}\right),\end{aligned}$$

Hence,

$$\frac{1}{\mu''}\left(\frac{1}{\log(\lambda/\mu'')}\log\left(\frac{1-\overline{q}}{1-(\overline{q}\lambda/\mu'')}\right) - 1\right) > \frac{1}{\mu'}\left(\frac{1}{\log(\lambda/\mu')}\log\left(\frac{1-q_L'}{1-(q_L'\lambda/\mu')}\right) - 1\right) \geq \theta,$$

which implies $\overline{q} > q_P''$.[28] Therefore, $q_L'' > \overline{q} > q_P''$. ■

Proof of Proposition 5B:

Welfare is higher under perfect monitoring following Lemma 4 and the arguments described in the text. Namely, the planner can emulate the utilization of juniors in the limited-monitoring environment when there is perfect monitoring and generate lower wait times.

We compare the solutions (q_L^*, μ_L^*) and (q_P^*, μ_P^*). In the proof of Proposition 1, we wrote the planner's problem under perfect monitoring as:[29]

$$\begin{aligned}[PM] \max_{\phi\in[\frac{1}{a},1+\frac{1}{a})} &\left(1 - \frac{1}{a\phi}\right)\lambda\theta - \mathbf{E}[Q_P], \\ \text{where} \quad \mathbf{E}[Q_P] &= \begin{cases} \frac{(a-1)(a-2)}{2a}, & \text{if } \phi = 1, \\ \frac{1}{a} + \frac{1}{1-\phi}\left(\phi^2 + \frac{\log(a(1-\phi)+1)}{a\log\phi}\right) & \text{otherwise,} \end{cases}\end{aligned}$$

which is strictly convex and continuously differentiable in ϕ, including at $\phi = 1$. The planner's

[28] From our analysis before, recall that $\frac{1}{\log(\lambda/\mu)}\log\left(\frac{1-q}{1-(q\lambda/\mu)}\right)$ is strictly increasing in q.

[29] Recall that $\phi = \frac{1}{a(1-q)} \geq \frac{1}{a}$, and the steady-state constraint $q\lambda < \mu$ (i.e., $q\phi < 1$) implies that $\phi < 1 + \frac{1}{a}$.

problem under limited monitoring is

$$[LM] \quad \max_{\phi \in [\frac{1}{a}, 1+\frac{1}{a})} \left(1 - \frac{1}{a\phi}\right) \lambda\theta - \mathbf{E}[Q_L],$$

where, by Little's formula,

$$\mathbf{E}[Q_L] = q\lambda \mathbf{E}[W_L] = \frac{(q\lambda)^2}{\mu(\mu - q\lambda)} = \frac{(q\phi)^2}{1 - q\phi} = \frac{(a\phi - 1)^2}{a(a(1-\phi)+1)}.$$

The following Lemmas H and I ultimately guarantee that $\frac{d\mathbf{E}[Q_L]}{d\phi} > \frac{d\mathbf{E}[Q_P]}{d\phi}$. Since both $\mathbf{E}[Q_L]$ and $\mathbf{E}[Q_P]$ are continuously differentiable, including at $\phi = 1$, it is without loss of generality to consider $\phi \in (\frac{1}{a}, 1 + \frac{1}{a}) \backslash \{1\}$.

Lemma H *For any $\phi \neq 1$, $\frac{a(1-\phi)^2}{\log^2(a(1-\phi)+1)}$ is strictly positive and strictly decreasing in ϕ and $\frac{-1}{\log(a(1-\phi)+1)}\left(1 + \frac{1-\phi}{\log\phi}\right)$ is increasing in ϕ.*

Proof of Lemma H: Let $g(\phi) \equiv \frac{-a(1-\phi)}{\log(a(1-\phi)+1)}$, which is strictly negative for any $\phi \neq 1$. By Lemma D, we have $g'(\phi) > 0$. Thus, $\frac{(g(\phi)^2}{a} = \frac{a(1-\phi)^2}{\log^2(a(1-\phi)+1)}$ is strictly positive and strictly decreasing in $\phi \neq 1$.

Next, let $h(\phi) \equiv \frac{1}{1-\phi} + \frac{1}{\log\phi}$. For any $\phi \neq 1$, $\log\phi < \phi - 1$ and $(1-\phi)\log\phi < 0$, which imply $h(\phi) = \frac{\log\phi + 1 - \phi}{(1-\phi)\log\phi} > 0$. Note that

$$h'(\phi) < 0 \iff \frac{1}{(1-\phi)^2} < \frac{1}{\phi \log^2\phi} \iff \phi \log^2\phi < (1-\phi)^2.$$

We differentiate each side of the last inequality. The first derivatives are equal at $\phi = 1$. For every $\phi \neq 1$, since $\log\phi < \phi - 1$, the second derivative of the left-hand side is smaller than that of the right-hand side: $2(\log\phi)(1/\phi) + 2/\phi < 2$. Thus, if $\phi > 1$, we have $\log^2\phi + 2\log\phi < -2(1-\phi)$, and if $\phi < 1$, we have $\log^2\phi + 2\log\phi > -2(1-\phi)$. Therefore, we obtain $\phi\log^2\phi < (1-\phi)^2$, and $h'(\phi) < 0$. Hence, Lemma H follows from

$$\left(-\frac{1}{\log(a(1-\phi)+1)}\left(1 + \frac{1-\phi}{\log\phi}\right)\right)' = \frac{g'(\phi)h(\phi) + g(\phi)h'(\phi)}{a} > 0.$$

■

Lemma I *For any $\phi \neq 1$, $\mathbf{E}[Q_L] - \mathbf{E}[Q_P]$ is strictly increasing in ϕ.*

Proof of Lemma I: We have

$$\mathbf{E}[Q_L] - \mathbf{E}[Q_P] = \frac{(a\phi - 1)^2}{a(a(1-\phi)+1)} - \frac{1}{a} - \frac{1}{1-\phi}\left(\phi^2 + \frac{\log(a(1-\phi)+1)}{a\log\phi}\right).$$

Since

$$\frac{(a\phi-1)^2}{a(a(1-\phi)+1)} - \frac{1}{a} - \frac{\phi^2}{1-\phi} = \frac{a\phi^2-\phi-1}{a(1-\phi)+1} - \frac{\phi^2}{1-\phi} = \frac{-1}{(a(1-\phi)+1)(1-\phi)},$$

we get

$$\mathbf{E}[Q_L] - \mathbf{E}[Q_P] = \frac{-1}{(a(1-\phi)+1)(1-\phi)} - \frac{\log(a(1-\phi)+1)}{a(1-\phi)\log\phi}.$$

Therefore,

$$\left(\frac{a(1-\phi)^2}{\log^2(a(1-\phi)+1)}\right)(\mathbf{E}[Q_L] - \mathbf{E}[Q_P])$$
$$= \left(\frac{-a(1-\phi)}{(a(1-\phi)+1)\log^2(a(1-\phi)+1)} + \frac{1}{\log(a(1-\phi)+1)}\right) - \frac{1}{\log(a(1-\phi)+1)}\left(1 + \frac{1-\phi}{\log\phi}\right).$$

The right-hand side of the last equation is increasing in ϕ (see the expression $r''(x)$ in the proof of Lemma D, where we substitute $-a(1-\phi)$ for x, and Lemma H). From $\mathbf{E}[Q_L] - \mathbf{E}[Q_P] > 0$ and Lemma H, it follows that $\mathbf{E}[Q_L] - \mathbf{E}[Q_P]$ is strictly increasing in ϕ, which concludes the proof of Lemma I. ■

We are now ready to show Proposition 5B. Lemma I implies that

$$\frac{\lambda\theta}{a\phi_L^*} = \frac{d\mathbf{E}[Q_L](\phi_L^*)}{d\phi} > \frac{d\mathbf{E}[Q_P](\phi_L^*)}{d\phi}, \quad \text{and} \quad \frac{\lambda\theta}{a\phi_P^*} = \frac{d\mathbf{E}[Q_P](\phi_P^*)}{d\phi}.$$

From the strict convexity of $\mathbf{E}[Q_P]$ that we showed in the proof of Proposition 1, we obtain that $\phi_L^* < \phi_P^*$. Therefore, $q_L^* < q_P^*$. ■

9 References

Acemoglu, Daron, 1997, "Training and Innovation in an Imperfect Labour Market," Review of Economic Studies, 64(3), 445-464.

Acemoglu, Daron, and Jörn-Steffen Pischke, 1999, "The Structure of Wages and Investments in General Training," Journal of Political Economy, 107(3), 539-572.

Arnosti, Nicholas, Ramesh Johari, and Yash Kanoria, 2015, "Managing Congestion in Dynamic

Matching Markets," mimeo.

Ashlagi, Itai, Maximilien Burq, Patrick Jaillet, and Vahideh Manshadi, 2019, Operations Research, 67(4), 927-949.

Baccara, Mariagiovanna, and Leeat Yariv, 2021, "Dynamic Matching," in "Online and Matching-Based Market Design," edited by Federico Echenique, Nicole Immorlica, and Vijay V. Vazirani.

Bird Daniel, and Alexander Frug, 2021, "Optimal Contracts with Randomly Arriving Tasks," Economic Journal, 131(637), 1905-1918.

Bloch, Francis and David Cantala, 2017, "Dynamic Assignment of Objects to Queuing Agents," American Economic Journal: Microeconomics, 9, 88-122.

Bray, Robert, Decio Coviello, Andrea Ichino, and Nicola Persico, 2016, "Multitasking, Multi-Armed Bandits, and the Italian Judiciary," Manufacturing and Service Operations Management, 18(4), 545-558.

Chari, Varadarajan and Hugo Hopenhayn, 1991, "Vintage Human Capital, Growth, and the Diffusion of New Technology," Journal of Political Economy, 99(6), 1142-1165.

Coviello, Decio, Andrea Ichino, and Nicola Persico, 2014," Time Allocation and Task Juggling," American Economic Review, 104(2), 609-23.

Echenique, Federico, Nicole Immorlica, and Vijay Vazirani, 2021, "One-Sided Matching Markets," Online and Matching-Based Market Design, forthcoming.

Elit, Lorraine M., Erin M. O'Leary, Gregory R. Pond, Hsien-Yeang Seow, 2014, "Impact of Wait Times on Survival for Women With Uterine Cancer," Journal of Clinical Oncology, 32(1) 27-33.

Garicano, 2000, "Hierarchies and the Organization of Knowledge in Production," Journal of Political Economy, 108(5), 874-904.

Garicano, Luis, and Luis Rayo, 2017, "Relational Knowledge Transfers," American Economic Review, 107(9), 2695-2730.

Garicano, Luis, and Esteban Rossi-Hansberg, 2012, "Organizing Growth," Journal of Economic Theory, 147(2), 623-656.

Gavazza, Alessandro and Alessandro Lizzeri, 2007, "The Perils of Transparency in Bureaucracies," American Economic Review, 97(2), 300-305.

Hassin, Rafael and Moshe Haviv, 2003, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Springer Science & Business Media.

Kaltenmeier, Christof, Chengli Shen, David S. Medich, David A. Geller,David L. Bartlett, Allan Tsung, and Samer Tohme, 2019, "Time to Surgery and Colon Cancer Survival in the United States," Annals of Surgery, doi: 10.1097/SLA.0000000000003745.

Leshno, Jacob, 2021, "Dynamic Matching in Overloaded Waiting Lists," mimeo.

Leon-Garcia, Alberto, 2008, Probability, Statistics, and Random Process for Electrical Engineering, Third Edition, Pearson Education, Inc.

Lizzeri, Alessandro, and Marciano Siniscalchi, 2008, "Parental Guidance and Supervised Learning," Quarterly Journal of Economics, 123(3), 1161–1195.

Ortoleva, Pietro, Evgenii Safonov, and Leeat Yariv, 2021, "Who Cares More? Allocation with Diverse Preference Intensities," mimeo.

Puterman, Martin L., 2005, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley.

Ünver, Utku, 2010, "Dynamic Kidney Exchange," Review of Economic Studies, 77, 372-414.

Wijeysundera Harindra C., William W.L. Wong, Maria C. Bennell, Stephen E. Fremes, Sam Radhakrishnan, Mark Peterson, and Dennis T. Ko, 2014, "Impact of Wait Times on the Effectiveness of Transcatheter Aortic Valve Replacement in Severe Aortic Valve Disease: A Discrete Event Simulation Model," Canadian Journal of Cardiology, 30, 1162-1169.