# Task Allocation and On-the-job Training

Mariagiovanna Baccara[*]    SangMok Lee[†]    Leeat Yariv[‡][§]

8th October 2020

**Abstract**

We study dynamic task allocation when service providers' expertise evolves. Clients arrive sequentially seeking service. Seniors provide superior service but entail waiting in a queue, which progresses at a speed dependent on their volume. Juniors offer service without wait and become seniors with experience. We show that clients choose senior service too frequently, generating longer waits and little training relative to the social optimum. Welfare gains from centralization are greater for larger institutions, better training technologies, and lower waiting costs. Finally, monitoring the seniors' queue increases welfare but may decrease training. Methodologically, we explore a matching setting in which agents' types are endogenous, and illustrate the usefulness of queueing theory techniques.

**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 Heski Bar-Isaac and Francisco Buera 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. More experienced providers typically offer more reliable service, but junior employees are often more available and benefit from hands-on practice. In this paper, we offer a framework for studying allocation problems in environments with on-the-job training and identify policies that can improve organizations' performance.

We consider environments in which clients—medical patients, individuals seeking legal aid, etc.—get allocated to service providers—medical doctors, lawyers, and so on. Service can be provided by either juniors or seniors. Importantly, the distribution of available juniors and seniors is endogenous: juniors who gain on-the-job training through serving clients become seniors over time.

At the heart of the allocation problems we study is a trade-off between clients' service quality and wait times. If an organization insists on providing only senior service, the quality of service is high, but there is no training. Over time, few seniors are added to the organization due to limited training. Natural attrition, through retirement or job changes, would then result in a scarcity of experienced personnel, leading to prolonged wait times, to the detriment of clients' satisfaction.[2] This trade-off cannot be analyzed using standard tools offered by the matching literature. Extant models generally consider agents' "types"—in our context, their expertise—as exogenously given, rather than as the result of past assignments.

We propose a new and tractable framework to address the task-allocation problem when agents' types are endogenously determined through training. We characterize optimal protocols

[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, for example, http://www.absurgery.org/default.jsp?certgsqe_training).

[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 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/ .

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.

In our model, clients seeking service arrive over time at a Poisson rate. The organization is comprised of junior and senior service 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 may be provided on a first-come-first-served basis. We assume the processing speed of clients by seniors depends on the seniors' volume in the organization. The more seniors available, the speedier the service.

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.

We consider two alternative information environments. First, we study the case in which decision makers, the social planner or the clients themselves, do not observe the current state of the queue when selecting service providers. This constitutes what we refer to as the *limited-monitoring* case. It captures settings in which organizations need to design assignment protocols whose provisions do not depend on how many clients are currently crowding the system. For example, when drafting a curriculum for all surgical residents 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 to which local emergency room to drive to. The second information environment we consider is one in which decision makers can monitor the queue over time and can condition their choices on its current state. This is 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.

Our first set of results pertains to the limited-monitoring case. Decision makers, the social planner in the centralized setting or individual clients in the discretionary environment, choose the probability with which they enter the seniors' queue. Absent any training, the model corresponds to what is often termed an M/M/1 queue in the queuing literature (see, for instance, Leon-Garcia, 2008). Training, however, links the decision maker's allocation choices to the volume of seniors, which impacts their service speed. Analyzing such training capabilities requires new techniques.

The training constraint determines the fraction of clients that juniors need to handle in order to sustain any volume of seniors. In steady state, the optimal policy maximizes the expected welfare subject to the training constraint. In the discretionary equilibrium, the fraction of clients joining the queue for senior service makes any client indifferent between the two types of service.

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. The social planner internalizes such externalities, while individuals acting on their own do not. It follows that discretionary settings are associated with more clients seeking senior service than is optimal. Consequently, discretionary settings feature higher average service quality but longer wait times relative to what is socially optimal.

Our characterization allows for some natural comparative statics regarding outcomes, the quality of service and expected wait times. Increases in the quality differential between senior and junior service, or decreases in wait costs, naturally make senior service more appealing. They result in a higher average service quality, longer wait times, and less training.

The impacts of changes in clients' arrival rate or improvements in the training technology are more subtle. More clients arriving, as a consequence of, say, a merger between hospitals or firms, could potentially cause more congestion, but at the same time present additional training opportunities. Improved technology, due to the introduction of online training options, simulated activities such as surgeries for doctors or mock 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 technology yield increases in average service quality. Furthermore, they generate decreased wait times for senior service in the centralized setting. Equilibrium wait times for senior service in the discretionary setting, however, does not change.

The welfare gap between the optimal and discretionary settings increases as clients' arrival rate grows or the training technology improves. Therefore, such changes in the environment make centralized interventions more impactful on clients' welfare.

In the perfect-monitoring environment, both the optimal policy and the discretionary equilibrium follow protocols that direct clients to the senior queue as long as the queue does not exceed a certain threshold. If the queue is sufficiently long, any arriving client is directed to junior providers. Absent training, the resulting queue falls under the rubric of what is often termed an M/M/1/k queue in the queueing literature. As before, the endogeneity of service speeds requires the development of some new methodological tools.

As it turns out, the social planner's optimal threshold shares features with her optimal policy when monitoring is limited. It is, again, derived from an optimization of the clients' welfare subject to the same training constraint. 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. The externalities present in the limited-monitoring case are also present when queues are monitored perfectly. In particular, the optimal threshold is set so that fewer clients seek senior service relative to what they would choose in the discretionary equilibrium.[3]

Our analysis of the perfect-monitoring case allows us to evaluate, in our last set of results, the effects of monitoring precision on centralized and discretionary processes. With limited monitoring, agents can only rely on expectations of queue length. In contrast, perfect monitoring allows decision makers to utilize the senior queue only when it is sufficiently short.

In the centralized setting, monitoring can only help in terms of welfare. With perfect monitoring, the social planner could certainly 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 would maintain service quality and reduce 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.

In the discretionary setting, results are more nuanced. If the training technology is relatively inefficient, 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, we show that the equilibrium welfare is always higher under perfect monitoring.

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

[3] The comparative statics pertaining to the perfect-monitoring environment by and large mimic those of the limited-monitoring setting, with some minor exceptions.

allocation problems. 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.[4] 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

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 are explored in several papers. Lizzeri and Siniscalchi (2008) consider parents who decide how much to shelter their children from mistakes, which are risky but provide useful learning opportunities. The result is that parental intervention occurs as differences between parents and children grow. 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. Fudenberg and Rayo (2019) introduce apprentices' effort into a similar setting and study the equilibrium dynamics of effort provision over time with endogenously evolving participation constraints. Again, none of these papers examine the implications of training considerations on task allocation and on potential delays in service.[5]

[4] 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.

[5] 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

Work in organization theory has studied how to allocate opportunities to heterogeneous individuals that 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.[6] None of these papers examines the interaction between task allocation and employees' training, and how lack of training may yield delays in service.

Our paper is also related to a recent and growing literature on dynamic allocation and matching. Leshno (2019) studies a one-sided market in which objects—for example, public houses—are sequentially allocated to agents waiting in a queue. He focuses on cases in which individuals' preferences are unknown to the planner. For related work, see Anderson, Ashlagi, Gamarnik, and Kanoria (2017), Bloch and Cantala (2017), and references therein.

On dynamic two-sided matching, Baccara, Lee, and Yariv (2020) focus on a setting in which heterogenous agents arrive at the market over time, and incur a waiting cost until they are matched to an agent on the other side. The paper explores the trade-off between waiting for a thicker market, allowing for higher-quality matches, and minimizing agents' waiting costs in both centralized and discretionary settings. Ünver (2010), and Akbarpour, Li, and Oveis Gharan (2020) focus on the organ-donation application, in which donors and recipients arrive stochastically, and preferences are compatibility-based. Zenios (1999) employs a queueing model to explain waiting times across different categories of patients on kidney transplant wait lists.[7]

## 2 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 $\lambda$. 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

example, through the training of employees.

[6] 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.

[7] More recent contributions in related settings are Herbst and Schickner (2016), Doval and Szentes (2019), Margaria (2020), and Loertcher, Muir, and Taylor (2019). There is also a theoretical literature that studies discretionary matching processes that are dynamic (see, e.g., Ferdowsian, Niederle, and Yariv, 2020, Haeringer and Wooders, 2011, and Pais, 2008).

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. Let $\mu_t \in \mathbb{R}_+$ denote the mass of seniors at time $t$. While this mass evolves over time, our analysis will mostly concentrate on the limiting mass of seniors within the settings we analyze. As the mass of senior providers increases, senior service becomes more rapid. Formally, the completion of service provided by seniors of mass $\mu_t$ follows a Poisson process with parameter $\mu_t$.[8] Clients directed at seniors form a queue and are served on a first-in-first-out (FIFO) basis.

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 $c \geq 0$ is the waiting cost and $W$ is the client's wait time in the queue. Implicitly, we assume that a client does not experience waiting costs while receiving service. This is done for presentation purposes, and to make the link with the queueing literature more direct, but does not impact our results qualitatively.

We consider centralized and discretionary allocation problems. In a *centralized allocation*, a planner allocates clients to seniors or juniors. The planner's objective is to maximize the average client payoff. 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 *limited monitoring*, decision makers—the social planner or the clients—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. With *perfect monitoring*, decision makers observe the queue for senior service and allocation decisions can be contingent on its current length.

Last, we consider a reduced-form model of training, which captures the process by which juniors transition to senior positions. In either the discretionary or centralized setting, each allocation policy gives rise to a time-average number of clients $x$ directed at juniors, and to a mass of seniors $\mu$ determined by a production function $f : \mathbb{R}_+ \to \mathbb{R}_+$. That is, $\mu = f(x)$. 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]

[8] Considering service rates that are proportional to the mass or quality of senior employees with an arbitrary multiplier amounts to a normalization and does not change our results.

[9] By a well-known queueing theory result (see details in the Appendix), if all clients join the queue, namely

Since $\mu$ represents the speed at which seniors process clients, for a given mass of clients seeking junior service, any technology that improves either the training or the service speed directly, corresponds to an upward shift of the function $f$.

The production 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 training 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 volume of clients and $l$ represents the average skill level in the general intern population. The production function captures the fraction of interns that are then selected for permanent positions.

There are several special cases that serve as important benchmarks:

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

- **Costless delay:** As $c \to 0$, delaying clients comes at no cost. 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 training:** When the production function $f$ is constant, there is no on-the-job training and the mass of seniors remains fixed at some exogenous level $\overline{\mu}$. 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 turnover is high. For instance, many departments employ post-docs that are not destined to receive a tenure-track position.

Several assumptions in our setting merit discussion. 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

$x = 0$, the average waiting time in the queue with a processing rate $\mu(> \lambda)$ is $\frac{1}{\mu-\lambda} - \frac{1}{\mu}$, which becomes arbitrarly large as $\mu$ approaches $\lambda$. The other potential "corner" allocation, $x = \lambda$, never occurs since $h > l$.

discounting. There are two reasons for considering flow costs in our setting. 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.[10] This is hardly the case when clients are medical patients seeking treatment, students awaiting their grades, etc. As we soon discuss, 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 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.

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.[11] 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.[12]

Finally, 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

[10] See Ortoleva, Safonov, and Yariv (2020) for a discussion of how discounting impacts optimal allocations of items.

[11] 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.

[12] 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 serviced first, while others who have been waiting remain in the queue. See Margaria (2019) and references therein.

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. Indeed, in some settings, the volume of juniors is a choice: in the medical world, hospitals decide on the size of their residency programs; in universities, departments decide how many junior professors to hire. In principle, considering a model involving two queues could generate some insights on the implications of the volume of unskilled labor an organization decides to employ.

# 3 Limited Monitoring

We start by analyzing the case in which 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.

## 3.1 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). Therefore, the characterization in both settings with limited monitoring boils down to the fraction $q \in [0, 1]$ of clients that are served by seniors. The remaining fraction $1 - q$ of clients is served by juniors. Therefore, seniors serve clients at a rate $\mu$, where $\mu = f((1 - q)\lambda)$. We call this last equality the *training constraint.*[13]

[13] One could also model the evolution of training explicitly. Our analysis then corresponds to dynamics of the form:

$$\frac{d\mu_t}{dt} = -\delta\mu_t + g(x),$$

A discretionary allocation or a centralized allocation policy are characterized by a pair $(q, \mu)$. Specifically, a discretionary equilibrium is defined through two constraints. First, each client optimizes her expected payoff, which we soon spell out, given all other clients' (symmetric) strategy of seeking senior service with probability $q$ and the mass $\mu$ 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 $(q, \mu)$ 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 $q\lambda$. The *utilization* of senior workers, given by $\frac{q\lambda}{\mu}$, is always in $(0,1)$. Indeed, 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 (see, for instance, Leon-Garcia, 2008). We provide the relevant preliminaries in the Appendix, which suggest that the average waiting time in the queue, conditional on entry, is

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

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. This, together with the production 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 $\mu$ 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 linear production technology case, where $f(x) = ax$, with $a > 0$, is a particularly useful

where $x = (1-q)\lambda$, and $\delta \in (0,1)$ is a decay parameter, reflecting retirement, job transitions, and so on. The resulting limit of $\mu_t$ is the unique steady state $\mu = \frac{g(x)}{\delta}$.

one to consider. While it is in many ways special, it fits well with numerous applications. For instance, it can reflect applications in which juniors' performance on tasks serves as a screening instrument, and only a fixed fraction justify promotion. Alternatively, it can capture a fixed flow of seniors and some fixed fraction of turnover or retirement of senior employees. The parameter $a$ is then a proxy of the training efficacy.

With a linear training technology, an increase in $\lambda$, keeping $q$ constant, implies a decrease in $\mathbf{E}[W]$. To see this, suppose clients' arrival rate $\lambda$ doubles while the fraction $q$ of clients served by seniors stays fixed. For a linear training technology, the fraction of seniors exactly doubles as well. Furthermore, an increase in $a$ increases $\mu$, and therefore decreases $\mathbf{E}[W]$, for any given $q$.

The training constraint yields the feasible set of $(q, \mu)$ 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:

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

Let $\theta \equiv \frac{h-l}{c}$ denote the quality differential per unit cost.

**Proposition 1 (Limited Monitoring)** 1. *In the discretionary setting, the unique equilibrium is governed by $(q_L^e, \mu_L^e)$ that solves:*

$$\theta = \frac{\lambda q}{\mu(\mu - q\lambda)} \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{1}$$

2. *In the centralized setting, the planner has a unique optimal policy governed by $(q_L^*, \mu_L^*)$ that solves:*

$$\theta = \frac{q\lambda}{\mu}\frac{2\mu - q\lambda}{(\mu - q\lambda)^2}\left(\frac{q\lambda}{\mu}f' + 1\right) \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{2}$$

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 1, we show that the first-order approach is valid for optimizing this objective. At the optimum, we then have $\theta = \frac{d(q\mathbf{E}[W])}{dq}$. This condition translates into:

$$\lambda(h-l) = c\left(\lambda\mathbf{E}[W] + (q\lambda)\left(\frac{\partial\mathbf{E}[W]}{\partial q} + \left|\frac{\partial\mathbf{E}[W]}{\partial\mu}\right|\left|\frac{d\mu}{dq}\right|\right)\right). \tag{3}$$

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 $\mu, q$, and $\lambda$, simple calculus generates the characterization appearing in the proposition.

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, Little's formula implies that $\mathbf{E}[Q] = \lambda q\mathbf{E}[W]$. We can therefore write the first-order condition as $\lambda\theta = \frac{d(\mathbf{E}[Q])}{dq}$. This formulation will appear when we consider more detailed monitoring of markets.

## 3.2 Comparative Statics and Welfare Comparisons

We now turn to some comparative statics resulting from our characterization.

### 3.2.1 Quality Differential and Waiting Costs

As $\theta$ 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. That is,

**Corollary 1 (Limited Monitoring – Service Quality and Wait Costs)** *As $h-l$ increases or $c$ decreases, both $q_L^e$ and $q_L^*$ increase, while the induced masses of seniors $\mu_L^e$ and $\mu_L^*$ decrease. Expected wait times increase in both settings.*

Of particular interest are cases in which waiting costs $c$ take on extreme values. The case of $c \to \infty$ corresponds to the case of no delay. In this case, in both the centralized and discretionary settings, waiting is minimized and clients choose junior service. This yields $q^e, q^* \to 0$ and $\mu^e, \mu^* \to f(\lambda)$.

In contrast, as $c$ approaches 0, clients naturally seek senior service more and 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 of $q$. Therefore, $q^e, q^* \to \overline{q}$ and $\mu^e, \mu^* \to f((1-\overline{q})\lambda)$.

### 3.2.2 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. 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 mergers of different service units: hospitals, law firms, etc. The resulting overall arrival rate of clients in the post-merger organization would presumably be higher than the arrival rate at each of the original organizations. 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—improved 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.[14]

In general, an increase in clients' arrival rate $\lambda$ 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 $\lambda$ and the efficacy of training.

**Proposition 2 (Limited Monitoring – Arrival Rates and Training Efficacy)** *Suppose $f(x) = ax$, for some $a > 0$. In both the discretionary and centralized settings, there*

[14] See https://trainingmag.com/ for annual reports on training expenditures across industries in the U.S.

*is higher quality, more training, and a greater mass of seniors as the arrival rate of clients increases or the training technology improves. That is, $q_L^e$ and $q_L^*$ as well as $\mu_L^e$ and $\mu_L^*$ increase with $\lambda$ and $a$. Moreover, expected wait time in the discretionary setting, $\mathbf{E}[W_L^e]$, is constant in both $\lambda$ and $a$, while expected wait time in the centralized setting, $\mathbf{E}[W_L^*]$, decreases in both $\lambda$ 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. As described before, the expected waiting time is half the original waiting time, making senior service more desirable, and leading to an increase in the optimal fraction of clients to be served by seniors in both the discretionary and centralized settings. If, however, the training technology has decreasing marginal returns, the mass of seniors would less than double. This would imply, in both the discretionary and centralized settings, that the impact on the expected waiting time depends on the slope of the training function, generating either an increase or a decrease in the fraction of clients served by seniors.

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.

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 ((2)), tantamount to $\theta = \frac{d(q\mathbf{E}[W])}{dq}$, which becomes

$$\theta = \frac{z(q;a)}{\lambda} + \frac{qz'(q;a)}{\lambda}. \tag{4}$$

It is easy to verify that each term on the right-hand side of (4) increases in $q$.[15] Since $z'(q;a) <$

[15] Indeed, $\frac{z(q;a)}{\lambda} + \frac{qz'(q;a)}{\lambda}$ increases in $q$ due to the convexity of $\mathbf{E}[W]$.

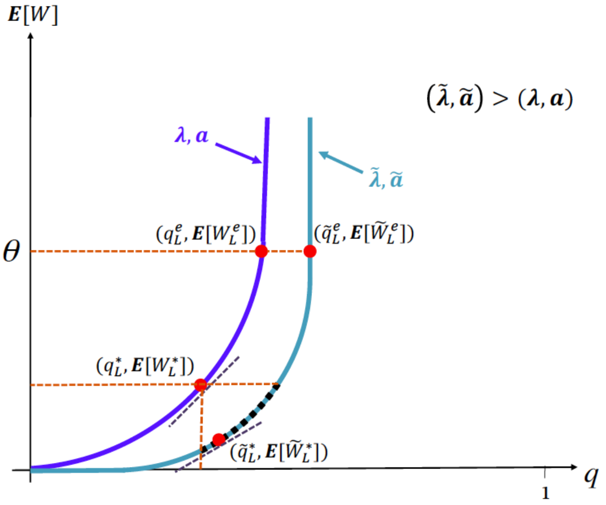

Figure 1: Impacts of Changes in Arrival Rates or Training Efficacy on Quality and Wait Times

$z'(q;a) + qz''(q;a) = [qz'(q;a)]'$, we have that $z(q;a)$ increases in $q$ slower than $qz'(q;a)$. This implies that if either $\lambda$ or $a$ increase, the optimal $q$ increases, but the expected wait time $\mathbf{E}[W] = \frac{z(q;a)}{\lambda}$ decreases.

Figure 1 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.[16] It depicts wait times as functions of $q$ for two values of arrival rates and training efficacies, the unique discretionary equilibrium choices—points in which expected times coincide with $\theta$, and the resulting optimal solutions—points in which the slope of $q\mathbf{E}[W]$ is fixed at $\theta$.

To conclude, Proposition 2 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 $\lambda$ would cause $q$ to decrease and $\mathbf{E}[W]$ to remain unchanged in both the discretionary and centralized settings.

[16] 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 $\lambda$.

#### 3.2.3 Welfare Comparison

We now turn to the impact of some parameters of our model on clients' expected welfare. The average utility 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 increased 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 changes in $\lambda$ could go either way. However, for linear training technologies, Proposition 2 indicates that the optimal fraction of clients directed at senior service increases, while wait times decrease. Consequently, $V_L^*$, and thus $V_L^* - V_L^e$, increase in $\lambda$. Similar comparative statics follow for the training efficacy. We therefore have the following corollary.

**Corollary 2 (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 $\lambda$ and $a$.*

Our discussion above considers the per-person 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 2 would then continue to hold. However, as arrival rates increase, the benefits of centralization would become even more pronounced as more clients are impacted.

This discussion suggests that organizations obtain greater advantages from an intervention toward centralization when the quality of senior service improves, waiting costs decrease, and when either the arrival rate or the training technology efficacy increase.

# 4 Perfect Monitoring

In this section, we assume 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 discretionary setting. For example, in many organizations, administrative staff are assigned duties based on their current workload. Analogously, patients may call clinics to learn about their expected wait times in advance of deciding whom 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 additional monitoring yields and later compare them with the limited-monitoring benchmark.

## 4.1 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 $k$.[17]

Formally, we consider a continuous-time Markov chain for the number $Q_t$ of clients waiting in the queue at time $t$, excluding any client currently being served. The processes we analyze correspond to those referred to as $M/M/1/k$ queues in the queueing literature, with arrival rate $\lambda$, and service rate $\mu$ (see our primer in the Appendix). In the perfect-monitoring case, the rate $\lambda/\mu$ can exceed 1, as the size of the queue is bounded above by $k$. The unique steady-state distribution across the relevant states of the queue, $\{0, 1, \ldots, k\}$, is as follows.

**Lemma 1 (Steady-state Distribution under a Threshold Policy)** *The probability $p_j$ of having $j$ clients in the queue for senior service in the unique steady state is*

$$p_j = \begin{cases} \frac{1}{k+1} & \text{if } \mu = \lambda \\ \frac{(\lambda/\mu)^j(1-(\lambda/\mu))}{1-(\lambda/\mu)^{k+1}} & \text{if } \mu \neq \lambda. \end{cases} \quad \forall j = 0, 1, \ldots, k. \tag{5}$$

In steady state, the average number of clients waiting in the senior queue is

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

[17] 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.

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

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

In steady state, the time-average number of clients joining the seniors' queue is $\lambda(1 - p_k)$. As in the previous section, let $\mathbf{E}[W]$ denote these clients' average waiting time before being served by seniors, conditional on joining the (possibly empty) queue. By Little's formula,

$$\mathbf{E}[Q] = \lambda(1 - p_k)\mathbf{E}[W].$$

In steady state, the time-average number of clients served by juniors is $\lambda p_k$. The mass of seniors is governed by the training technology and given by $\mu = f(\lambda p_k)$.[18]

The average fraction of clients served by seniors is $q \equiv 1 - p_k$. 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 approaching juniors is likely to be closely followed by another.

Without loss of generality, we assume that $k \geq 1$ throughout our analysis. Whenever the queue for seniors is empty, it is optimal for any individual client and the planner to seek senior service, which comes at a higher quality.

For exposition sake, it is convenient to relax the integer constraints on $k$.[19] 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 $\mu$.

**Lemma 2 (Service Quality under a Threshold Policy)** *For all $\mu$,*

$$q(k; \mu, \lambda) \equiv 1 - p_k = \begin{cases} \frac{k}{k+1} & \text{if } \lambda = \mu, \\ \frac{1-(\lambda/\mu)^k}{1-(\lambda/\mu)^{k+1}} & \text{if } \lambda \neq \mu \end{cases} \tag{6}$$

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

[18] One could model the evolution of training explicitly. Consider a setting in which every $T$ periods, the number of seniors adjusts as follows. A fraction is lost to decay—reflecting retirement, job transitions, etc.—and a random number of seniors is added through training, which naturally depends on the training technology in place. Under standard assumptions, this Markov process is ergodic and, for any fixed $T$, converges to a steady-state distribution over the number of seniors. As the size of the adjustment window $T$ grows, the steady-state distribution converges to a degenerate distribution centered at some $\mu$, which is the focus of our analysis.

[19] For any given $\lambda$ and $\mu$, 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.

Lemma 2 allows us to describe any outcome either by $(k, \mu)$ or by $(q, \mu)$. In what follows, we characterize solutions in terms of $(q, \mu)$ as it facilitates a direct comparison of solutions under perfect and limited monitoring. For a given pair $(q, \mu)$, the corresponding threshold $k$ is identified by the inverse of (6):

$$k(q; \mu, \lambda) \equiv \begin{cases} \frac{q}{1-q} & \text{if } \lambda = \mu \\ \frac{\log(1-q)-\log(1-(q\lambda/\mu))}{\log(\lambda/\mu)} & \text{if } \lambda \neq \mu. \end{cases} \tag{7}$$

## 4.2 Discretionary and Centralized Allocations

Consider first the case in which clients have discretion over which service to seek upon entering the market. In analogy to the limited-monitoring case, a symmetric equilibrium $(k, \mu)$ 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 $\mu$ of available 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 $\mu$ 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, it must be that $\mu = f((1-q)\lambda)$.

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.[20]

[20] 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.

From the social planner's perspective, threshold-based policies are optimal within the set of stationary policies that depend only on the number of clients waiting or being served in the senior queue.[21] 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 $\mu = f(\lambda p_k)$. Using the notation introduced before, this problem is equivalent to:

$$\max_{q,\mu} q\lambda\theta - \mathbf{E}[Q].$$

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:*

$$\mathbf{E}[Q] = \begin{cases} \frac{q(2q-1)}{2(1-q)}, & \text{if } q = 1 - \frac{1}{a}, \\ \frac{1}{a(1-q)-1}\left[\frac{q}{a(1-q)} - (1-q)^{\frac{\log(1-\frac{q}{a(1-q)})-\log(1-q)}{\log a + \log(1-q)}}\right] & \text{otherwise.} \end{cases} \tag{8}$$

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 $q\lambda < \mu = a(1-q)\lambda$.[22] In the Appendix, we show that the objective is continuously differentiable and single-peaked. Therefore, the problem is solvable using a first-order condition approach.

We have the following characterization of allocations under perfect monitoring.

[21] 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. By Theorem 7.1.9 of Puterman (2005), for any fixed $\mu$, an optimal policy is indentified 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.

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

**Proposition 3 (Perfect Monitoring)** 1. *In the discretionary setting, the unique equilibrium is governed by* $(q_P^e, \mu_P^e)$ *that solves:*

$$k(q, \mu; \lambda) = \mu\theta + 1 \quad \textit{and} \quad \mu = f((1-q)\lambda). \tag{9}$$

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

$$\lambda\theta = \frac{d\mathbf{E}[Q]}{dq} \quad \text{and} \quad \mu = a(1-q)\lambda. \tag{10}$$

The second part of Proposition 3 is similar in nature to the characterization pertaining to the planner's optimal policy with limited monitoring, as described in Proposition 1. The optimal policy satisfies a first-order condition and a training constraint. The crucial difference between the two is the derivation of the expected number of clients waiting in the queue, or relatedly, the expected wait time in the seniors' queue.

In the Online Appendix, we illustrate that most of the comparative statics pertaining to the optimal policies under perfect monitoring resemble those of the limited-monitoring setting.

# 5 The Impacts of Centralization

We can now compare the impacts of centralization on outcomes for each of our monitoring settings. Intuitively, for both the limited- and the perfect-monitoring scenarios, 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. Formally, assume $q_P^* > \frac{\mu_P^*}{\mu_P^* + \lambda}$, so that $k_P^* > 1$.

**Corollary 3 (Impacts of Centralization)** $q_X^* < q_X^e$ and $\mu_X^* > \mu_X^e$ for $X = L, P$. *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.

Corollary 3 holds even absent training, when the production function $f$ is constant. In other words, the inverse link between service quality and wait times is not a pure artifact of training by doing. Indeed, in the discretionary setting, there are 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 efficiency and yields greater welfare in both the discretionary and centralized settings. We also identify how monitoring affects outcomes in terms of quality, training, and waiting times.

At the heart of the effects of perfect monitoring is the ability, either of 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* $(q, \mu)$*,*

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

Lemma 4 also suggests that, when expected wait times are similar with limited or with perfect monitoring, the quality of service afforded by perfect monitoring is higher. This observation underlies our welfare comparisons.

## 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 will see, this feature implies that training can go up or down with improved monitoring, depending on the efficacy of the training technology. 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:

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

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 $(q''_L,\mu'') \in G_L$ and $\mu'' \neq \mu'$, either $\mu'' < \mu'$ and $q''_L < q''_P$ or $\mu' < \mu''$ and $q''_P < q''_L$.

**Proposition 4A (Impacts of Monitoring in Discretionary Settings)** *$G_P$ strictly single crosses $G_L$ from below. Furthermore, welfare is greater when monitoring is perfect.*

Proposition 4A, the intuition for which we soon describe, suggests that the comparison of $q^e_L$ and $q^e_P$ depends on the training technology. Consider the linear training characterized by $f(x) = ax$ for $a > 0$. Single crossing of the indifference curves under limited and perfect monitoring suggests that the ranking of equilibrium training under limited and perfect monitoring depends on the training efficacy, as depicted in Figure 2. 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 2 illustrates the threshold $a^*$ at which the impact of monitoring reverses. We therefore have the following corollary.

**Corollary 4 (Training Efficacy and Monitoring in Discretionary Settings)** *There exists $a^* > 0$ such that if $f(x) = ax$, then for $0 < a < a^*$, $q^*_L < q^*_P$ and $\mu^*_L > \mu^*_P$, while for $a > a^*$, $q^*_L > q^*_P$ and $\mu^*_L < \mu^*_P$.*

Proposition 4A implies that a similar conclusion to that of Corollary 4 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

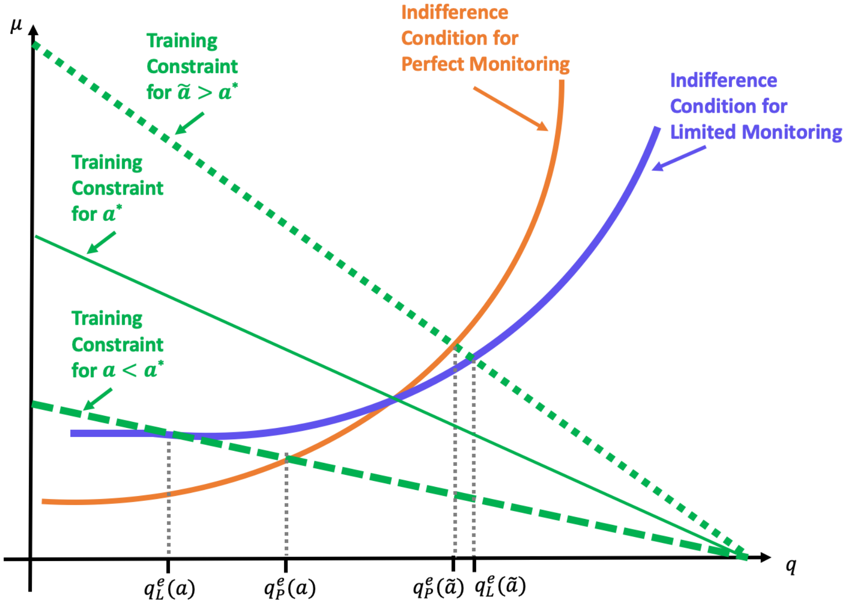

Figure 2: Impacts of Monitoring on Service Quality and Training

lower service quality under $g$, it also does so under $f$.[23]

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.

To gain some intuition for the single-crossing property identified by Proposition 4A as well as for Corollary 4, 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,

[23] 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.

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 is strictly higher than $l$.

## 6.2 Centralized Settings

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

**Proposition 4B (Impacts of Monitoring in Centralized Settings)** *Suppose $f(x) = ax$, for some $a > 0$. Then $q_L^* < q_P^*$ and $\mu_L^* > \mu_P^*$. Furthermore, welfare is greater when monitoring is perfect.*

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. Now, 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, at least when the planner governs clients' allocation to services, improving monitoring of the senior queue is beneficial to clients. The increase in welfare, however, 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, be it 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 limited and perfect monitoring of the seniors' queue.

Several insights follow from our analysis. First, when clients can choose whom to seek service from, the average service quality is inefficiently high and queues are too long. Second, as clients' arrival rate or the training technology's efficacy increase, both service quality and the scope of training increase. Consequently, average wait times decrease. Third, 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.

Pricing of service is also absent in our framework. This fits environments in which service has to be provided, as is largely the case for hospitals and some commercial settings that cannot discriminate clients based on their needs or demands. Nonetheless, in many settings it is natural to consider differential prices for varying service qualities. Such analysis could also shed light on the impacts of competition between organizations, which is altogether absent in our model. We view these as interesting angles for future research. We hope the tools we introduce are useful for inspecting such alternative goals.

# 8 Appendix

## 8.1 Primer on Queueing

The limited-monitoring case in Section 3 employs what is termed the M/M/1 queue in the queueing literature, while the perfect-monitoring case in Section 4 employs an M/M/1/k queue. Here we provide a summary of the results relevant to our analysis. For more details, see, for example, Leon-Garcia (2008).

### 8.1.1 M/M/1 Queue

Clients seeking service arrive at the market over time $t \in [0, \infty)$ according to a Poisson process with arrival rate $\rho$. One provider can serve at most one client at any given time. Service completion times are independent across clients and follow an exponential distribution with parameter $\mu$. Upon arrival, each client joins a queue until the provider becomes available. If the provider is not busy helping other clients, the client is served immediately. There is no limit on the possible length of the clients' queue. Clients are served according to a first-in-first-out (FIFO) protocol.

The total number of clients in the system, either waiting in the queue or being served, at time $t$, $N_t$, is a continuous-time Markov chain and takes values in $\{0, 1, \ldots, \}$. When $N_t = 0$, there are no clients being served or waiting. When $N_t = 1$, the system has only one client who is being served. When $N_t \geq 2$, at least one client is waiting in the queue. The number of clients $N_t$ increases by one when a client arrives, which occurs at a rate $\rho$. It decreases by one when the service of a client is completed, which occurs at a rate $\mu$. The ratio $\psi \equiv \frac{\rho}{\mu}$ denotes the provider's *utilization.* As long as $\psi < 1$, $N_t$ has a stationary distribution denoted by $\{p_0, p_1, \dots\}$ such that exactly $i$ clients are in the system with probability $p_i$.[24] The inflow-outflow equalities, known as the global balance equations, are

$$\rho p_0 = \mu p_1,$$
$$(\rho + \mu) p_j = \rho p_{j-1} + \mu p_{j+1}, \quad \forall j = 1, 2, \dots.$$

They yield the stationary distribution

$$p_j = (1 - \psi)\psi^j, \quad j = 0, 1, 2, \dots.$$

The average number of clients waiting in the queue, excluding the client currently being served, is

$$\mathbf{E}[Q] \equiv \sum_{j=1}^{\infty} (j-1) p_j = \frac{\psi}{1-\psi} - p_0 = \frac{\psi^2}{1-\psi}.$$

Let $\mathbf{E}[W]$ be the average waiting time in the queue. *Little's formula* guarantees that

$$\mathbf{E}[W] = \frac{\mathbf{E}[Q]}{\rho} = \frac{1}{\mu - \rho} - \frac{1}{\mu}.$$

The intuition behind the formula is the following. Take any time interval, say $[s, s+t)$ during which the system is at the steady state. The total time clients spend waiting in the queue is approximately $(\rho t)\mathbf{E}[W]$, so the average number of clients waiting in the queue at any given time is $\mathbf{E}[Q] = \rho \mathbf{E}[W]$.

### 8.1.2 M/M/1/k Queue

An M/M/1/k queue is similar to an M/M/1 queue but assumes the service provider can accommodate up to $k$ clients, with one client being served, and at most $k-1$ clients waiting in the queue. If a client finds $k$ others present upon arrival, she is turned away. As before, clients' arrival follows a Poisson distribution with parameter $\rho$, and the provider completes each client's service at times following an exponential distribution with parameter $\mu$. Since, by construction, the length of the queue is bounded, $\psi \equiv \frac{\rho}{\mu}$ need not be lower than 1.

The total number of clients in the system $N_t$ follows a continuous-time Markov chain over $\{0, 1, 2, \dots, k\}$. The inflow-outflow equalities, which we omit, yield the following restrictions on

[24] The restriction $\psi < 1$ is necessary because, if $\psi \to 1$, the average wait time, which we address shortly, diverges.

the stationary distribution:

$$p_j = \psi^j p_0, \quad \forall j = 1, \ldots, k.$$

Since

$$1 = \sum_{j=0}^{k} p_j = p_0 \sum_{j=0}^{k} \psi^j = \begin{cases} p_0(k+1) & \text{if } \psi = 1, \\ p_0\left(\frac{1-\psi^{k+1}}{1-\psi}\right) & \text{if } \psi \neq 1, \end{cases}$$

the stationary distribution is given by:

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

Similarly, the average number of clients in the queue is

$$\begin{aligned} \mathbf{E}[Q] &= \sum_{j=1}^{k} (j-1)p_j = \sum_{j=1}^{k} (j-1)\psi^j p_0 \\ &= \begin{cases} \frac{k(k-1)}{2(k+1)} & \text{if } \psi = 1, \\ p_0\left(\psi^2 + 2\psi^3 + \cdots + (k-1)\psi^k\right) & \text{if } \psi \neq 1. \end{cases} \end{aligned}$$

In particular, if $\psi \neq 1$, the above expression can be written as:

$$\mathbf{E}[Q] = \frac{1}{1-\psi^{k+1}}\left(\frac{\psi^2 - \psi^{k+1}}{1-\psi} - (k-1)\psi^{k+1}\right).$$

Finally, in the steady state, since a new client is turned away only when $N_t = k$, the average number of clients that join the queue over a unit of time is $\rho(1-p_k)$. Let $\mathbf{E}[W]$ denote those clients' average waiting time. By Little's formula, we have $\mathbf{E}[W] = \frac{\mathbf{E}[Q]}{\rho(1-p_k)}$.

## 8.2 Proofs for Limited Monitoring

**Proof of Proposition 1**:

1. For any given mass of seniors $\mu$ 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

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

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 $\infty$.

A unique solution $q_L^e$ of the indifference condition exists, and $\mu_L^e = f((1-q_L^e)\lambda)$ follows.

2. We write the planner's problem as follows:

$$\max_{q\in(0,\mu/\lambda)} q\left(h - c\mathbf{E}[W]\right) + (1-q)l \iff \max_{q\in(0,\mu/\lambda)} q\theta - \frac{\lambda q^2}{\mu(\mu-\lambda q)},$$

subject to

$$\mu = f((1-q)\lambda).$$

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 $(q,\mu)$ 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 $(q_1,\mu_1)$ and $(q_2,\mu_2)$ 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 $(q,\mu)$ is convex because $f$ is (weakly) concave. Also, there is no corner solution. Indeed, if $q \to 0$, 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:

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

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 2**: 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 $q < a(1-q)$.

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

In the centralized setting, the planner's optimal choice $q_L^*$ solves (3), which becomes

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

Observe that $z(q;a)$ is strictly increasing in $q$, and strictly decreasing in $a$ because, for any $a_l < a_h$,

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

which clearly holds. On the other hand,

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

and each term of the above expression for $z'(q;a)$ is strictly increasing in $q < a(1-q)$ and strictly decreasing in $a$. Thus, if either $\lambda$ or $a$ increase, $q$ has to increase to satisfy (11).

In fact, when $q$ increases, the first term on the right-hand side of (11) increases more slowly than the second term because $z'(q;a) < (qz'(q;a))' = z'(q;a) + qz''(q;a)$. Hence, when the solution $q$ increases in $\lambda$ or $a$, the expected waiting time $\mathbf{E}[W] = \frac{z(q;a)}{\lambda}$ decreases. ■

## 8.3 Proofs for Perfect Monitoring

In the following proofs, we let $\phi \equiv \frac{\lambda}{\mu}$. When $\lambda$ is fixed, $\phi$ and $\mu$ uniquely determine one another. In the case of linear production, where $f(x) = ax$ for some $a > 0$, we have $\mu = a(1-q)\lambda$ and $\phi = \frac{1}{a(1-q)}$.

**Proof of Lemma 2**: We omit a trivial proof for the case of $\phi = 1$. If $\phi \neq 1$, then $1 - p_k = 1 - \frac{1-\phi}{\phi^{-k}-\phi}$. In either case of $\phi > 1$ 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 (8) and substitute $k$ with (7). 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,

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

**Proof of Proposition 3**:

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

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

   The graph $G_1$ is continuous and downward sloping: as $q$ increases from 0 to 1, $\mu$ decreases from $f(\lambda)$ to $f(0)$.

   Consider the graph $G_2$. The function $k(q, \mu; \lambda)$ is continuous in $q$ and $\mu$, and strictly increasing in $q$, see Lemma 2 and (7). 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 $\mu$. Therefore, the graph $G_2$ is continuous and upward sloping: as $q$ increases from 0 to 1, $\mu$ increases from $-\frac{1}{\theta}$ to $\infty$. 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

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

   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 $\phi$ while $\lambda$ 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 of $\phi$ 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, $\underline{\phi} = \frac{1+\sqrt{1+4a}}{2a}$.

In the rest of the proof, we show that the objective in $[P']$ is continuously differentiable and strictly concave in the single-choice variable $\phi$. Then, the objective of the original problem $[P]$ is continuously differentiable and single-peaked in the single choice variable $q$, which concludes the proof. We divide our arguments into two steps. The first step shows that $\mathbf{E}[Q]$ is continuously differentiable at $\phi = 1$, which allows us to focus on the functional form for the case of $\phi \neq 1$, by taking the value at $\phi = 1$ as $\lim_{\phi \to 1} \mathbf{E}[Q]$, and similarly for $\frac{d\mathbf{E}[Q]}{d\phi}$ at $\phi = 1$. The second step shows that $\mathbf{E}[Q]$ is strictly convex.

*Step 1: $E[Q]$ is continuously differentiable at $\phi = 1$.*

To show Step 1, we first present the following lemmas A-C. To ease expositions, we denote $z \equiv a(1-\phi)+1$ and note that $dz/d\phi = -a$, $\lim_{\phi \to 1} z = 1$, and $\lim_{\phi \to 1} \frac{\log z}{1-\phi} = \lim_{\phi \to 1} \frac{a}{z} = a$.

**Lemma A** $\mathbf{E}[Q]$ *is continuous at $\phi = 1$.*

**Proof of Lemma A**: Using L'Hopital's rule, we obtain

$$\begin{aligned} \lim_{\phi \to 1} \mathbf{E}[Q] &= \frac{1}{a} + \lim_{\phi \to 1} \frac{1}{1-\phi}\left((\phi^2 - 1) + 1 + \frac{\log z}{a \log \phi}\right) \\ &= \frac{(a-1)(a-2)}{2a}. \ \blacksquare \end{aligned}$$

**Lemma B** *Let* $\varepsilon_\phi \equiv \frac{\phi}{\phi - 1} - \frac{1}{\log \phi} - \frac{1}{2}$. *Then,* $\lim_{\phi \to 1} \frac{\varepsilon_\phi}{\log \phi} = \frac{1}{12}$.

**Proof of Lemma B**: The proof follows directly from repeat applications of L'Hopital's rule. ■

**Lemma C** $\lim_{\phi\to 1}\frac{d\mathbf{E}[Q]}{d\phi}$ *exists in* $\mathbb{R}$.

**Proof of Lemma C**: For $\phi \neq 1$, we have

$$\begin{aligned}\frac{d\mathbf{E}[Q]}{d\phi} &= \frac{d}{d\phi}\left(\frac{\phi^2}{1-\phi} + \frac{\log z}{a(1-\phi)\log\phi}\right)\\ &= \frac{2\phi-\phi^2}{(1-\phi)^2} + \frac{-a/z}{a(1-\phi)\log\phi} - \frac{\log z}{a(1-\phi)^2(\log\phi)^2}\left(-\log\phi + \frac{1-\phi}{\phi}\right)\\ &= -1 + \frac{1}{(1-\phi)^2} - \frac{1}{z(1-\phi)\log\phi} - \frac{\log z}{a\phi(1-\phi)\log\phi}\left(\frac{3\phi+1}{2(\phi-1)} - \varepsilon_\phi\right).\end{aligned}$$

It follows from $\lim_{\phi\to 1}\frac{\log z}{1-\phi} = a$ and Lemma B that

$$\lim_{\phi\to 1}\frac{\varepsilon_\phi \log z}{a\phi(1-\phi)\log\phi} = \frac{1}{a\phi}\frac{\log z}{1-\phi}\frac{\varepsilon_\phi}{\log\phi} = \frac{1}{12}.$$

Therefore,

$$\lim_{\phi\to 1}\frac{d\mathbf{E}[Q]}{d\phi} = -\frac{11}{12} + \lim_{\phi\to 1}\left(\frac{\log\phi}{1-\phi}\right)^2 \cdot \lim_{\phi\to 1} h(\phi) = -\frac{11}{12} + \lim_{\phi\to 1} h(\phi),$$

where

$$h(\phi) \equiv \frac{1}{(\log\phi)^3}\left[\log\phi - \frac{1-\phi}{z} + \left(\frac{3\phi+1}{2a\phi}\right)\log z\right].$$

Using L'Hopital's rule repeatedly,

$$\begin{aligned}\lim_{\phi\to 1} h(\phi) =& \lim_{\phi\to 1}\frac{\phi}{3(\log\phi)^2}\left[\frac{1}{\phi} + \frac{1}{z^2} + \frac{-2a}{4a^2\phi^2}\log z + \frac{3\phi+1}{2a\phi}\frac{(-a)}{z}\right]\\ =& \frac{2a^2+(3/2)a-1}{6}.\end{aligned}$$

Therefore, $\lim_{\phi\to 1}\frac{d\mathbf{E}[Q]}{d\phi}$ exists in $\mathbb{R}$, which proves Lemma C. ■

By Lemma A, $\mathbf{E}[Q]$ is continuous at $\phi = 1$. It is also differentiable at every $\phi \neq 1$. Then, the Mean Value Theorem implies that, for any $\phi \neq 1$, there exists $x_\phi \in (1,\phi)$ such that $\frac{\mathbf{E}[Q](\phi)-\mathbf{E}[Q](1)}{\phi-1} = \frac{d\mathbf{E}[Q](x_\phi)}{d\phi}$. As $\phi \to 1$, we have $x_\phi \to 1$. Hence, by Lemma C, $\frac{d\mathbf{E}[Q](1)}{d\phi} \equiv \lim_{\phi\to 1}\frac{\mathbf{E}[Q](\phi)-\mathbf{E}[Q](1)}{\phi-1} = \lim_{x_\phi\to 1}\frac{\mathbf{E}[Q](x_\phi)}{d\phi}$ exists in $\mathbb{R}$. That is, $\mathbf{E}[Q]$ is continuously differentiable at $\phi = 1$, which concludes the proof of Step 1.

*Step 2: $E[Q]$ is strictly convex.*

To show Step 2, we first present the following lemmas D-G.

**Lemma D** *The function* $r(x) = \frac{x}{\log(1-x)}$ *is increasing in* $x$ *and strictly convex on* $(-\infty, 0)$ *and* $(0, 1)$.

**Proof of Lemma D**: Derivating,

$$r'(x) = \frac{1}{\log(1-x)} + \frac{x}{(1-x)\log^2(1-x)} = \frac{1}{\log^2(1-x)}\left(\log(1-x) + \frac{x}{1-x}\right) > 0.^{25}$$

Derivating again,

$$\begin{aligned} r''(x) &= \left(\frac{1}{\log(1-x)} + \frac{x}{(1-x)\log^2(1-x)}\right)' \\ &= \frac{(2-x)\log(1-x) + 2x}{(1-x)^2\log^3(1-x)}. \end{aligned}$$

If $x < 0$, we have $(2-x)\log(1-x) + 2x > 0$, so $r''(x) > 0$. If $0 < x < 1$, we have $g(x) \equiv \log(1-x) + \frac{2x}{2-x} < 0$, because $g(0) = 0$ and $g'(x) = \frac{-1}{1-x} + \frac{4}{(2-x)^2} = \frac{(-x^2+4x-4)+4(1-x)}{(1-x)(2-x)^2} < 0$. Thus, $r''(x) > 0$. ■

**Lemma E** *The function* $(a(\phi+1)-1)\frac{\phi-1}{\log(a(1-\phi)+1)}$ *is strictly convex on* $[\underline{\phi}, 1)$ *and* $(1, 1+1/a)$.

**Proof of Lemma E**: Let $r(\phi) = a(\phi+1) - 1$ and $g(\phi) = \frac{\phi-1}{\log(a(1-\phi)+1)}$. Note that $a\phi \geq 1$ because of the training constraint $a(1-q)\phi = 1$. Thus, $r(\phi) > 0$. Moreover, Lemma D implies that $g$ is increasing and strictly convex (using $x = a(\phi - 1)$). Therefore,

$$(r(\phi)g(\phi))'' = r''(\phi)g(\phi) + 2r'(\phi)g'(\phi) + r(\phi)g''(\phi) \geq r(\phi)g''(\phi) > 0. \blacksquare$$

**Lemma F** *The function* $r(\phi) = \frac{1}{\log(a(1-\phi)+1)} + \frac{1}{a\log\phi}$ *is strictly convex on* $[\underline{\phi}, 1)$ *and* $(1, 1+1/a)$.

**Proof of Lemma F**: Recall that $z \equiv a(1-\phi)+1$, and that $z' \equiv \frac{dz}{d\phi} = -a$. The second derivative of $r(\phi)$ follows from

$$\begin{aligned} &\left(\frac{1}{\log\phi}\right)'' = -\left(\frac{1}{\phi\log\phi}\right)' = \frac{1}{\phi^2\log^2\phi} + \frac{2}{\phi^2\log^3\phi}, \quad \text{and} \\ &\left(\frac{1}{\log(a(1-\phi)+1)}\right)'' = a^2\left(\frac{1}{\log z}\right)'' = a^2\left(\frac{1}{z^2\log^2 z} + \frac{2}{z^2\log^3 z}\right). \end{aligned}$$

[25] For any $y \neq 1$, $-\log y < \frac{1}{y} - 1$, which implies that $\log y > \frac{1-y}{y}$. We substitute $1-x$ for $y$.

Thus,

$$
\begin{aligned}
r''(\phi) &= \frac{a^2}{z^2 \log^2 z}\left(1+\frac{2}{\log z}\right)+\frac{1}{a\phi^2\log^2\phi}\left(1+\frac{2}{\log\phi}\right) \\
\Longrightarrow (\log^2\phi) r''(\phi) &= \frac{a^2}{z^2}\left(\frac{\log\phi}{\log z}\right)^2\left(1+\frac{2}{\log z}\right)+\frac{1}{a\phi^2}\left(1+\frac{2}{\log\phi}\right).
\end{aligned}
$$

We make two claims:

*Claim 1*: $-\frac{\log\phi}{\log z} > 1$ *for every* $\phi \neq 1$.

*Proof of Claim 1:* By the training constraint, $a(1-q)\phi = 1$, so $a\phi \geq 1$. Hence, if $\phi > 1$, then $(a\phi)(1-\phi) > (1-\phi)$ and $\phi > \frac{1}{a(1-\phi)+1}$. If $\phi < 1$, then $(1-\phi) > (a\phi)(1-\phi)$, so $\frac{1}{\phi} > a(1-\phi)+1$.

*Claim 2:* $\lim_{\phi\to 1} -\frac{\log z}{\log\phi} = a$ *and is increasing in* $\phi$.

*Proof of Claim 2:* First, $\lim_{\phi\to 1} -\frac{\log z}{\log\phi} = \lim_{\phi\to 1}\frac{a/(a(1-\phi)+1)}{1/\phi} = a$. Now, define for any $\phi \neq 1$

$$
h(\phi) \equiv \phi\left(-\frac{\log z}{\log\phi}\right)' = \frac{a\phi\log\phi}{a(1-\phi)+1} + \log(a(1-\phi)+1).
$$

Then,

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

which is strictly negative for $\phi < 1$ and strictly positive for $\phi > 1$. Thus, for any $\phi \neq 1$, $h(\phi) > \lim_{\phi\to 1} h(\phi) = 0$, which concludes the proof of Claim 2.

If $\phi \in [\underline{\phi}, 1)$, then $\log\phi < 0$, $\log z > 0$, and

$$
\begin{aligned}
(\log^2\phi)(\log x) r''(\phi) &= \frac{a^2}{z^2}\left(\frac{\log\phi}{\log z}\right)^2(\log z + 2) + \frac{1}{a\phi^2}\left(\log z + \frac{2\log z}{\log\phi}\right) \\
&> \left(\frac{a^2}{z^2}+\frac{1}{a\phi^2}\right)\log x + 2\left(\frac{a^2}{z^2}+\frac{1}{a\phi^2}\frac{\log z}{\log\phi}\right) \\
&> \left(\frac{a^2}{z^2}+\frac{1}{a\phi^2}\right)\log x + 2\left(\frac{a^2}{z^2}-\frac{1}{\phi^2}\right),
\end{aligned}
$$

where the first inequality follows from Claim 1 and the second from Claim 2. Since $\phi \geq \underline{\phi} = \frac{1+\sqrt{1+4a}}{2a}$ implies $(a\phi)^2 - z^2 = (a\phi)^2 - (a(1-\phi)+1)^2 = (2a\phi - a - 1)(a+1) > 0$, we obtain $r''(\phi) > 0$.

If $\phi \in (1, 1 + 1/a)$, then $\log \phi > 0$, $\log z < 0$, and

$$(\log^2 \phi)\phi^2 r''(\phi) = \left(\frac{a\phi \log \phi}{z(-\log z)}\right)^2 \left(1 + \frac{2}{\log z}\right) + \frac{1}{a}\left(1 + \frac{2}{\log \phi}\right).$$

Suppose that $1 + \frac{2}{\log z} < 0$, as for otherwise it is clear that $r''(\phi) > 0$. It must be that $a\phi \log \phi + z \log z > 0$ since the left-hand-side is zero at $\phi = 1$, and the derivative $a(\log \phi + 1) - a(\log z + 1) = a(\log \phi - \log z) > 0$. Thus, by Claim 1 above,

$$(\log^2 \phi)\phi^2 r''(\phi) > \left(1 + \frac{2}{\log z}\right) + \frac{1}{a}\left(1 + \frac{2}{\log \phi}\right) = 1 + \frac{1}{a} + 2\left(\frac{1}{\log z} + \frac{1}{a \log \phi}\right) > 0,$$

and $r''(\phi) > 0$, which concludes the proof of Lemma F. ■

**Lemma G** *Suppose that $r : \mathbb{R} \to \mathbb{R}$, is positive and decreasing in $x$ and satisfies $g(x)r(x) = h(x)$, with $h(x)$ strictly convex and $g(x)$ strictly positive, concave, and decreasing in $x$. Then, $r(x)$ is strictly convex.*

**Proof of Lemma G**: Take any $x_1, x_2 \in \mathbb{R}$ and $\bar{x} = \beta x_1 + (1 - \beta)x_2$ for some $\beta \in (0, 1)$. Then,

$$\begin{aligned}
&g(\bar{x}) \cdot (\beta r(x_1) + (1 - \beta)r(x_2)) \\
&\geq (\beta g(x_1) + (1 - \beta)g(x_2)) \cdot (\beta r(x_1) + (1 - \beta)r(x_2)) \\
&= \beta^2 g(x_1)r(x_1) + (1 - \beta)^2 g(x_2)r(x_2) + \beta(1 - \beta)(g(x_1)r(x_2) + g(x_2)r(x_1)),
\end{aligned}$$

where the first inequality is guaranteed by the concavity of $g$. Since $r(x)$ is increasing and $g(x)$ is decreasing in $x$,

$$g(x_1)r(x_2) + g(x_2)r(x_1) \geq g(x_1)r(x_1) + g(x_2)r(x_2).$$

Thus,

$$\begin{aligned}
g(\bar{x}) \cdot (\beta r(x_1) + (1 - \beta)r(x_2)) &\geq \beta g(x_1)r(x_1) + (1 - \beta)g(x_2)r(x_2) \\
&= \beta h(x_1) + (1 - \beta)h(x_2) > h(\bar{x}) = g(\bar{x})r(\bar{x}),
\end{aligned}$$

which implies $\beta r(x_1) + (1 - \beta)r(x_2) > r(\bar{x})$. ■

We are now ready to show that

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

where the value at $\phi = 1$ is given by $\lim_{\phi\to 1}\mathbf{E}[Q]$, which is strictly convex in $\phi$.

For any $\phi \in [\underline{\phi}, 1) \cup (1, 1+1/a)$, we have

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

By Lemma D, $\frac{1-\phi}{\log(a(1-\phi)+1)}$ is decreasing and concave. By Lemmas E and F, the right-hand side is strictly convex. It follows from Lemma G that $\mathbf{E}[Q]$ is strictly convex on $[\underline{\phi}, 1)$ and $(1, 1+1/a)$. Last, by Lemmas A and C, we know that $\mathbf{E}[Q]$ is continuously differentiable at $\phi = 1$, which completes the proof of Step 2, and therefore of Proposition 3. ■

## 8.4 Proofs for Impacts of Centralization

**Proof of Corollary 3**: Consider (1) and (2) in Proposition 1 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 $(q_P^*, \mu_P^*)$ determines $\phi_P^* = \lambda/\mu_P^*$ and $k_P^* = k(q_P^*; \mu_P^*, \lambda)$ by (7). The proof of Corollary 3 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 (6) and the first-order condition of Proposition 3 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 $(q, \phi)$, with $\phi = \frac{1}{a(1-q)}$, a function of $q$. Then, the first-order condition (10) corresponds to

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

Observe that $\frac{d\phi}{dq} > 0$, and $\frac{\partial\mathbf{E}[Q]}{\partial\phi} > 0$ because an increase of $\phi \equiv \frac{\lambda}{\mu}$, while holding $q, a$, and $\lambda$

fixed, corresponds to a decrease in $\mu$, which increases $\mathbf{E}[Q]$. Thus, 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 (7).

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 (7), 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 $\mu_P^*\theta + 1 > k_P^*$. Hence, $(q_P^*, \mu_P^*)$ lies on the left-hand side of the graph $G_2$ defined in the proof of Proposition 3. It follows that $q_P^* < q_P^e$ and $\mu_P^* > \mu_P^e$. ■

## 8.5 Proofs for Impacts of Monitoring

**Proof of Lemma 4:** For any choice of $(q, \mu)$, 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 (7).

If $\lambda = \mu$, then $\frac{\mathbf{E}[W_P]}{\mathbf{E}[W_L]} = \frac{2q-1}{2q} < 1$. If $\lambda \neq \mu$, $\phi \neq 1$ and $\mu\mathbf{E}[W_L] = \frac{q\phi}{1-q\phi}$. Since $k \log\phi = \log\left(\frac{1-q}{1-q\phi}\right)$, or equivalently $\phi^k = \frac{1-q}{1-q\phi}$, we have

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

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

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

and the last inequality holds for every $\phi \neq 1$. The expression's minimum is 0 at $\phi = 1$. ■

**Proof of Proposition 4A:** By Propositions 1 and 3, $(q_L^e, \mu_L^e)$ solves $\theta = \frac{q\lambda}{\mu(\mu-q\lambda)}$ and $\mu = f((1-q)\lambda)$, while $(q_P^e, \mu_P^e)$ 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

$$\begin{aligned} \frac{q\lambda}{\mu(\mu - q\lambda)} > \frac{k(q,\mu;\lambda) - 1}{\mu} &= \frac{1}{\mu}\left(\frac{1}{\log(\lambda/\mu)}\log\left(\frac{1-q}{1-(q\lambda/\mu)}\right) - 1\right) \\ \iff \frac{q\lambda}{\mu - q\lambda} &> \frac{1}{\log(\lambda/\mu)}\left(\frac{1-q}{1-(q\lambda/\mu)} - 1\right) - 1 \quad (\forall x \neq 1, \log x < x - 1) \\ \iff \frac{(\lambda/\mu)\log(\lambda/\mu)}{(\lambda/\mu) - 1} &> \frac{q\lambda}{\mu}, \end{aligned}$$

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 $\overline{q} > q''_P$.[26] Therefore, $q''_L > \overline{q} > q''_P$. ■

**Proof of Proposition 4B:** We compare the solutions $(q^*_L, \mu^*_L)$ and $(q^*_P, \mu^*_P)$ in terms of $\phi \equiv \frac{\lambda}{\mu} = \frac{1}{a(1-q)}$. The steady-state constraint $q\lambda < \mu$ implies that $q\phi = (1 - \frac{1}{a\phi})\phi = \phi - \frac{1}{a} < 1$. Hence, $\phi \in [\frac{1}{a}, 1 + \frac{1}{a})$. In the proof of Proposition 3, we wrote the planner's problem under perfect monitoring as

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

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

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

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

[26] 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 $q$.

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 $\phi$ and $\frac{-1}{\log(a(1-\phi)+1)}\left(1 + \frac{1-\phi}{\log\phi}\right)$ is increasing in $\phi$.*

**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 $\phi$.*

**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 $\phi$ (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 $\phi$, which concludes the proof of Lemma I. ■

We are now ready to show Proposition 4B. 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 3, 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.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan, 2020, "Thickness and Information in Dynamic Matching Markets," Journal of Political Economy, 128(3), 783-815.

Anderson, Ross, Itai Ashlagi, David Gamarnik, and Yash Kanoria, 2017, "Efficient Dynamic Barter Exchange," Operations Research, 65(6), 1446-1459.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv, 2020, "Optimal Dynamic Matching," Theoretical Economics, 50, 1221-1278.

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.

Doval, Laura and Balász Szentes, 2019, "On the Efficiency of Queueing in Dynamic Matching Markets," mimeo.

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.

Ferdowsian, Andrew, Muriel Niederle, and Leeat Yariv, 2020, "Discretionary Matching with Aligned Preferences," mimeo.

Fudenberg, Drew, and Luis Rayo, 2019, "Training and Effort Dynamics in Apprenticeship," American Economic Review, 109(11), 3780-3812 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.

Haeringer, Guillaume and Myrna Wooders, 2011, "Discretionary Job Matching," International Journal of Game Theory, 40, 1-28.

Herbst, Holger and Benjamin Schickner, 2016, "Dynamic Formation of Teams: When Does Waiting for Good Matches Pay Off," mimeo.

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, 2019, "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.

Loertscher, Simon, Ellen V. Muir, and Peter G. Taylor, 2019, "Optimal Market Thickness and

Clearing," mimeo.

Margaria, Chiara, 2020, "Queueing to Learn," mimeo.

Ortoleva, Pietro, Evgenii Safonov, and Leeat Yariv, 2020, "Who Cares More? Allocation with Diverse Preference Intensities," mimeo.

Pais, Joana V., 2008, "Incentives in Discretionary Random Matching Markets," Games and Economic Behavior, 64, 632-649.

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.

Zenios, Stefanos A, 1999, "Modeling the Transplant Waiting List: A Queueing Model with Reneging," Queueing Systems, 31, 239-251.