Agents self-routing to balance inbound and outbound services

Abstract

We analyze a service system with inbound and outbound customers. Inbound customers arrive over time depending on the offered waiting time while outbound customers can be contacted at all times. The agents are in control of the routing decisions. Knowing system state, they decide whether to serve an inbound customer, an outbound one, or to idle. The system manager seeks to provide a good trade-off between the performance of inbound and outbound customers by incentivizing the agents' actions through linear payouts. Our aim is to evaluate the cost of agents' self-routing. This problem is a novel variant of the principal–agent problem where the agents' effort consists of selecting their routing policy. From a Markov decision process, we show that the optimal policy for the agents is a reservation threshold policy for inbound customers and we express the compensation parameters that minimize the staffing cost.

We conclude that motivating idling decisions through linear payouts is very costly with respect to the improvement provided for inbound customers. This justifies the current practice at most call centers of using automated routing. Moreover, paying for idling cannot reduce staffing cost. Nevertheless, discriminating between delayed and non-delayed customers in the reward structure has a high potential to reduce the agents' pay. Finally, when agents do not know the status (idle or busy) of their colleagues, our analysis argues in favor of not revealing the system state to the agents through delay announcements when the objective waiting time for inbound customers is low.
Keywords: Blended queue; incentives; reservation; self-routing.

1 Introduction

Many organizations, such as call centers and hospitals, face highly unpredictable demand that often results in long waits for customers or long idling times for agents. To improve the level of customer service, alleviate congestion, and to make a better use of the available resources, these systems face challenging routing issues. When assigning customers to agents, firms must decide on performance objectives, priority rules, queue discipline, and agent selection while considering staffing cost, service urgency, demand forecast, agent skills and preferences, and past performances. The complexity of routing problems has led to a large number of studies, particularly in the field of call centers (Aksin et al., 2007). The underlying assumption in these studies is the ability of the firm to tell agents when and on which task to work. One advantage of exercising control over agents is its simplicity. By generating automatic decisions, employees do not waste time rationalizing their actions. This approach is also supported by the idea that employees can only make decisions that are beneficial in the short-term, and unable to make decisions that consider long-term goals.

Nevertheless, agents do not necessarily adhere to their assigned tasks. This generates human resource issues like turn-over and absenteeism (Hillmer et al., 2004). Alternatively to this practice, some firms allow agents to decide when to work. This is called self-scheduling. It results in better morale and improved work–life balance (Koning, 2014). In service-based organizations, such as ride-sharing services, self-scheduling is increasingly used. In hospitals, nurses often choose self-scheduling (Hung,

2002; Bailyn et al., 2007; Rönnberg and Larsson, 2010). In work-from-home call centers, like LiveOps, agents can decide when they wish to be available (Stouras et al., 2014; Gurvich et al., 2019; Brunelli, 2020). Although increasingly used, self-scheduling is often restricted to the decision of when to work and does not encompass the decision of on which task to work. This is the case in work-from-home call centers where the routing of customers remains under external control. Letting agents be in charge of customer selection for service is nevertheless a trend as it allows agents to take a responsible and active role in the company and could liberate the system manager from the burden of exercising control. Without direct control of customers selection, the system manager relies on payouts to incentivize the agents' actions. Therefore, asking agents to be in charge with the routing of customers to service, in the so-called the strategy of self-routing, has a cost for the service system. The aim of this paper is to evaluate this cost and to determine the efficiency of agents' self-routing.

To this end, we consider the blended queue model with inbound and outbound customers of Gans and Zhou (2003) and Bhulai and Koole (2003). Inbound customers arrive randomly over time, adjust their joining strategy to the offered waiting time and are urgent while outbound customers can be contacted at any point in time and are non-urgent. Moreover, the agents can serve both inbound and outbound customers. This flexibility allows the system to provide a good service quality for the two classes of customers although it results in higher training costs than when agents are specialized in one type of service (Echchakoui, 2016). The blended queue well models multi-channel call centers with inbound and outbound calls. CarFinance247, for instance, the UK's largest online car finance broker, helps customers to find cars online and to get approval for financing by utilizing both inbound and outbound channels. The economic importance of blended contact centers has been demonstrated over the last few decades by the high number of patents specifically devoted to staffing and scheduling issues (Dumas et al., 1996; Villena et al., 2004; Anisimov et al., 2017). A lack of resources also led hospital emergency departments to adopt blended strategies for the available resources (operating rooms, nurses, anesthesiologists, and surgeons) to treat both elective patients (i.e., outbound patients, already present or scheduled in advance) and non-elective patients (i.e., inbound patients arriving randomly over time).

Due to the competition between the two classes of customers for the same resource, setting an efficient payout contract to incentivize the agents' actions is complex. For instance if inbound and outbound services are equally rewarded, agents could be tempted to initiate outbound services at all times to maximize their profit. However, if all agents are constantly working, all inbound customers will be delayed, leading to a poor quality of service for these customers. This can be avoided if a

sufficient number of agents remains available for the service of inbound customers. Thus, the challenge in setting the payouts for inbound and outbound services is not only in inducing a priority rule between the two classes of customers but also inducing a reservation policy where some agents could choose to remain idle for a future inbound customer service instead of initiating an outbound service.

We analyze the issue of agents self-routing in this context using a principal–agent framework with moral hazard and linear compensation (Laffont and Martimort, 2009) in the first- and second-best settings. We assume that the principal (the system manager) wants to minimize the long-run expected staffing cost of inducing the agents to choose the reservation policy which provides the optimal trade-off between the rate of served outbound customers and the expected waiting time of served inbound customers. The agents know the state of the system and coordinate their effort to select the reservation policy that maximizes their expected utility defined as the difference between their revenue and their effort-cost.

As a first step of the analysis, we evaluate the parameters of the state-dependent equilibrium joining strategy of customers. In the second-best setting, it allows us to develop a Markov decision process to derive the optimal reservation policy. In the case of highly wait-averse customers, we prove that the optimal reservation policy is a deterministic threshold one. This result indicates that self-routing cannot induce the first-best effort as the optimal policy in the first-best setting is non-deterministic and randomizes in between adjacent threshold levels. In a context with equal service rates between inbound and outbound customers, we prove that there exists a unique local maximum of the agents' utility and that the reservation level increases with the reward for serving an inbound customer. This in turn enables us to express the compensation parameters which minimize the staffing cost in the first- and second-best settings.

Our numerical investigations reveal that the cost of achieving a low waiting time for inbound customers can grow extremely high in the second-best setting as compared to the first-best one. In support of this observation, we prove the convexity of the variable part of the staffing cost in the objective waiting time for inbound customers. The high staffing cost with linear piece rate compensation is mainly due to the high cost for serving an inbound customer as this one should compensate agents for the idling time before service. This result justifies the current situation in call centers where routing decisions are automated by the system. Even if self-routing may be more satisfying from a human-resources management perspective, its cost precludes implementation.

Thus, we question whether the linear piece rate compensation model could be modified to reduce

the staffing cost. For the purpose of reducing the payout of inbound customers, we consider the possibility of paying the agents to idle before service as an extra motivation to remain available for inbound customers. However, we prove that this strategy does not reduce the staffing cost and should consequently be excluded. Next, since the non-delayed inbound customers are only those who benefit from the agents' reservation, we consider the possibility to discriminate between the delayed and the non-delayed customers in the agents' pay. In this way, the customers' service level influences the reward structure. This modification of the payment contract significantly reduces the staffing cost, suggesting a stronger potential for implementation, especially when the objective waiting time for inbound customers is low. Another advantage of this change is that the staffing cost is close to being a constant function of the objective waiting time for inbound customers. Thus, for the system manager, when selecting a waiting time objective for inbound customers, the resulting cost is no longer an issue.

One assumption of our analysis is that agents can observe the state of the system and may make decisions based on this observation. However, in virtual systems like call centers, agents may not know the status (idle or busy) of their colleagues. To account for such systems, we investigate the non-observable case for the agents. We show that the optimal policy is a randomizing policy and that there exists a unique maximum of the agents' utility in the randomizing parameter. Next, as in the observable case, we express the reward parameters in the second-best setting. Although non-observable by nature, call centers can be made observable to the agents by announcing their idling delays before receiving an inbound customer. This possibility allows us to question whether announcing delays to agents is profitable to the system. Our analysis argues in favor of announcing delays to the agents only when the objective waiting time for inbound customers is high. On the contrary, not announcing delays has the potential to reduce the staffing cost when the objective waiting time is low as agents would then make decisions based on the average idling duration and not on the actual and potentially long expected idling duration.

Structure of the article. We end this section with a review of the literature. Section 2 formulates the optimzation problem. Section 3 evaluates the cost of self-routing for reservation policies. Section 4 analyzes extensions of the compensation model. Section 5 considers a context where agents do not know the state of the system and questions the opportunity of announcing idling delays to agents. Finally, Section 6 concludes the paper. The proofs of the main results and the closed-form expressions of the performance measures in the case of equal service rates are provided in the appendix.

Literature review. As our study relates to the analysis of compensation design, we first present studies in this field from non-queueing contexts. Next, in relation to the queueing model considered in this paper, we detail the literature on blended queues. As customers' and agents' decision actions are endogenous in our study, we then briefly describe important references on customers' joining decisions and elaborate on agents' service speed selection and agents' self-scheduling problems.

There is a long history of compensation analyses in the fields of marketing, economics, heath care, and operations management (e.g., Lal and Srinivasan (1993); Herweg et al. (2010); Jain (2012); Chen et al. (2016); Suen et al. (2018); Li et al. (2020)). This literature stream focuses on linear commission and quota-bonus contracts (i.e., the employee receives a bonus for meeting a performance quota). Jain (2012) showed that firms can reduce the negative consequences of self-control by employing multiperiod quotas (such as annual quotas) to compensate employees for their cumulative performance instead of receiving direct rewards. In a producer–seller relationship, Chen et al. (2016) compared forecast-based and linear compensation contracts. They showed that with an endogenous information-acquisition effort, forecast-based contracts can outperform linear compensation ones. Suen et al. (2018) described how payouts can be employed to induce a socially-optimal behavior in a context of patients' adhesion to antibiotic therapy and showed the inefficiency of linear payouts. Li et al. (2020) also compared linear and non-linear contracts and showed that the feature of fairness plays a role in the potential outcomes realized, leading to a reduction in the benefits of non-linear contracts. Meanwhile, Long and Nasiry (2020) discussed contexts where making wages transparent to employees were beneficial to the firms. The incentive-design issue becomes more complicated with multitasking agents. As in our work, Dai et al. (2021) considered a principal–agent framework where the agent can exercise two types of tasks, operational and marketing. They characterized the optimal compensation plan, where a bonus is paid when either all the inventory above a threshold is sold or the sales quantity meets an inventory-dependent target. In this paper, we investigate linear and performance-based payouts used as a tool to induce a reservation policy. Our analysis shows a novel context where performance-based payouts are more efficient than linear ones.

The focus of the literature on blended queues with inbound and outbound customers is on performance evaluation, staffing, and routing decisions, primarily for applications in call centers. By analyzing various continuous time Markov chains, Deslauriers et al. (2007) demonstrated the value of a threshold reservation policy. The authors showed that by reserving some agents for inbound customers, the system can achieve a good trade-off between the rate of served outbound customers and the waiting

time distribution of inbound customers. Later, in the case of identical service rates for inbound and outbound customers, Gans and Zhou (2003) and Bhulai and Koole (2003) provided formal proofs that this threshold type reservation policy is optimal for maximizing the rate of served outbound customers with a service level constraint on the inbound customers' waiting time. In this paper, we extend their proofs to the case of unequal service rates and highly wait-averse customers. Pang and Perry (2014) investigated a large call blending model and proposed a logarithmic safety staffing rule, combined with a threshold reservation policy which manages simultaneously having agents' utilization close to one with idle agents almost always present. Extensions of the blended queue model with reservation are investigated with time-dependent parameters (Legros et al., 2015), retrials (Phung-Duc et al., 2016), reservation for arriving customers where delayed ones are viewed as outbound ones (Legros, 2017), or in combination with outsourcing decisions in a sales environment (Legros et al., 2021). In the above references, reservation is controlled by the system manager. Instead, this paper shows that reservation can be the outcome of agents' individual decisions motivated by payouts.

There is a large body of literature on queueing games models to capture customers' joining decisions, including Gavirneni and Kulkarni (2016), Dimitrakopoulos and Burnetas (2016), Cui and Veeraraghavan (2016), and Hassin and Roet-Green (2017). As in this paper, this literature stream focuses on the impact of customers acting strategically (i.e., deciding whether to join or to balk) when trying to obtain the best trade-off between the value of a service and the cost of waiting. Hassin and Haviv (2003)'s book explains the main principles of decision making from the customers' perspectives. We follow their approach based on the expected waiting time to build a simple utility model to determine the customers' joining parameters.

Customers' joining problem can be viewed as an arrival rate optimization question. The symmetrical problem from the agents' perspective is the service speed selection. Fewer references are found in this area compared to the customers' joining literature. Due to the mathematical complexity of obtaining structural results, most analyses are made for systems with fewer than two servers. In the single-server case, Zhan and Ward (2018) considered a queue with abandonment. They proved that there exists a unique maximum of the agent's utility defined as the product of the value of the service speed multiplied by the agent's utilization rate. In a principal–agent framework similar to ours, Baiman et al. (2010) studied a single server queue with finite capacity where the principal controls the payment parameters and the buffer size while the agent decides for their service speed. They showed that decreasing the buffer size may exacerbate or mitigate the agent's moral hazard problem, depending on

the level of blocking in the system. In the two-server case, Kalai et al. (1992) studied a queue with competing exponential servers and Poisson arrivals. They found that in situations where the expected waiting time is finite, there exists a unique symmetric strategic equilibrium. Christ and Avi-Itzhak (2002) extended this model to a situation with a state-dependent Poisson arrival process, showing that when the cost function is increasing and convex, there exists a unique pure symmetric Nash equilibrium strategy. For the same model, Avi-Itzhak et al. (2006) showed that the unique Nash equilibrium is generally strictly inferior to a globally optimal solution. Geng et al. (2015) also considered a queue with two servers seeking fairness and proved the existence and uniqueness of the Nash equilibrium for some routing policies. In the multi-server case, Gopalakrishnan et al. (2016) explored the effect of routing rules when each agent can select the service rate which maximizes a trade-off between the effort-cost and the proportion of idling time. They showed that with fair routing disciplines, all agents adopt the same service speed. Zhan and Ward (2019) further investigated the multi-server case to find a joint staffing, routing, and payment policy that leads to the optimal service-system performance. By solving the centralized control problem under fluid scaling, they found that critically loaded, efficiency driven, quality driven, and intentional idling regimes were economically optimal. In contrast to the above references, the agents' speed of service is exogenous in this paper, but the routing decisions are decided by the agents.

Another field of research where agents partially exert control over their work activities is the analysis of self-scheduling issues. Self-scheduling is in line with a growing stream of literature on the management of on-demand service platforms (Cachon et al., 2017; Taylor, 2018; Bimpikis et al., 2019; Braverman et al., 2019; Hu and Zhou, 2019; Özkan and Ward, 2020). In a queueing context, Ibrahim (2018) studied the challenges of staffing and controlling queues with an uncertain number of servers and impatient customers. The author showed how managers can use three forms of control from their toolbox, namely, staffing, compensation, and announcements, to effectively control their system. Gurvich et al. (2019) studied capacity management when workers self-schedule, when the firm controls its capacity indirectly through compensation. They showed that to guarantee an adequate capacity, the firm must offer a high compensation to their agents.

In contrast, the number of agents in our study is fixed, but their availability for either inbound or outbound customers is self-determined. In this way, our analysis can be defined as a self-routing issue in contrast to the aforementioned self-scheduling problems. We mention Lu et al. (2009) who also investigated a self-routing problem in a principal–agent setting with a different focus than ours. The

authors considered a queueing model with rework routing. They compared the incentives of different allocation schemes and showed that self-routing of rework will never induce the first-best effort. In the context of reservation strategy, we also make the conclusion that self-routing cannot induce the first-best effort. Nevertheless, the authors showed that with a large capacity, dedicated routing and cross routing can both achieve the first-best profit rate.

2 Problem formulation

We analyze a blended queue with a team of s homogeneous agents who can serve two types of customers, namely, inbound and outbound. We refer to inbound customers as class-1 customers, and to outbound ones as class-2 customers. Class-1 customers arrive at the system according to a Poisson process with rate λ. If class-1 customers are not served immediately upon arrival, then they wait in an infinite capacity queue before being served, under a first-come-first-served queue discipline. We assume that there is an infinite supply of class-2 customers. The idea of this simplifying assumption is to consider the number of accessible customers to contact as very large, such that an agent can find an available customer to contact at any point in time. The service time of class-i customers ($i = 1, 2$) is assumed to be exponentially distributed with rate μ_i and service preemption is not permitted. When the service rates are equal (i.e., $\mu_1 = \mu_2$), we omit the index i, so the service rate is denoted by μ. In this case, the offered load for class-1 customers, a, is defined by $a = \frac{\lambda}{\mu}$.

Class-1 customers' joining decisions is determined by the utility model of Naor (1969) defined as the difference between a reward for service and a cost proportional to the waiting time. In this model customers are risk-neutral, that is, they maximize the expected value of their utility. Specifically, the customers' reward from completed service is R_S and the cost for staying in the queue is C_W per unit of waiting time. Moreover, the system is directly observable by customers or is made observable through delay announcements. Thus, the expected waiting time at arrival can be estimated by customers. If the expected waiting time of a given customer at arrival, called Customer n, is $E(W_n)$, the expected utility of joining the queue is then $R_S - C_W E(W_n)$. Given that the utility of balking is zero, Customer n joins the queue if $R_S - C_W E(W_n) > 0$. Due to the first-come-first-served discipline, the remaining expected waiting time of a customer who joins the queue reduces over time. Thus, the expected utility increases over time. Consequently, the joining decision is irrevocable, and reneging does not happen.

The system manager and the agents agree to a contract, which will govern their employment relation. As in Baiman et al. (2010), we consider a compensation per agent and per time unit which

consists of a fix wage F plus a piece-rate wage based on the realized rate of served class-1 and class-2 customers. Specifically, serving a class-i customer is rewarded with $r_i > 0$ for $i = 1, 2$. The system manager has discretion in setting r_1, r_2, and F. We call this compensation model the base model. The resulting expected staffing cost, SC, is then given by $SC = s(r_1T_1 + r_2T_2 + F)$, where T_i is the expected rate of served class-i customers by an agent for $i = 1, 2$.

We call by reservation policy the function which associates a decision action among idling, serving a class-1 or serving a class-2 customer, to each agent and to each state of the system. For a given reservation policy, termed Policy π, we associate an obedient effort per agent, P_π, which corresponds to the expected proportion of time spent on serving customers. We assume that the individual effort cost EC_π of employing Policy π is proportional to P_π: $EC_\pi = e \times P_\pi$ with $e > 0$. With this assumption, we state that the effort per time unit of work is identical when serving a class-1 or a class-2 customer. We then define the agents' expected utility of employing Policy π as the difference between the revenue and the effort-cost: $U = s(r_1T_1 + r_2T_2 + F - EC_\pi)$.

The agents have control over the reservation policy. The reservation policy is thus non-contractible and subject to moral hazard. Furthermore, we assume that agents can observe the system state and coordinate their decisions as a way to maximize their expected utility. As such, agents are risk-neutral. For the reason of fairness, the longest-idle-first discipline is applied when selecting an agent for the service of a class-1 customer. Combining this fair discipline, coordination and the assumption that all agents are identical implies that each agent makes an identical decision in each identical situation. Thus, in the long-run, each agent spends the same proportion of time on idling, serving class-1 or serving class-2 customers. Finally, the agents are subject to a limited liability constraint which translates to the idea that their realized revenue should be higher than or equal to their effort-cost, for any realization of the revenue. As random variables, the realized rate of served class-1 and class-2 customers can be zero. Consequently, the minimal revenue for an agent is F. So, the limited liability constraint translates into $F \geq EC_\pi$.

The system manager is risk-neutral and wants to minimize the long-run expected staffing cost, SC, of inducing the agents to choose a reservation policy. Our analysis can then be interpreted as a particular case of the principal–agent problem (Laffont and Martimort, 2009). The principal's

optimization problem of inducing Policy π can be expressed as

$$\begin{aligned} &\underset{r_1,r_2,F}{\text{Minimize}}\ s(r_1T_1 + r_2T_2 + F), \\ &\text{subject to } \begin{cases} F \geq EC_\pi, \text{ and,} \\ \text{Policy } \pi \in \arg\max\{r_1T_1 + r_2T_2 + F - EC_\pi\}, \end{cases} \end{aligned} \tag{1}$$

We focus on the first-best and second-best solutions of Problem (1). The first-best solution solves the principal's optimization problem subject to all constraints except the incentive constraints, assuming that the agents' actions are contractible. The second-best solution solves the principal's optimization problem subject to all constraints.

There remains to specify the objective reservation policy for the system manager. As in Bhulai and Koole (2003), the objective of the system manager is to induce a reservation policy which maximizes the expected rate of served class-2 customers per agent, while maintaining the expected waiting time of served class-1 customers $E(W)$ below a threshold level $\overline{w}$:

$$\begin{cases} \text{Maximize } T_2, \\ \text{subject to } E(W) \leq \overline{w}. \end{cases} \tag{2}$$

Since the two classes of customers are served by the same team of agents, improving the performance of one class of customers is detrimental to the other class of customers. Therefore, to maximize the rate of served class-2 customers per agent, the expected waiting time of class-1 customers should be as close as possible to the threshold level $\overline{w}$. In this way, $\overline{w}$ can be viewed as an objective waiting time for class-1 customers.

After solving Problem (1) in Section 3, we explore how the base model can be extended to provide solutions of Problem (1) with lower staffing cost. In Section 4, we investigate the possibility to directly pay for the idling time of an agent with $r_3 \geq 0$ monetary unit per unit of idling time. We also consider the possibility to discriminate between delayed and non-delayed class-1 customers in the agents' payout by introducing the rewards r_1^d and r_1^{nd} per served delayed and non-delayed customer, respectively. Finally, to account for service systems like call centers that are not observable by agents, in Section 5 we study the solutions of Problem (1) in a setting where agents cannot observe the state of the system. This, in turn, allows us to discuss the opportunity of revealing the system state to the agents through idling delay announcements.

We conclude our model description with three remarks. First, customers and agents are assumed to be aware of the system's characteristics. This means that customers are able to estimate their expected

waiting time upon arrival. This evaluation is possible if customers know the system in terms of service speed and number of agents. The same stipulation holds for the agents. To decide on an appropriate reservation strategy, the agents should be aware of the arrival rate of class-1 customers. This means that our study is valid in a context involving experienced agents and regular customers, or when the system can provide accurate evaluations of the performance measures through delay announcements or idling time estimates. If the system is unknown by the agents or customers, their decisions may be different from those presented in this paper (Debo and Veeraraghavan, 2014).

Second, we assume that customers' utility is based on their expected waiting time. While this choice is standard in the queueing literature, it may not correspond to human psychology regarding the waiting time aversion. For instance, a percentile of the waiting time or the expected excess may also be considered (Maister et al., 1984). Determining the most appropriate metric to capture customers' perception of the wait is an ongoing research issue with an impact on scheduling and staffing decisions (Koole, 2003). However, we decided not to pursue the analysis of the system through other performance measures as they are unlikely to contribute significantly to our observations.

Third, we assume that each individual agent wants to maximize their individual utility in a context where coordination between the agents is possible. Without coordination, the policy which maximizes the agents' utility is not necessarily the one that each agent would potentially adopt in an individual revenue maximizing context. However, it is likely that this is not the case as demonstrated for service rate optimization in Gopalakrishnan et al. (2016) due to the fair longest-idle-first discipline. Yet, this outcome is very difficult to prove in our case as customers are delay-sensitive. Moreover, defining a reservation policy for one agent independently from the policy of the other agents does not seem feasible. Thus, we also decided not to pursue the analysis of a system without coordination.

We end this section with a table of the notations used throughout the article (Table 1).

3 On the efficiency of agents self-routing

In this section, we determine the efficiency of agents self-routing in the framework described in Section 2. First, in Section 3.1, we determine the customers' equilibrium joining decisions. Next in Section 3.2, we characterize the agents' reservation policy in the first- and second-best settings. This, in turn, allows us to determine the optimal payout parameters in Section 3.3. Finally in Section 3.4, we provide numerical illustrations of our analysis revealing the high cost of reservation when agents' self-routing is implemented.

Table 1: Table of notations

	System state
x	Number of customers in the system (class-1 + class-2)
y	Number of class-2 customers in service
	Parameters of the queueing model
λ	Class-1 customer arrival rate
μ_i	Service rate of class-i customers for $i = 1, 2$
μ	Service rate when $\mu_1 = \mu_2$
s	Number of agents
a	Offered load for class-1 customers when $\mu_1 = \mu_2$: $a = \frac{\lambda}{\mu}$
	Customers' utility and decision parameters
R_S	Reward for being served
C_W	Cost per unit of waiting time
n_y	Joining threshold on the number of class-1 customers in the queue when y class-2 customers are in service ($y = 0, 1, \cdots, s$)
n	Joining threshold on the number of class-1 customers in the queue when $\mu_1 = \mu_2$
	Agents' reservation policy
Policy π_{fb}	Reservation policy in the first-best setting with parameters p and $\tilde{c}$
Policy π_{sb}	Reservation policy in the observable second-best setting with threshold level $\tilde{c}$
Policy π_{sb}^{no}	Reservation policy in the non-observable second-best setting with threshold level q
$\tilde{c}$	Rank of Policy π_{sb} when policies are sorted in ascending order of their rate of served class-1 customers with $\tilde{c} = 0, 1, \cdots, \overline{c}$ (i.e., reservation level of Policy π_{sb})
$\overline{c}$	Maximal value for $\tilde{c}$ (i.e., when all agents are reserved for class-1 customers)
c	Reservation threshold for the agents when $\mu_1 = \mu_2$ with $c = 0, 1, \cdots, s$
p	Probability to select reservation level $\tilde{c} + 1$ at service completion for Policy π_{fb}, while $1 - p$ is the probability to select reservation level $\tilde{c}$ with $0 \leq p \leq 1$
q	Probability to idle at service completion with Policy π_{sb}^{no} with $0 \leq q \leq 1$
	Agents' utility, payments and effort
r_i	Reward per served class-i customer for $i = 1, 2$
r_3	Reward per unit of idling time
r_1^d, r_1^{nd}	Reward per served delayed and non-delayed class-1 customer, respectively
F	Fix wage per agent
EC_π	Effort cost per agent under Policy π ($EC_\pi = eP_\pi$)
e	Cost per unit of non-idling time
P_π	Proportion of busy time for an agent under Policy π
SC	Staffing cost in the base model
VC	Variable part of the staffing cost in the base model
SC^*	Staffing cost in the model which discriminates delayed and non-delayed class-1 customers
U	Agents' utility
	Performance measures
T_i	Expected rate of served class-i customers per agent for $i = 1, 2$
P_3	Proportion of idling time for an agent
T_1^d, T_1^{nd}	Expected rate of served delayed and non-delayed class-1 customers per agent, respectively
$E(W)$	Expected waiting time of served class-1 customers
$\overline{w}$	Objective expected waiting time
$W_y(z)$	Expected waiting time upon arrival when z customers are in the queue and y class-2 customers are in service

3.1 Customers' equilibrium joining decisions

Bhulai and Koole (2003) proved that one property of the optimal reservation policy which solves Problem (2) is the priority for class-1 customers. This property means that at service completion, an available agent always chooses to start the service of a class-1 customer if there is at least one customer waiting in the queue. In the first-best setting, this priority rule is contractible. In the second-best

setting, we show in Section 3.3 that the payout parameters can be selected such that the priority for class-1 customers can also be ensured. Thus assuming priority for class-1 customers, we determine the equilibrium joining strategy of the customers. Theorem 1 proves that customers' equilibrium joining policy is of the threshold type and provides the threshold parameters. The notation $\lceil v \rceil$ defines the first integer above a given real v.

Theorem 1. *Customers' equilibrium joining policy is a deterministic threshold policy with queue-length threshold parameters n_y for $0 \leq y \leq s$, where y represents the number of class-2 customers in service. Under this policy, an arriving customer joins the system if they observe $n_y - 1$ or fewer customers in the queue and balks if they observe n_y customers or more with*

$$n_y = \left\lceil W_y^{-1}\left(\frac{R_S}{C_W}\right)\right\rceil, \tag{3}$$

where $W_y(z) = \sum_{j=1}^{y} A_{j,y}\frac{1-p_j^{z+1}}{1-p_j} + (z+1)t_0$, *with* $p_y = \frac{(s-y)\mu_1}{(s-y)\mu_1+y\mu_2}$ *and* $t_y = \frac{1}{(s-y)\mu_1+y\mu_2}$ *for* $0 \leq y \leq s$, $A_{1,1} = t_1 - t_0$, $A_{j,y} = \frac{1-p_y}{p_j-p_y}A_{j,y-1}$ *for* $j < y$, *and* $A_{y,y} = t_y - \left(t_0 + \sum_{j=1}^{y-1}\frac{1-p_y}{p_j-p_y}A_{j,y-1}\right)$.

Next, Corollary 1 specifies the monotonicity properties of the joining thresholds. As expected, if the service time of class-2 customers is longer than the one of class-1 customers, then a situation with a high number of class-2 customers in service (i.e., a high value for y) may lead to long waiting times for which customers react by selecting a low joining threshold. In the case of equal service rates, the optimal joining policy results in a threshold policy with parameter n, which does not depend on the number of class-2 customers in service. This threshold is also the maximum possible number of customers in the queue.

Corollary 1. *The following holds:*

- *The threshold n_y is strictly increasing (respectively, strictly decreasing) in y if $\mu_1 > \mu_2$ (respectively, if $\mu_1 < \mu_2$).*
- *When $\mu_1 = \mu_2 = \mu$, customers' equilibrium joining policy is a threshold policy with parameter $n = \left\lceil \frac{s\mu R_S}{C_W} \right\rceil$. Under this policy, an arriving customer joins the system if they observe $n-1$ or fewer customers in the queue and balks if they observe n customers or more.*

3.2 Agents' reservation policy

We now determine the optimal reservation policy for the agents in the first- and second-best settings.

First-best setting. In the first-best setting, the agents' actions are contractible. Therefore, the objective reservation policy is the one that solves Problem (2) under external control. In the case $\mu_1 = \mu_2$, Bhulai and Koole (2003) proved that this policy, termed Policy π_{fb}, (i) gives priority to class-1 customers, (ii) provides a threshold type reservation for class-1 customers, and (iii) randomizes in between adjacent deterministic threshold policies. Properties (i), (ii) and (iii) determine the possible actions of an agent who just completed service. Property (i) states that if there is at least one class-1 customer waiting in the queue, then the first customer in line in the queue directly starts service. Property (ii) determines the decision actions when the queue is empty through a reservation threshold c for $c = 0, 1, \cdots, s$. Specifically, if the number of idle agents (excluding the one who just completed service) is at least c, then the newly available agent initiates the service of a class-2 customer, otherwise she remains idle. This means that there are c agents that are reserved for class-1 customers. Hence, there are at least $s-c$ agents working at any time. Property (iii) is defined by a randomizing parameter p for $0 \leq p \leq 1$ such that at service completion either the deterministic reservation threshold c is selected with probability $1-p$ or the reservation threshold $c+1$ is selected with probability p.

In the case $\mu_1 \neq \mu_2$, there is no proof of the form of the optimal policy in the queueing literature. Yet, numerical investigations from Markov decision process analyses indicate that Policy π_{fb} is also optimal when $\mu_1 \neq \mu_2$ with the particularity that the reservation threshold should depend on the number of class-1 and class-2 customers in service (Bhulai and Koole, 2003). Therefore, we assume that Policy π_{fb} with the reservation threshold possibly being state-dependent, is the objective reservation policy in the first-best setting. Due to the competition between class-1 and class-2 customers for the same resources (the agents), if one policy, termed Policy π, provides a higher rate of served class-1 customers than another one, termed Policy π', then the rate of served class-2 customers is lower with Policy π than with Policy π'. Therefore, we can sort the admissible deterministic reservation policies by the resulting rate of served class-1 and class-2 customers. We write that Policy π has more reservation than Policy π' if the rate of served class-1 customers (respectively, the rate of served class-2 customers) with Policy π is higher (respectively, lower) than the one with Policy π'. We call by level of reservation, the rank of each admissible policy when admissible policies are sorted in ascending order of their rate of served class-1 customers. The level of reservation is denoted by $\tilde{c}$, where $\tilde{c}$ is an integer such that $0 \leq \tilde{c} \leq \overline{c}$. With $\tilde{c} = 0$, agents never idle, and with $\tilde{c} = \overline{c}$, all agents are reserved for class-1 customers. Note that $\overline{c}+1$ is the number of admissible deterministic policies. In the case $\mu_1 = \mu_2$, we have $\tilde{c} = c$ and $\overline{c} = s$.

In the case $\mu_1 \neq \mu_2$, Policy π_{fb} is then defined by the parameters $\tilde{c}$ and p, where $\tilde{c}$ is the deterministic reservation level and p is the randomizing parameter. As in the case $\mu_1 = \mu_2$, at each service completion, either a policy with reservation level $\tilde{c}+1$ is implemented with probability p or a policy with reservation level $\tilde{c}$ is implemented with probability $1-p$. With this definition, the sum $\tilde{c}+p$ can be viewed as the level of reservation of Policy π_{fb}.

Second-best setting. We now determine how agents select their reservation policy when they make decisions in a utility maximizer perspective. To derive the agents' reservation policy, we develop an iterative Markov decision process approach. This approach applies for our model as the queueing model under consideration can be represented by a Markov chain and the maximal event rate, $\lambda + s\max(\mu_1, \mu_2)$, is bounded. Therefore, uniformization is possible. Consequently, we assume, without loss of generality, that $\lambda + s\max(\mu_1, \mu_2) = 1$, such that the transition rates in the continuous time Markov process can be viewed as transition probabilities in the equivalent discrete time one (Koole, 2007). The state of the system is defined by the couple (x, y), where x is the number of customers present (class-1+class-2) and y is the number of class-2 customers in service. We assume that the parameters r_1, r_2, and F are chosen such that class-1 customers have service priority. We can then define the value function, $V_k(x, y)$, over k steps in order to capture the agents' utility. We choose $V_0(x, y) = 0$ and for $k \geq 0$,

$$\begin{aligned} V_{k+1}(x,y) = sF + \lambda \left(\mathbb{1}_{x-s<n_y} \left(I_k(x+1,y) + r_1 - \frac{e}{\mu_1} \right) + \mathbb{1}_{x-s=n_y} I_k(x,y) \right) \\ + \min(s-y, x-y)\mu_1 I_k(x-1,y) + y\mu_2 I_k(x-1,y-1) \\ + (1 - \lambda - \min(s-y, x-y)\mu_1 - y\mu_2) I_k(x,y), \end{aligned} \tag{4}$$

with $x \geq 0$, and $0 \leq y \leq \min(x, s)$, where $I_k(x, y) = V_k(x, y)$, if $x \geq s$ and $I_k(x, y) = \max\left(V_k(x, y), V_k(x+1, y+1) + r_2 - \frac{e}{\mu_2}\right)$, if $x < s$. The operator I_k controls the decision to initiate a class-2 service. We obtain the long-run average optimal actions by applying the value iteration technique introduced by Bellman (1957) and Howard (1960), by recursive evaluation of V_k using (4) for $k \geq 0$. As k tends to infinity, the optimal policy converges to the unique average optimal policy, which maximizes the agents' utility (i.e., Policy π_{sb}) and the difference $V_{k+1}(x, y) - V_k(x, y)$ converges to the long-run optimal utility for the agents (Puterman, 1994).

From various numerical experiments, we observe that Policy π_{sb} is a deterministic state-dependent threshold one with threshold level $\tilde{c}$ as defined in the first-best setting. This means that self-routing

cannot induce the first-best effort as Policy π_{sb} differs from Policy π_{fb} by the impossibility to randomize in between adjacent threshold levels. We partially prove the form of the optimal reservation policy in Theorem 2. Specifically, we prove that Policy π_{sb} is deterministic. Moreover, we prove its threshold form when customers are highly wait-averse, translated by the condition $\frac{R_S s \max(\mu_1,\mu_2)}{C_W} < 1$. This condition translates into customers joining the system only if at least one agent is available. It means that Theorem 2 proves the threshold form of the reservation policy in a loss system. This represents a novel contribution in the analysis of blended queues as the case $\mu_1 \neq \mu_2$ has never been tackled.

Theorem 2. *The optimal policy for the agents is a deterministic stationary policy. For* $\frac{R_S s \max(\mu_1,\mu_2)}{C_W} < 1$*, the optimal policy for the agents is a state-dependent reservation threshold policy, defined by the function* $y = c(x)$ *such that if the system is in state* (x, y) *with* $x < s$ *and* $y < s - c(x)$*, then* $s - c(x) - y$ *class-2 services should be initiated.*

When $\mu_1 = \mu_2$, Theorem 2 can easily be proven by adjusting the proof of Theorem 3.2 in Bhulai and Koole (2003). The reservation policy then becomes a threshold policy with reservation threshold c for $c = 0, 1, \cdots, s$ as defined in the first-best setting.

3.3 Optimal payout parameters

We now evaluate how the payout parameters should be selected in the first- and second-best settings to solve Problem (1). First, we provide conditions for the priority for class-1 customers in the second-best setting. Using the definition of EC_π, the utility of an agent can be rewritten as $\left(r_1 - \frac{e}{\mu_1}\right) T_1 + \left(r_2 - \frac{e}{\mu_2}\right) T_2 + F$. So, if $r_1 < \frac{e}{\mu_1}$, then the effort for serving a class-1 customer is not compensated by the reward r_1. In this case, an agent would prefer to idle than to serve a class-1 customer. The system manager should instead select r_1 such that $r_1 \geq \frac{e}{\mu_1}$ in order to create a preference for serving class-1 customers instead of idling. Next, having $r_2\mu_2 > r_1\mu_1$ creates an incentive for agents to serve only class-2 customers as these customers are always available for service. The system manager should instead select r_1 such that $r_1\mu_1 \geq r_2\mu_2$ in order to induce a priority for class-1 customers over class-2 customers.

After ordering the admissible policies in the second-best setting according to their reservation level, we observe that the maximum of the agents' utility in $\tilde{c}$ is unique. Having a unique maximum is important as it determines how agents may choose their reservation policy. If there was more than one local maximum, agents could choose a local maximum that is not necessarily the global one. We also observe that the level of reservation increases with r_1. This second observation is intuitive and

represents a necessary step for optimizing the reward r_1 in Theorem 4. These results are proven in Theorem 3 in the case $\mu_1 = \mu_2$.

Theorem 3. *In the case $\mu_1 = \mu_2$, the agents' utility has a unique maximum in the reservation threshold c. Moreover, the optimal reservation threshold c increases with r_1.*

We now provide the solutions of Problem (1) for the system manager in the first- and in the second-best settings in Theorem 4. This theorem is proven assuming that Theorem 3 is also valid when $\mu_1 \neq \mu_2$.

Theorem 4. *The optimal compensation parameters in the first-best and in the second-best settings to achieve Policy π_{fb} and Policy π_{sb} with reservation level $\tilde{c}$, respectively, are as follows:*

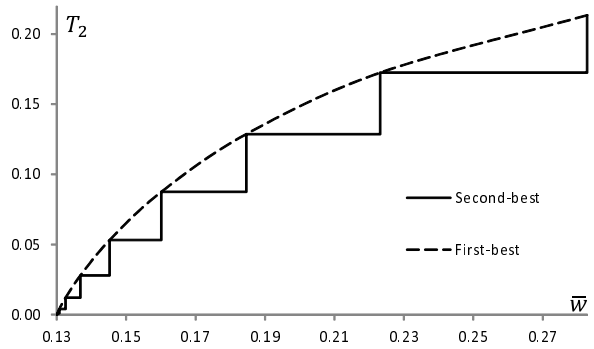
- *First-best setting: $r_1 = r_2 = 0$, and $F = EC_{\pi_{fb}}$.*
- *Second-best setting: $F = EC_{\pi_{sb}}$, and*
  - *If agents should not be reserved for class-1 customers (i.e., if $\tilde{c} = 0$), then $r_1 = \frac{e}{\mu_1}$ and $r_2 = \frac{e}{\mu_2}$.*
  - *For non-extreme reservation policies (i.e., if $0 < \tilde{c} < \overline{c}$), $r_1 = \frac{e}{\mu_2}\frac{T_2(\tilde{c}-1)-T_2(\tilde{c})}{T_1(\tilde{c})-T_1(\tilde{c}-1)}$ and $r_2 = \frac{e}{\mu_2}$, where $T_i(\tilde{c})$ and $T_i(\tilde{c}-1)$ are the rates of served class-i customers per agent with reservation level $\tilde{c}$ and $\tilde{c}-1$, respectively, for $i = 1, 2$.*
  - *If all agents should be reserved for class-1 customers (i.e., if $\tilde{c} = \overline{c}$), then $r_1 = \frac{e}{\mu_1}$ and $r_2 = 0$.*

From Theorem 4, we deduce the expression of the optimal staffing cost, $SC_{\tilde{c}}$:

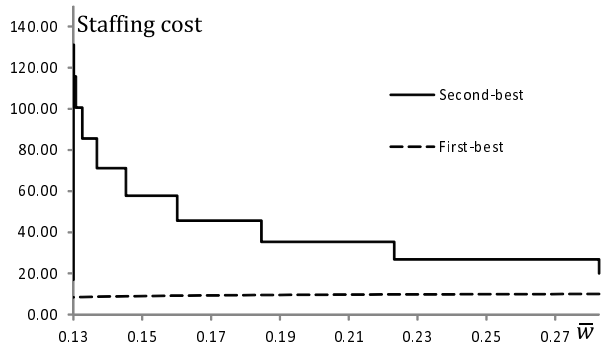
$$
\begin{aligned}
SC_0 &= \frac{2es}{\mu_1}T_1(0) + \frac{2es}{\mu_2}T_2(0) \text{ for } \tilde{c} = 0, \\
SC_{\tilde{c}} &= \left(\frac{es}{\mu_2}\frac{T_2(\tilde{c}-1) - T_2(\tilde{c})}{T_1(\tilde{c}) - T_1(\tilde{c}-1)} + \frac{es}{\mu_1}\right)T_1(\tilde{c}) + \frac{2es}{\mu_2}T_2(\tilde{c}) \text{ for } 0 < \tilde{c} < \overline{c}, \text{ and} \\
SC_{\overline{c}} &= \frac{2es}{\mu_1}T_1(\overline{c}) \text{ for } \tilde{c} = \overline{c}.
\end{aligned}
\tag{5}
$$

3.4 Numerical illustration

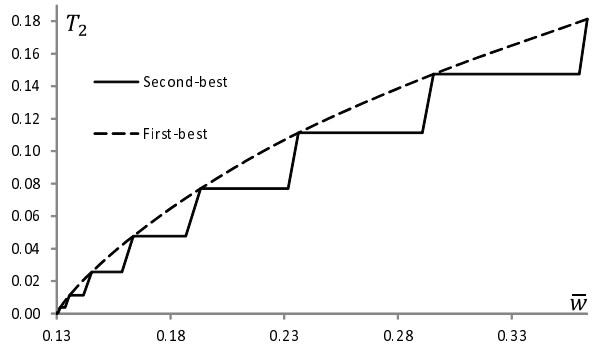
We now investigate the impact of achieving a service level objective $\overline{w}$ for class-1 customers. In Proposition 1, we prove that the variable part of the staffing cost is decreasing and convex in $\overline{w}$ for $E(W)_{s-1} \leq \overline{w} \leq E(W)_0$ when $\mu_1 = \mu_2$, where $E(W)_c$ is the expected waiting time associated with



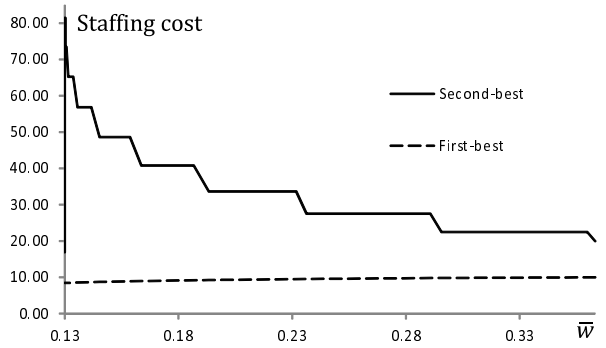
(a) Rate of served class-2 customers per agent with $\mu_2 = 1$



(b) Staffing cost with $\mu_2 = 1$



(c) Rate of served class-2 customers per agent with $\mu_2 = 0.8$



(d) Staffing cost with $\mu_2 = 0.8$

Figure 1: Numerical illustrations ($\lambda = 9$, $s = 10$, $\mu_1 = 1$, $e = 1$, $R_S = 5$, $C_W = 3$)

the reservation threshold c for $c = 0, 1, \cdots, s$. Numerically, we observe that the same result holds for the overall staffing cost when $\mu_1 \neq \mu_2$. However, it is difficult to prove this result as the effort cost is neither convex nor concave in the reservation threshold c. Note also that the decreasing property does not hold between reservation thresholds s and $s-1$ as the principal does not need to provide an incentive for serving class-2 customers when all agents should be reserved for class-1 customers.

Proposition 1. *The variable part of the staffing cost is decreasing and convex in $\overline{w}$ in the second-best setting when $\mu_1 = \mu_2$ for $E(W)_{s-1} \leq \overline{w} \leq E(W)_0$.*

In Figure 1 we provide a numerical illustration for the solutions of Problems (1) and (2) in the first- and second-best settings as functions of the service level objective for the expected waiting time of served class-1 customers $\overline{w}$. Figures 1(a) and 1(c) present the optimal rate of served class-2 customers per agent (i.e., solution of Problem (2)) and Figures 1(b) and 1(d) provide the corresponding staffing cost (i.e., solution of Problem (1)).

Policy π_{sb} does not allow for randomization. This explains why the staffing cost and the rate of served class-2 customers are step functions. Each step corresponds to a change in the reservation

threshold. On the contrary, Policy π_{fb} allows for randomizing in between adjacent thresholds in order to saturate the constraint on the expected waiting time (i.e., $E(W) = \overline{w}$ under Policy π_{fb}). This means that Policy π_{fb} behaves as if the reservation threshold was a real whereas Policy π_{sb} is restricted to a finite set of integer reservation thresholds. Consequently, in situations where we cannot obtain $E(W) = \overline{w}$ with an integer reservation threshold, Policy π_{fb} provides a higher rate of served class-2 customers than Policy π_{sb}. This is an illustration that self-routing cannot induce the first-best effort. The difference between the two policies can be significant in small systems, where the number of achievable deterministic reservation policies is low. Having different service rates, as in Figures 1(c) and 1(d), allows for a wider range of admissible reservation policies. However, the benefits of having more admissible policies are limited. We observe that the admissible policies can be divided into $s+1$ subsets where each subset contains one reservation policy of the case $\mu_1 = \mu_2$. Within one subset of policies, the solutions of Problems (1) and (2) do not differ much. This explains why the staffing cost and the rate of served class-2 customers are close to be step functions also when $\mu_1 \neq \mu_2$. This means that the fixed reservation policy –which is optimal in the case $\mu_1 = \mu_2$– is close to optimal in the case $\mu_1 \neq \mu_2$.

Second, we observe that the staffing cost reaches extremely high values in the second-best setting as compared to the first-best one when the objective waiting time for class-1 customers is low. This observation is supported by the convexity property of the variable part of the staffing cost proven in Proposition 1. For the system manager, this means that above a certain reservation threshold, further increasing reservation as a means to reduce the expected waiting time is very costly for only a limited improvement. Finally, as shown in Figure 1(d), having a slow speed of service for class-2 customers reduces the attractiveness of these customers for the agents that then reduces the need to pay a high price for the service of class-1 customers. The impact of the other system parameters are intuitive and are confirmed by other numerical investigations not presented here. For instance, it is known that the expected waiting time increases with the arrival rate of class-1 customers and reduces with the system size. Therefore, low traffic or large systems only require a low reservation level to achieve a sufficiently low expected waiting time, resulting in a low staffing cost.

Our analysis justifies why most call centers and service systems modeled by blended queues do not implement agents' self-routing. Although more satisfying at the human resource level, linear incentives result in too high an increase in the agents' pay to achieve a certain service level as compared to automated routing. In the next section, we explore extensions of the base model that could possibly

reduce the agents' pay.

4 Improvement of the incentive model

In the base model, agents face the dilemma to either gain a direct reward r_2 for serving a class-2 customer or a postponed one r_1 for a more profitable service. If the principal wants an agent to take the latter alternative, then the parameter r_1 should be set sufficiently high to compensate the zero profit idling period before receiving a class-1 customer. The analysis in the previous section shows that r_1 reaches a very high value when the objective waiting time is low. This reveals the question of redefining the payment structure as a means to reduce the staffing cost. We propose two extensions of the base model. In the first one, we investigate the possibility of directly paying for idling. In the second one, we propose to discriminate between the delayed and non-delayed class-1 customers.

Should we pay for idling? We extend the base model by including a reward for idling of $r_3 \geq 0$ monetary unit per unit of idling time. The idea is that having $r_3 > 0$ could create an additional motivation for idling, which would allow the principal to reduce the value of r_1 to achieve Policy π_{sb}. However, we prove in Proposition 2 that this does not reduce the staffing cost. In this proposition, we prove that paying for idling would force the system manager to pay a higher price for serving class-1 and class-2 customers, which consequently increases the staffing cost as compared to the case with $r_3 = 0$. Therefore, paying for idling should be excluded.

Proposition 2. *The staffing cost cannot be reduced in the second-best setting by including a reward for idling. In other words,* $r_3 = 0$ *is the optimal compensation for idling.*

Should we discriminate between delayed and non-delayed class-1 customers? Recall that Policy π_{sb} has two properties: priority for class-1 customers and threshold reservation policy. Reservation is stronger than priority. However, not all class-1 customers benefit from a reserved agent. Customers who benefit from a reserved agent are those who do not wait. Delayed class-1 customers only benefit from the priority over class-2 customers. Therefore, we propose discriminating between delayed and non-delayed customers in the reward structure as non-delayed customers benefit from a stronger property of Policy π_{sb} than delayed ones. We introduce the reward parameters r_1^d and r_1^{nd} to reward the service of a delayed and a non-delayed class-1 customer, respectively. This makes this compensation model non-linear as the service of customers of the same class is rewarded as a function

of the service quality provided. In Proposition 3, we provide the optimal values for r_1^d, r_1^{nd}, and r_2 in the second-best setting. We denote by $T_1^{nd}(\tilde{c})$ and $T_1^d(\tilde{c})$ the rates of served non-delayed and delayed class-1 customers, respectively, under Policy π_{sb} at reservation level $\tilde{c}$.

Proposition 3. *The optimal compensation parameters to achieve Policy π_{sb} in the second-best setting with reservation level $\tilde{c}$ are given by $F = EC_{\pi_{sb}}$ and:*

- *If agents should not be reserved for class-1 customers (i.e., if $\tilde{c} = 0$), then $r_1^{nd} = r_1^d = \frac{e}{\mu_1}$ and $r_2 = \frac{e}{\mu_2}$.*
- *For non-extreme reservation policies (i.e., if $0 < \tilde{c} < \overline{c}$), $r_1^{nd} = \frac{e}{\mu_2}\frac{T_2(\tilde{c}-1)-T_2(\tilde{c})}{T_1^{nd}(\tilde{c})-T_1^{nd}(\tilde{c}-1)} + \frac{e}{\mu_1}\frac{T_1^d(\tilde{c}-1)-T_1^d(\tilde{c})}{T_1^{nd}(\tilde{c})-T_1^{nd}(\tilde{c}-1)}$, $r_1^d = \frac{e}{\mu_1}$ and $r_2 = \frac{e}{\mu_2}$.*
- *If all agents should be reserved for class-1 customers (i.e., if $\tilde{c} = \overline{c}$), then $r_1^{nd} = r_1^d = \frac{e}{\mu_1}$ and $r_2 = 0$.*

From Proposition 3, we express the optimal staffing cost $SC_{\tilde{c}}^*$ for this compensation model as follows:

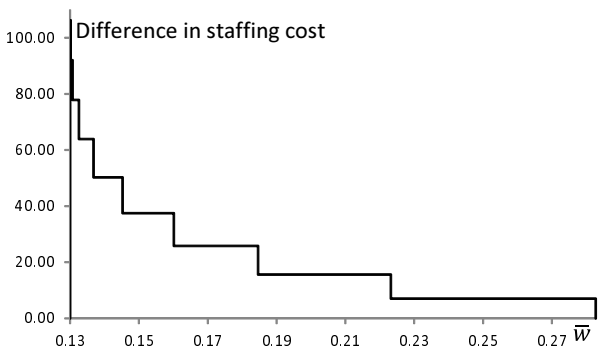
$$
\begin{aligned}
SC_0^* &= \frac{2es}{\mu_1}T_1(0) + \frac{2es}{\mu_2}T_2(0) \text{ for } \tilde{c} = 0, \qquad (6)\\
SC_{\tilde{c}}^* &= \left(\frac{es}{\mu_2}\frac{T_2(\tilde{c}-1)-T_2(\tilde{c})}{T_1^{nd}(\tilde{c})-T_1^{nd}(\tilde{c}-1)} + \frac{es}{\mu_1}\frac{T_1^d(\tilde{c}-1)-T_1^d(\tilde{c})}{T_1^{nd}(\tilde{c})-T_1^{nd}(\tilde{c}-1)} + \frac{es}{\mu_1}\right)T_1^{nd}(\tilde{c}) + \frac{2es}{\mu_1}T_1^d(\tilde{c}) + \frac{2es}{\mu_2}T_2(\tilde{c})\\
&\qquad \text{for } 0 < \tilde{c} < \overline{c}, \text{ and}\\
SC_{\overline{c}}^* &= \frac{2es}{\mu_1}T_1(\overline{c}) \text{ for } \tilde{c} = \overline{c}.
\end{aligned}
$$

In Corollary 2, we prove that by discriminating between delayed and non-delayed class-1 customers in the incentive structure, we reduce the staffing cost except in the extreme cases $\tilde{c} = 0$ and $\tilde{c} = \overline{c}$. The proof of Corollary 2 follows from the expressions of $SC_{\tilde{c}}$ and $SC_{\tilde{c}}^*$.

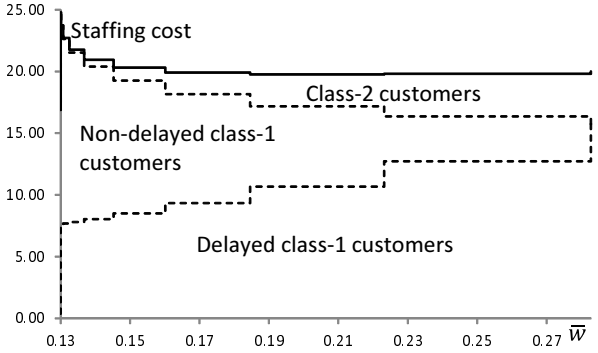
Corollary 2. *Discriminating between delayed and non-delayed class-1 customers in the incentive structure reduces the staffing cost except when $\tilde{c} = 0$ or $\tilde{c} = \overline{c}$. Specifically,*

$$
\begin{aligned}
&SC_{\tilde{c}} - SC_{\tilde{c}}^* = 0 \text{ for } \tilde{c} = 0 \text{ and } \tilde{c} = \overline{c}, \text{ and}\\
&SC_{\tilde{c}} - SC_{\tilde{c}}^* = s\frac{T_1^{nd}(\tilde{c})T_1^d(\tilde{c}-1) + T_1^{nd}(\tilde{c}-1)T_1^d(\tilde{c})}{T_1^{nd}(\tilde{c}) - T_1^{nd}(\tilde{c}-1)}\left(\frac{e}{\mu_2}\frac{T_2(\tilde{c}-1)-T_2(\tilde{c})}{T_1(\overline{c})-T_1(\overline{c}-1)} - \frac{e}{\mu_1}\right) > 0 \text{ for } 0 < \tilde{c} < \overline{c}.
\end{aligned}
$$

In Figure 2(a), we present the difference in staffing cost between the base model and the non-linear one studied in this section. This illustrates the result of Corollary 2. We observe that the difference



(a) Difference in staffing cost between the two compensation models ($SC_{\bar{c}} - SC^*_{\bar{c}}$)



(b) Decomposition of the staffing cost

Figure 2: Numerical illustration ($\lambda = 9$, $s = 10$, $\mu_1 = \mu_2 = 1$, $e = 1$, $R_S = 5$, $C_W = 3$)

between the two contracts is high, especially when $\overline{w}$ is low. This suggests that delay-dependent rewards should be implemented instead of a non-discriminating pay-per-task contract. In Figure 2(b), we delineate the staffing cost of the three components related to the compensation of delayed class-1, non-delayed class-1, and class-2 customers. As expected, the parts related to class-2 and delayed class-1 customers increase with the objective expected waiting time for class-1 customers while the part related to non-delayed customers reduces. Moreover, we observe that the overall staffing cost is close to be a constant function of $\overline{w}$. This is another difference from the base model (Figure 1(b)). When $\overline{w}$ increases, the overall rate of served customers increases as agents spend less time idling. This may increase the staffing cost. Yet, the rate of the most rewarded services reduces, which tends to reduce the staffing cost. These two competing phenomena result in a close to constant staffing cost in the model with non-linear rewards. With the base model, the staffing cost was mainly driven by the high payment due to served class-1 customers. This further argues in favor of the non-linear model. With a close to constant staffing cost, the decision to select a given value for $\overline{w}$ becomes mainly driven by the service quality that the system manager wants to provide to class-1 and to class-2 customers and less by the cost it could induce.

5 Analysis when agents cannot observe the system state

In some systems, like call centers, agents do not know the state of the system. Therefore, their decisions cannot be made on the number of available agents as was the case in the analysis in Section 3. To account for such systems, we analyze Problem (1) with the base model in a context where agents are blind to the system state. Nevertheless, call centers can easily be made observable to the agents by

announcing their idling delay before receiving a class-1 customer. Thus, at the end of the section, we compare the solutions of Problem (1) in the observable and non-observable cases for the agents in order to discuss the opportunity to announce idling delays.

Remark: As for agents, call centers' state is non-observable to customers but can be made observable through delay announcements. The question of announcing delays to customers has received a lot of attention in the field of call centers (Guo and Zipkin, 2007; Allon and Bassamboo, 2011; Yu et al., 2017, 2018). We also investigated this question in our context. The results of this analysis were expected from the literature on delay announcement. In particular, in terms of staffing cost, we observe that announcing delays to customers is beneficial in congested situations whereas in a light traffic context, it is better to keep the system non-observable. Thus, we decided to not present the results of this analysis.

The optimal policy in the first-best setting is unchanged as the system manager knows the system state. Thus, we only focus on the second-best policy, termed Policy π_{sb}^{no}. As in the observable case, the priority for class-1 customers is ensured by the inequalities $r_1 \geq \frac{e}{\mu_1}$ and $r_1\mu_1 \geq r_2\mu_2$. Customers' joining decisions are also unchanged due to the priority for class-1 customers. When agents do not know the system state, the only possible reservation policy is a randomizing policy with parameter q. At service completion, an agent decides to idle with probability q or to initiate a class-2 service with probability $1-q$ for $0 \leq q \leq 1$. Thus, the parameter q translates to a level of reservation: when $q = 100\%$ all agents are reserved for class-1 customers, whereas when $q = 0\%$ agents are never idle. In Proposition 4, we extend the result of Theorem 3 in the non-observable case by considering the closed-form expressions of the performance measures in the reservation parameter q. In particular, we prove a novel property of the Erlang loss-function defined as $\gamma(q) = \sum_{x=0}^{s-1} \frac{s!\left(\frac{q}{a}\right)^{s-x}}{x!}$. That is

$$q\gamma(q)\frac{\partial^2\gamma(q)}{\partial q^2} - 2q\left(\frac{\partial\gamma(q)}{\partial q}\right)^2 + 2\gamma(q)\frac{\partial\gamma(q)}{\partial q} \leq 0. \tag{7}$$

Proposition 4. *In the case $\mu_1 = \mu_2$, the agents' utility has a unique maximum in the reservation level, q. Moreover, the optimal reservation level q increases with r_1.*

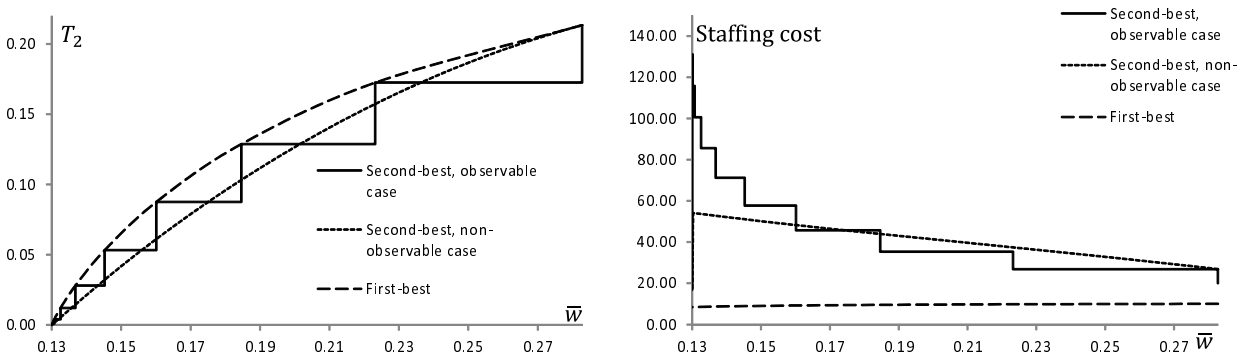
Assuming that Proposition 4 is also valid in the case $\mu_1 \neq \mu_2$, we deduce the optimal incentive parameters r_1 and r_2 in the second-best setting in Proposition 5 from which the staffing cost can be deduced. The proof follows similar arguments as Theorem 4.

Proposition 5. *The optimal compensation parameters in the non-observable second-best setting to*

achieve Policy π_{sb}^{no} *with a reservation level* q *are* $F = EC_{\pi_{sb}^{no}}$*, and*

- *If agents should not be reserved for class-1 customers (i.e., if* $q = 0$*), then* $r_1 = \frac{e}{\mu_1}$ *and* $r_2 = \frac{e}{\mu_2}$.
- *For non-extreme reservation policies (i.e., if* $0 < q < 1$*), then* $r_1 = -\frac{e}{\mu_2} \frac{\frac{\partial T_2(q)}{\partial q}}{\frac{\partial T_1(q)}{\partial q}}$ *and* $r_2 = \frac{e}{\mu_2}$*, where* $T_i(q)$ *is the rate of served class-i customers at reservation level* q *for* $i = 1, 2$.
- *If all agents should be reserved for class-1 customers (i.e., if* $q = 1$*), then* $r_1 = \frac{e}{\mu_1}$ *and* $r_2 = 0$.

Comparison with the observable case. In Figure 3, we compare the solutions in the observable and non-observable cases as functions of the expected waiting time objective, in terms of rate of served class-2 customers per agent (Figure 3(a)) and staffing cost (Figure 3(b)). From Figure 3(a), we



(a) Rate of served class-2 customers per agent

(b) Staffing cost

Figure 3: Comparison between the observable and the non-observable case ($\lambda = 9$, $s = 10$, $\mu_1 = \mu_2 = 1$, $e = 1$, $R_S = 5$, $C_W = 3$)

observe that the rate of served class-2 customers is higher in the observable case for most values of the objective waiting time. This is one positive outcome of making state-dependent decisions over blind ones. However, in the observable second-best setting, agents do not randomize in between adjacent thresholds. Therefore, to satisfy the constraint on the expected waiting time, agents tend to employ a higher reservation threshold than needed. This reduces the rate of served class-2 customers and creates situations where the non-observable model achieves a higher value for T_2 than the observable one. These cases occur more often when the objective expected waiting time is high.

From Figure 3(b), we remark that when the objective waiting time is low, the staffing cost in the non-observable case is significantly lower than the one of the observable case. In the observable case, agents are aware of the expected duration of the idling time before serving a class-1 customer. Specifically, if an agent becomes available and z other agents are already idling, then the expected duration before receiving a class-1 customer is $\frac{z+1}{\lambda}$. The decision variable z is unknown in the non-

observable case. Hence, at service completion, an agent makes an idling decision based on the mean idling duration. When the objective expected waiting time is low, idling times are long and the probability to significantly exceed the mean idling duration is high. As agents only make decisions based on mean values (they are risk-neutral), having a non-observable system allows the system manager to not pay for the longest idling times. This renders the staffing cost lower in the non-observable case as compared to the observable one when $\overline{w}$ is low. The opposite effect occurs when the objective expected waiting time is high. Uninformed agents tend to overestimate the idling time duration, leading the system manager to overpay for it.

Deciding on announcing delays to agents should be a function of the waiting time objective for class-1 customers. When the objective waiting time for class-1 customers is low, announcing delays results in an increased rate of served class-2 customers (i.e., a better solution of Problem (2)). Yet, the staffing cost may reach an extremely high level. This disfavors announcing delays. With a higher objective waiting time for class-1 customers, announcing delays tends to reduce the staffing cost but sometimes leads to a lower rate of served class-2 customers.

6 Conclusion

This paper investigates a multi-agent blended queue with inbound and outbound customers where inbound customers are delay-sensitive. Available agents know the system state and are in control of deciding whether to serve an inbound customer, an outbound one, or to remain idle. The system manager wants to provide a good trade-off between the expected waiting time of inbound customers and the rate of served outbound ones at minimal cost. To this end, they decide the agents' payouts. Agents are rewarded by a fix wage and piece rate compensation dependent on the nature of the served customer (inbound or outbound). The agents' possibility to self-route in this context is a principal–agent problem with moral hazard where the agents' effort consists of selecting their reservation policy. Methodologically, our analysis relies on the monotonicity properties of the performance measures and on Markov decision processes. We show that the optimal routing policy for the agents is a deterministic threshold policy in the second-best setting. From this result, we express the compensation parameters that minimize the staffing cost in the first- and second-best settings.

Agents self-routing with linear payouts results in high staffing costs, especially when the objective waiting time is low. Therefore, our study justifies why most call centers prefer the keep the system manager in charge of routing issues. Nevertheless, the negative consequences of automated routing on

motivation or absenteeism may lead to more companies considering the switch to self-routing. We then investigated extensions of the initial payment model. We found that directly paying for idling should be excluded as it cannot reduce the staffing cost. On the contrary, discriminating between delayed and non-delayed customers in the reward structure has a high potential to reduce the agents' payment. We finally investigated the case where agents cannot observe the system state as in call centers. Our analysis shows that not revealing the system state to the agents through delay announcements can significantly reduce staffing costs when the objective waiting time for inbound customers is low.

The study limitations open up several avenues for future research. It is important to determine how our conclusions would be modified considering other optimization problems or other disciplines of service. The model could be made more complex by including abandonment, non-exponential distribution, and loss-aversion. While we implicitly assumed a monopolistic system, it would be interesting to investigate the impact of competition on the agents' decisions. Finally, other ways of providing revenue to the agents could be investigated. In particular, it would be useful to determine the effect of heterogeneous rewards on the agents' behavior.

References

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

Allon, G. and Bassamboo, A. (2011). The impact of delaying the delay announcements. *Operations Research*, 59(5):1198–1210.

Anisimov, N., Fedorov, S., and Ristock, H. (2017). System and methods for scheduling and optimizing inbound call flow to a call center. US Patent 9,553,988.

Avi-Itzhak, B., Golany, B., and Rothblum, U. (2006). Strategic equilibrium versus global optimum for a pair of competing servers. *Journal of Applied Probability*, 43(4):1165–1172.

Bailyn, L., Collins, R., and Song, Y. (2007). Self-scheduling for hospital nurses: An attempt and its difficulties. *Journal of Nursing Management*, 15(1):72–77.

Baiman, S., Netessine, S., and Saouma, R. (2010). Informativeness, incentive compensation, and the choice of inventory buffer. *The Accounting Review*, 85(6):1839–1860.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton.

Bhulai, S. and Koole, G. (2003). A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 48(8):1434–1438.

Bimpikis, K., Candogan, O., and Saban, D. (2019). Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769.

Braverman, A., Dai, J., Liu, X., and Ying, L. (2019). Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452.

Brunelli, L. (2020). *Work at Home Call Center Salaries*. `https://www.thebalancecareers.com/how-home-call-centers-pay-3542389` (accessed June 20, 2020).

Cachon, G., Daniels, K., and Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384.

Chen, F., Lai, G., and Xiao, W. (2016). Provision of incentives for information acquisition: Forecast-based contracts vs. menus of linear contracts. *Management Science*, 62(7):1899–1914.

Christ, D. and Avi-Itzhak, B. (2002). Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management science*, 48(6):813–820.

Cui, S. and Veeraraghavan, S. (2016). Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Science*, 62(12):3656–3672.

Dai, T., Ke, R., and Ryan, C. (2021). Incentive design for operations-marketing multitasking. *Management Science*.

Debo, L. and Veeraraghavan, S. (2014). Equilibrium in queues under unknown service times and service value. *Operations Research*, 62(1):38–57.

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., and Avramidis, A. (2007). Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645.

Dimitrakopoulos, Y. and Burnetas, A. (2016). Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control. *European Journal of Operational Research*, 252(2):477–486.

Dumas, G., Perkins, M., and White, C. (1996). Call sharing for inbound and outbound call center agents. US Patent 5,519,773.

Echchakoui, S. (2016). Addressing differences between inbound and outbound agents for effective call center management. *Global Business and Organizational Excellence*, 36(1):70–86.

Gans, N. and Zhou, Y. (2003). A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271.

Gavirneni, S. and Kulkarni, V. (2016). Self-selecting priority queues with Burr distributed waiting costs. *Production and Operations Management*, 25(6):979–992.

Geng, X., Huh, W., and Nagarajan, M. (2015). Fairness among servers when capacity decisions are endogenous. *Production and Operations Management*, 24(6):961–974.

Gopalakrishnan, R., Doroudi, S., Ward, A., and Wierman, A. (2016). Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050.

Guo, P. and Zipkin, P. (2007). Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970.

Gurvich, I., Lariviere, M., and Moreno, A. (2019). Operations in the on-demand economy: Staffing services with self-scheduling capacity. In *Sharing Economy*, pages 249–278. Springer.

Hassin, R. and Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media.

Hassin, R. and Roet-Green, R. (2017). The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research*, 65(3):804–820.

Herweg, F., Müller, D., and Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5):2451–77.

Hillmer, S., Hillmer, B., and McRoberts, G. (2004). The real costs of turnover: Lessons from a call center. *Human Resource Planning*, 27(3):34–42.

Howard, R. (1960). *Dynamic Programming and Markov Processes*. Massachusetts Institute of Technology Press, Cambridge.

Hu, M. and Zhou, Y. (2019). Price, wage and fixed commission in on-demand matching. *Available at SSRN 2949513*.

Hung, R. (2002). A note on nurse-self-scheduling. *Nursing Economics*, 20(1):37.

Ibrahim, R. (2018). Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management*, 27(2):234–250.

Jain, S. (2012). Self-control and incentives: An analysis of multiperiod quota plans. *Marketing Science*, 31(5):855–869.

Kalai, E., Kamien, M., and Rubinovitch, M. (1992). Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163.

Koning, C. (2014). Does self-scheduling increase nurses' job satisfaction? An integrative literature review: Flexible work patterns can be beneficial for staff and employers. *Nursing Management*, 21(6):24–28.

Koole, G. (2003). Redefining the service level in call centers. *Technical Report, Department of Stochastics, Vrije Universiteit, Amsterdam*.

Koole, G. (2007). *Monotonicity in Markov reward and decision chains: Theory and applications*. Now Publishers Inc.

Laffont, J. and Martimort, D. (2009). *The theory of incentives: The principal-agent model*. Princeton University press.

Lal, R. and Srinivasan, V. (1993). Compensation plans for single-and multi-product salesforces: An application of the Holmstrom-Milgrom model. *Management Science*, 39(7):777–793.

Legros, B. (2017). Reservation, a tool to reduce the balking effect and the probability of delay. *Operations Research Letters*, 45(6):592–597.

Legros, B., Jouini, O., and Koole, G. (2015). Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430.

Legros, B., Jouini, O., and Koole, G. (2021). Should we wait before outsourcing? Analysis of a revenue-generating blended contact center. *Manufacturing & Service Operations Management*.

Li, S., Chen, K., and Rong, Y. (2020). The behavioral promise and pitfalls in compensating store managers. *Management Science*, 66(10):4899–4919.

Long, X. and Nasiry, J. (2020). Wage transparency and social comparison in sales force compensation. *Management Science*, 66(11):5290–5315.

Lu, L., Van Mieghem, J., and Savaskan, C. (2009). Incentives for quality through endogenous routing. *Manufacturing & Service Operations Management*, 11(2):254–273.

Maister, D. H. et al. (1984). *The psychology of waiting lines*. Harvard Business School Boston, MA.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, 37(1):15–24.

Özkan, E. and Ward, A. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70.

Pang, G. and Perry, O. (2014). A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91.

Phung-Duc, T., Rogiest, W., Takahashi, Y., and Bruneel, H. (2016). Retrial queues with balanced call blending: Analysis of single-server and multiserver model. *Annals of Operations Research*, 239(2):429–449.

Puterman, M. (1994). *Markov Decision Processes*. John Wiley and Sons.

Rönnberg, E. and Larsson, T. (2010). Automating the self-scheduling process of nurses in Swedish healthcare: A pilot study. *Health Care Management Science*, 13(1):35–53.

Stouras, K., Girotra, K., and Netessine, S. (2014). Liveops Inc.: The contact centre reinvented.

Suen, S., Negoescu, D., and Goh, J. (2018). Design of incentive programs for optimal medication adherence. *Available at SSRN 3308510*.

Taylor, T. (2018). On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720.

Villena, J., Tellez, A., Mathur, M., Stout, J., and Ben-Chanoch, E. (2004). Blended agent contact center. US Patent 6,775,378.

Yu, Q., Allon, G., and Bassamboo, A. (2017). How do delay announcements shape customer behavior? An empirical study. *Management Science*, 63(1):1–20.

Yu, Q., Allon, G., Bassamboo, A., and Iravani, S. (2018). Managing customer expectations and priorities in service systems. *Management Science*, 64(8):3942–3970.

Zhan, D. and Ward, A. (2018). The M/M/1+ M queue with a utility-maximizing server. *Operations Research Letters*, 46(5):518–522.

Zhan, D. and Ward, A. (2019). Staffing, routing, and payment to trade off speed and quality in large service systems. *Operations Research*, 67(6):1738–1751.