Should we wait before outsourcing? Analysis of a revenue-generating blended contact center

Benjamin Legros, Oualid Jouini, Ger Koole

Should we wait before outsourcing? Analysis of a revenue-generating blended contact center

Benjamin Legros[1] • Oualid Jouini[2] • Ger Koole[3]

[1] *EM Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France*

[2] *Laboratoire Genie Industriel, CentraleSupélec, Université Paris-Saclay, 9 rue Joliot Curie, 91190, Gif-sur-Yvette, France*

[3] *VU University Amsterdam, Department of Mathematics, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*

benjamin.legros@centraliens.net • oualid.jouini@centralesupelec.fr • ger.koole@vu.nl

Abstract

(1) ***Problem definition***: We consider a revenue-generating call center with inbound and outbound calls, where service and sales activities are blended. For maximizing the call center's revenue, the call center manager exercises two levels of control; agent reservation for inbound calls and call outsourcing. Given the influence of waits on purchase probability, we investigate the strategy of outsourcing customers who have waited already, as opposed to outsourcing customers directly at arrival.

(2) ***Academic / Practical relevance***: The main novelty of this article arises from the use of a single framework to investigate combining agent reservation with outsourcing decisions, and a waiting time-based outsourcing strategy. The existing literature only considers these two strategies in isolation and is restricted to quantity-based decisions. From a practical viewpoint, our results aim to provide decision support tools that are directly implementable in a call center's routing software.

(3) ***Methodology***: We apply a Markov decision process approach to optimize the manager's decisions. The particularity of our approach is that we use the experienced waiting time as a decision variable.

(4) ***Results***: We prove that the optimal policy for reservation and outsourcing is of threshold type. Our main conclusion is that outsourcing customers after letting them wait in-house generates higher revenue than outsourcing calls at arrival. However, it is also detrimental to service quality. In addition, we identify contexts where the difference between the two outsourcing strategies is significant.

(5) ***Managerial implications***: Contrary to standard call center practices, which either consist of specialized teams for one type of call, or only exercising one specific level of decision-making (reservation or outsourcing), we demonstrate the potential of partial outsourcing with partial reservation. Our study shows that small congested call centers are those where the benefits of implementing our results are the greatest.

Keywords: Call centers; Markov decision process; outsourcing; agent reservation; service and sales activities.

1 Introduction

Services and sales activities. Call centers are generally a firm's primary channel of interaction with its customers. Historically, call centers were mainly considered as a service delivery channel with inbound calls only. In computer hardware companies, for example, customers would contact the call center to obtain support for their installation. However, from a marketing perspective, a call center also has the potential to become

an ideal sales environment. Agents are not only considered as a passive workforce that responds to demand but are increasingly encouraged to look for new customers or new sales. In banks, for instance, agents might contact their customers to propose a new insurance policy or a new financial product. To illustrate this, Lerzan and Akşin (2010) note that 25% of bank transactions are projected to take place in call centers and that 80% of the bank's growth comes from selling additional products to existing customers. Consequently, inbound call centers introduced revenue generation as a strategic priority.

In this context, an inbound caller requesting a service can also become a potential source of revenue for the call center. Thus, standard performance measures such as waiting time no longer simply represent a poor level of service, but also may impact customer's reaction to sales offers. Given that a long wait is frustrating and reduces the trust given by customers to the company, the chance that a customer in need of assistance accepts an unexpected sales offer may decrease with the wait. This negative relation between the wait and the purchase probability is given the most consideration in this article. However, Ulku et al. (2017) may contradict this assumption by showing that a long wait is an incentive to consume more when customers have consumption as a primary objective. This case is also examined in this paper. In both cases, it is important for call center managers to control the system's congestion, especially if having fewer customers would result in more customers with greater buying potential.

Outsourcing. One strategy investigated to reduce the flow of inbound calls involves *outsourcing* some of them. Outsourcing is implemented as a way to provide a sufficient service quality for most customers and to reduce costs. The alternative to outsourcing is to hire more staff. For small call centers, it is however an expensive option as it takes time to train and manage new employees. Larger call centers are usually better structured to monitor them. Moreover, existing models for staffing in the call center literature are known to be more effective for large call centers (Harrison and Zeevi, 2005; Bassamboo et al., 2006; Whitt, 2006). This means that mistakes in staffing levels are often encountered in small call centers. Consequently, small call centers often face situations of high congestion. Finally, even with an appropriate staffing level, the server utilization and the risk of having long waits are worse in smaller call centers. For these reasons, small call centers are particularly concerned by outsourcing strategies.

The outsourcing decision is complex when sales activities are blended with service ones. An outsourced customer in need of a service no longer represents a sales opportunity. This means it may not be efficient to outsource too many calls, especially those which have a high purchase probability. Moreover, since the waiting time may influence the purchase probability, it should be taken into account in the outsourcing decision. The traditional solution used in practice and in existing models in the academic literature is to outsource an inbound call upon arrival (Akşin et al., 2008; Ren and Zhou, 2008; Koçağa and Ward, 2010; Schrieck et al., 2014). The decision to outsource a newly arrived call is based on the system's state, i.e., on its expected waiting time. This is referred to as *a priori* outsourcing. However, it is only one way of outsourcing. Another possibility, proposed here, is to accept the new call in the queue, but to allow it to be outsourced later according to its *experienced* waiting time. This is referred to as *a posteriori* outsourcing. Intuitively, both types of outsourcing (*a priori* and *a posteriori*) have advantages and disadvantages. The advantage of *a priori* outsourcing is that

it avoids any useless in-house waiting for outsourced customers, thereby reducing customer dissatisfaction due to excessive waiting. On the other hand, it can lead to a customer being outsourced who in fact could have begun a service within a reasonable time given the variability in service times. With *a posteriori* outsourcing, a decision is taken based on actual waiting time. This provides better control of outsourcing and possibly better sales potential. Despite the potential of *a posteriori* outsourcing policies, to the best of our knowledge, they have not previously been addressed in the call center literature.

A priori and *a posteriori* policies are also implemented in contexts without outsourcing. Because of capacity shortage, call centers commonly employ call rejection, either on arrival (*a priori* rejection) or after a certain waiting period (*a posteriori* rejection). *A posteriori* rejection may or may not be followed by an automatic message. For instance, Amazon employs *a posteriori* rejection without a message (simple disconnection of the call), while Dior plays a voice message after a 2-minute wait. Different types of messages exist, such as an invitation to call back later, to leave a message to be called back, or an invitation to use other resources such as chats or email. Our partner, Interactiv Group, offers CTI (Computer Telephony Integration) software which can reject calls after a certain waiting threshold. The threshold value is adjusted by the customers (the call centers) depending on their business requirements. For instance, it is 6 minutes for the energy company Primagaz, 5 minutes for the pharmaceutical company Sanofi, 3 minutes for the telecom operator Keyyo's sales call center, and 15 minutes for its technical hotline.

Blending. Outsourcing may therefore allow the call center to better serve and sell to customers. However, due to the variability in arrivals, we may encounter situations where agents are idling. It could then also be appropriate to let some agents initiate calls to propose sales offers so as to generate extra revenue for the call center. The operational value of outbound calls is that they can be initiated at a chosen time. This helps prevent idle overcapacity, and limits the need for extremely accurate forecasts. While the benefits of combining inbound and outbound calls in call centers seem clear, the implementation comes with significant operational challenges. Since the amount of work could be considerable, agents may be continually occupied, either answering inbound calls or initiating outbound ones. Unless staffing levels are adjusted, pushing agents to work in such conditions could lead to a degradation in service level in terms of the waiting time experienced by inbound customers. The delay probability would be close to one, for instance. Initiating outbound calls should therefore be limited in order to ensure the adequate service quality of inbound ones. One routing solution proposed in the literature for this type of problem is to develop a *reservation strategy* (Bhulai and Koole, 2003; Gans and Zhou, 2003). The idea is to keep a certain amount of idleness in the agents' team by not allowing agents to initiate outbound calls at all time.

Research question and contributions. We consider a call center with inbound and outbound calls in which the service can generate revenue. Inbound calls initially request a service but can also represent a sales opportunity. The willingness of inbound callers to buy is often related to their waiting experience. To avoid excessive congestion, an outsourcer is contractually engaged to receive a given quantity of calls from the call center per time unit. Therefore, an important challenge for the call center manager is to determine when an agent should initiate an outbound call, and which inbound calls should be outsourced, primarily with a

revenue maximizer perspective and, secondarily, with a consideration for the quality of the service provided. Our main aim is to evaluate the potential of letting customers wait before being outsourced as compared to outsourcing at arrival.

On the methodological level, we employ a Markov decision process approach to prove the threshold form of the optimal routing policy for agent reservation and call outsourcing. The particularity of our approach with *a posteriori* outsourcing is that we model the waiting time of the first customer in line as a decision variable. This helps us to identify new structural properties of the value function which differ from the classical convexity/concavity shown in a value iteration step approach. We also derive the call center's expected revenue and service quality, and prove their monotonicity properties in the control parameters. This allows us to determine the constraints that the reservation and outsourcing thresholds should satisfy. We then explicitly compute the relative value function under the optimal policy and prove that the optimal outsourcing threshold can be computed after a finite number of iterations. This allows us to construct an efficient algorithm to derive the optimal policy.

Next, we compare the two policy classes for outsourcing. The main proven result of the comparison is that *a posteriori outsourcing outperforms a priori outsourcing in a revenue maximizer perspective but not in the quality of service one*. In particular, the added value of *a posteriori* outsourcing is to ensure shorter waits for customers who are served in-house detrimentally to outsourced customers. As expected, the difference between the two outsourcing policies is shown to increase with the system congestion. Our numerical investigations also show that the difference between the two policy classes is mostly significant for call centers with less than 50 agents. For medium to large call centers, the two policies are virtually the same with less than 1% difference in generated revenue and almost no wait. This means that the results of our study are applicable to small call centers as in the Business-to-Business sector (Chevalier and Van den Schrieck, 2008) or larger call centers organized in small independent teams (Jouini et al., 2008).

Different extensions to the initial model are investigated. We consider (i) the abandonment feature, (ii) the possibility of the wait having a positive effect on purchase behavior, and (iii) the effect of having different service rates between inbound and outbound calls. These model extensions do not contradict our main finding but allow us to determine contexts where differences between the two policy classes are most significant. With abandonment, assuming a threshold policy for outsourcing and reservation, we show that the algorithm for computing optimal thresholds remains applicable. Extreme reservation/outsourcing strategies tend to be optimal when customers are highly impatient. This reduces the relative benefits of implementing an *a posteriori* outsourcing policy. When the wait has a positive effect, with or without abandonment, we prove that extreme choices should be made for the outsourcing threshold. For the call center manager, this means deciding either to outsource all inbound callers and become a specialized outbound contact center or serving all inbound calls in-house. Therefore, it does not make sense to have a contract where only a given proportion of inbound calls is outsourced. The effect of having different service rates between inbound and outbound calls is less significant. We show, however, that the difference between the two policy classes is highest when the service rates of inbound and outbound calls are close and when the arrival rate is sufficiently high.

Structure of the article. The rest of the paper is organized as follows. The first section ends with a literature review. Section 2 defines the model and the optimization problem. Section 3 identifies the optimal policies for reservation and outsourcing while Section 4 compares *a priori* outsourcing with *a posteriori* outsourcing. Section 5 investigates different model extensions. Finally, Section 6 concludes the paper and highlights avenues for future research. All proofs are given in the online supplement.

Literature review. We distinguish five streams of literature related to this paper. The first deals with the analysis of cross-selling opportunities in a queueing setting. The second is devoted to understanding outsourcing strategies. The third analyzes multi-channel call center queueing models with reservation policies. The fourth explores the nature and the impact of customer's abandonment. The last considers queueing models where decisions are based on customers' waiting experience.

The combination of sales and services activities is referred to as *cross-selling* and is widely used in retail banking call centers, for example. The empirical study by Aksin and Harker (1999) shows that although cross-selling may significantly improve a firm's revenue, it can have a detrimental effect on customer service due to the additional load it creates on the system. To tackle this congestion problem, different studies have focused on the development of optimal policies. The idea is to determine when cross-selling opportunities can be exercised in a way that will maximize expected profit. To this end, Byers and So (2007) developed a mathematical model which incorporates queueing congestion and customer profiles in order to determine the optimal control policy to maximize revenue, showing the usefulness of real-time information for control decisions in a cross-selling context. Güneş and Akşin (2004) investigated the different value-generation potential of customers and determined a market segmentation scheme which divides customers into two groups (high and low). Looking at various forms of customer segmentation, Gurvich et al. (2009) examined decisions on operational staffing, call routing, and cross-selling to define near optimal policies. Another approach developed by Lerzan and Akşin (2010) involved analyzing cross-selling issues as a dynamic service rate control problem, while Armony and Gurvich (2010) described asymptotically optimal control and staffing schemes implemented as the system load grows larger. Finally, Güneş et al. (2010) developed a model to show the negative impact a failed sales attempt can have on a customer's future behavior. This led to a new policy which took the customer's history into account. The policies developed in the aforementioned studies suggest that cross-selling opportunities should only be exercised below a given number of customers in the system. In our paper, we also develop a threshold type of control system. Our particularity is that a control on the customers' experienced wait can be exercised alternatively to a control on the queue length. Moreover, customers with a low purchase potential can be outsourced.

There is a large body of literature on *outsourcing* strategies in call centers. Some articles focus on helping firms to draw up the contract with an outsourcer (e.g., see Hasija et al. (2008) and Akşin et al. (2008)). In our paper, the definition the contract corresponds to the volume-based contract developed in Akşin et al. (2008). Outsourcing is often seen as a strategy to solve the problem of excessive demand. Therefore, the question is to determine whether it is costlier to employ an extra agent or to outsource a larger quantity of inbound calls. Ren and Zhou (2008) show that although a call center can coordinate staffing levels and outsourcing

decisions, the resulting service quality is frequently below its optimal level. To address this issue, they show the value of contracts where considerable attention is devoted to service quality. Koçağa et al. (2015) develop a joint policy for staffing and call outsourcing that minimizes the long-run average cost by solving a two-stage stochastic program. Schrieck et al. (2014) consider staffing issues in a setting where short-term variability and correlations in time-for-call-arrivals are taken into account. Their study leads to an extension of the square root staffing rule and another staffing method which makes use of the Hayward approximation principles. Other studies consider routing decisions and performance evaluation. For instance, Gans and Zhou (2007) consider a call center with high and low value calls, and evaluate routing schemes for outsourcing some of the low values calls. Gurvich and Perry (2012) consider a service network operated under a threshold-type overflow mechanism. If the waiting room is full, the call is overflowed to an outsourcer. The *a priori* outsourcing considered in our article follows a similar routing scheme as theirs.

A third stream of literature related to this paper analyzes *reservation strategy*. In most studies, reservation strategies are considered when two different job types, namely, inbound and outbound calls, have to be handled by a unique group of agents, involving an analysis of *call blended policies*. Some papers focus on performance evaluation, while others address analysis of blending policies or staffing decisions. Deslauriers et al. (2007) developed various continuous Markov chain models for a call center with inbound and outbound calls. The authors considered a threshold policy and characterized the rate of outbounds and the waiting time distribution of inbounds. Gans and Zhou (2003) and Bhulai and Koole (2003) prove that a threshold policy on the number of idle agents is optimal to maximize the outbound throughput under a service level constraint on inbound waiting time, when inbound and outbound calls have the same service rate. Pang and Perry (2014) consider a large call blending model and propose a logarithmic safety staffing rule, combined with a threshold control policy to ensure that agents' utilization is always close to one with idle agents always present. The common point between most studies on reservation strategy is the use of a reservation threshold policy. In our paper, we prove that such a policy is optimal when outsourcing decisions can be taken together with reservation decisions. Combining these strategies extends the range of options for improving the system's performance. Our study shows that employing a unique reservation or outsourcing strategy is optimal only in extreme workload situations.

In this article, we include the *abandonment feature* as an extension to our initial model. Queues with abandonment have often been studied in order to evaluate the performance of a service system (Zeltyn and Mandelbaum, 2005; Yao, 2016), or to make staffing and routing decisions (Mandelbaum and Zeltyn, 2007, 2009; Koçağa and Ward, 2010). However, in practice, queueing models without abandonment like the M/M/s queue (Erlang-C) are often employed for management issues in call center operations (Koole, 2013). One reason, revealed in the statistical analysis of Robbins et al. (2010), is that the Erlang-C formula gives a pessimistic evaluation of call center performance and therefore results in safe managerial decisions. Thus, our findings on the case without abandonment can be used to take routing decisions when abandonment is difficult to predict. Predicting or anticipating abandonment is particularly challenging. Past waiting experiences (Emadi and Swaminathan, 2017), customers' beliefs and expectations (Veeraraghavan et al., 2018), delay announcements (Akşin et al., 2013, 2016), and learning experience regarding the service speed while waiting (Cui et al., 2018)

influence customers' patience. The difficulty of capturing customers' reasons when they abandon the queue explains why abandonment is generally modeled as an exogenous parameter in most call center studies. In particular, *exponential* distribution serves as a reference in the call center literature (Koole, 2013). Assuming a memoryless distribution for abandonment might seem unrealistic. However, the statistical analysis by Brown et al. (2005) showed the robustness of the M/M/s+M queue (Erlang-A) to fit performance measures reasonably accurately. In this article, we also chose to model abandonment by an exponential distribution.

Finally, a specific feature of our queueing model under an *a posteriori* policy is that decisions are taken based on the experienced waiting time of the oldest customer in the queue. While it is common in call centers to use waiting time as a decision variable, the literature generally focuses on quantity-based policies where the number of customers is the decision variable. This is often due to the difficulty of providing a Markov chain analysis when the wait is the decision variable. To overcome this difficulty, Koole et al. (2012) created a tool to develop Markov decision processes analysis where the first-in-line waiting time is used as a decision variable. Later, Legros et al. (2017) extended this method to queueing models with abandonment. However, complexity of the transition structure makes it complicated to prove the optimality of a threshold policy using this method. In this paper, we tackle the issue by identifying new monotonicity properties of the value function operator, with the first-in-line waiting time as a decision variable. This in turn allows us to prove the optimality of a time-based threshold policy for our optimization problem. The proven monotonicity properties are general and could be used in other queueing contexts involving time-based decisions.

2 The Model and the Routing Problem

Below, we provide the model description and the routing problem. Our assumptions are partly driven by the actual problem that motivates the analysis, and partly by our concern to keep the model as simple as possible. The idea is to obtain an easy-to-implement reservation and outsourcing policy, and to gain insights into the environmental conditions that drive these routing decisions. We consider a system with a single pool of s homogeneous agents and two types of calls, namely, inbound and outbound. We sometimes refer to inbound calls as class-1 customers, and to outbound calls as class-2 (numbered in order of priority: class-1 customers have non-preemptive priority over class-2 customers). Class-1 customers arrive at the system according to a Poisson process with rate λ. If class-1 customers are not routed to the service immediately upon arrival, then either they wait in an infinite capacity queue for their turn to be served, with customers being served in order of arrival, or they are outsourced as explained below. Unlike class-1 customers, we assume that there is an infinite supply of class-2 customers, so an available agent can always serve such a customer, if desired. The service times of all class-i customers ($i = 1, 2$) are assumed to be exponential random variables with rate μ. We denote by a the ratio between the arrival rate and the service rate; $a = \frac{\lambda}{\mu}$.

The call center is engaged by a contract with an outsourcer, whereby a given proportion of class-1 customers, $\overline{P_S}$, can be outsourced per time unit for a given fee, C_{outs}. The call center may decide to outsource fewer calls than the contract would allow. The proportion of outsourced calls in the contract should be chosen in a way as to ensure the stability of the call center; $\lambda(1 - \overline{P_S}) < s\mu$. The call center's revenue is generated by the service

afforded to class-1 and class-2 customers. The revenue generated by the service of a class-2 customer, R_2, is random and depends on the customers' heterogeneity. Therefore, the revenue generated by class-2 customers per time unit is equal to $R_2 \times T$, where T is the random throughput of served class-2 customers. Class-1 calls service may also generate revenue. Unlike class-2 calls, class-1 calls may wait before being served. This wait, denoted by W_S, influences the callers' willingness to accept a purchase offer (Güneş et al., 2010; Lu et al., 2013). While callers' delay sensitivity can be understood through an analysis of abandonment (Akşin et al., 2013), the impact of waiting on purchase probability is not yet well understood in our context. Since customers are initially seeking for a service, their wait may generate frustration from not achieving their goal. This type of frustration may have detrimental consequences on satisfaction and loyalty as well as on the potential to accept an unexpected purchase offer. Therefore, to simplify the analysis, we assume a decreasing and linear relation between the revenue generated by the service for a class-1 call and its wait. The revenue from a class-1 call is then $R_1(1 - \omega W_S)$, for $\omega \geq 0$, where R_1 is a random variable independent of W_S. In view of the contract with the outsourcer, the long-run expected rate of served class-1 calls is $\lambda(1 - P_{\overline{S}})$. The long-run expected revenue per time unit, denoted by $E(G)$, can thus be written as

$$E(G) = r_2 E(T) + r_1 \lambda (1 - P_{\overline{S}})(1 - \omega E(W_S)) - C_{\text{outs}}, \tag{1}$$

where $r_i = E(R_i)$, for $i = 1, 2$, and where $E(X)$ denotes the expected value of a given random variable, X.

The system manager has discretion regarding routing jobs to the various servers and to the outsourcer. Treating the call center as a profit center, the system manager needs to choose a policy that maximizes the expected revenue subject to a limitation for outsourcing a proportion of class-1 calls to the outsourcer. This can be formulated as

$$\begin{cases} \text{maximize } E(G), \\ \text{subject to } P_{\overline{S}} \leq \overline{P_{\overline{S}}}, \end{cases} \tag{2}$$

where $P_{\overline{S}}$ is the proportion of outsourced calls. It is reasonable (although not required) to expect that an optimal policy to solve Problem (2) would be *non-idling for class-1 customers* in the sense that servers may idle only if the queue is empty. The infinite number of class-2 customers could allow a full servers' utilization. However, if all agents are constantly working, all class-1 customers will be delayed in the queue before entering service. This can be avoided if there is idleness in the system, which can be controlled through a reservation strategy. For outsourcing, two classes of policies are considered; *a priori* outsourcing and *a posteriori* outsourcing. With *a priori* outsourcing, the decision to outsource a call is taken at customer's arrival. With *a posteriori* outsourcing, all calls are admitted into the system. The decision to outsource a call is taken if the call has waited too long. We denote the set of policies for reservation and *a priori* outsourcing by Ω_e and the set of policies for reservation and *a posteriori* outsourcing by Ω_l. The letters e and l refer to early (*a priori*) or late (*a posteriori*) outsourcing respectively. Index e and l are also used to indicate whether a policy π belongs to Ω_e or to Ω_l. We attempt to determine the optimal policies in Ω_e or Ω_l, Policy π_e^* and Policy π_l^*, and to compare between them. While the system manager is in a revenue-maximizing perspective, service quality should also be reported when comparing the two policy classes for outsourcing. Thus, the wait

of outsourced callers should not be ignored. Service quality is evaluated by the expected wait, denoted by $E(W)$, of both outsourced and served in-house class-1 calls.

The setting described above allows us to prove the optimal outsourcing and reservation policy (Section 3), and to compare the two policy classes (Section 4). However, some of our assumptions may seem too restrictive. In Section 5, we thus propose investigating some extensions that generalize our analysis. First, we suggest including customer abandonment in the model. We assume that a waiting class-1 customer has finite patience and will abandon if the waiting time exceeds a random time that is exponentially distributed with mean $1/\beta$. In this case, the percentage of abandonment, P_A, is considered as an additional measure of the call center's service quality. Second, we reconsider the relation between the wait and the purchase probability. In the context of retail stores, Lu et al. (2013) show a negative correlation between customers' sensitivity to waiting and price sensitivity. Moreover, Ulku et al. (2017) demonstrate that the consumption quantity increases with the wait. Therefore, customers who can withstand a long wait may consume more with a preference for cheaper products. It is however difficult to conclude whether the revenue per served customer would increase with the wait, especially in our context where customers do not have the initial intention to buy a product. Although a positive relation between the wait and the purchase probability is less likely to happen, by investigating the case $\omega < 0$, our article also aims to provide a routing solution for this case. Finally, we reconsider the assumption of equal service rates for class-1 and class-2 calls. The assumption of equal service rates makes sense in a context where class-2 calls are performed from a list of waiting customers that are a subset of customers who have previously phoned the call center. However, in other contexts, class-1 and class-2 calls may be independent groups of customers with different service rates. To explore this issue, we assume that the service times of all class-i customers are exponential random variables with rate μ_i, $i = 1, 2$.

We conclude our model description with three remarks. First, as mentioned above, the proportion of outsourced calls in the contract, $\overline{P_S}$, is chosen such that the system is stable. Therefore, situations where the optimization problem has no solution should not occur. However, due to mistakes in the forecasting of the arrival rate, instability could happen. If customers are patient and the call center managers wish to avoid too long waits, it is possible to decide for a penalty to pay per outsourced call in case the call center needs to outsource more calls than what was initially decided in the contract. Analysis of this possibility in the contract can be made in a similar way to the one studied in the present article and leads to similar conclusions. We therefore decided not to pursue this analysis.

Second, when agents initiate outbound calls, customers may not pick up the phone directly. This waste of capacity may be significant if agents initiate outbound calls only at service completion. One way to reduce these idling times is to employ an automatic dialer. Specifically, at many modern outbound contact centers, automatic dialers initiate outbound calls even when all agents are busy, using predictive dialing software with the purpose of minimizing agents' idling times (Pang and Perry (2014)). However, automatic dialers are not perfect for estimating the remaining service time of an agent or a customer's availability. Therefore, one unintended consequence of using this software is to drop calls if there is no agent available or to call some customers and make them wait before service. This extension will not be pursued here.

Finally, the arrival rate λ is assumed to be fixed. This is unrealistic since in most service systems, such as

call centers, there is strong variation depending on the time of the day, promotional offers and the customers' history. Moreover, the customers' waiting experience may influence their future behavior with phenomena of retention and acquisition that can affect the arrival rate in the long-run. However, for the real-time routing operations considered in this article, these effects may be ignored. Moreover, if the arrival rate gradually varies relative to the system dynamics, then the call center can be analyzed using a point-wise stationary approximation, where performance at a given time is approximated by the steady state performance of the stationary system with a constant arrival rate (Green and Kolesar, 1991; Jennings et al., 1996). Therefore, the extension of a time-varying or a state-varying arrival rate will not be pursued here.

3 Optimal policy

In this section, we determine the optimal policy to maximize $E(G)$ as defined in Equation (1) for each class of policies for outsourcing (Ω_e and Ω_l). For this purpose, Section 3.1 proves the form of the optimal policy within the sets Ω_e and Ω_l for a given proportion of outsourced calls. Section 3.2 provides the performance measures and their monotonicity properties in the control parameters. This allows us to determine how the search for the control parameters should be initiated. Section 3.3 explains how to compute the control parameters under each policy in order to answer our optimization question.

3.1 Form of the optimal policy

We formulate the routing problem as a Markov decision process (MDP) and next use the value iteration technique to prove the form of the optimal reservation and outsourcing policy. We formulate the problem via the transition structure and the possible actions.

The transition structure. The two classes of policies for outsourcing (Ω_e and Ω_l) require different definitions for the state space. For a policy in Ω_e, let us denote a state of the system by x, where $x \geq -s$. States with $-s \leq x \leq 0$ correspond to an empty queue and $s + x$ busy agents. States with $x > 0$ correspond to the number of class-1 calls waiting in the queue. The transition rate from state x to state x' is denoted as $t_{x,x'}$. So, for $x, x' \geq -s$, we have

$$t_{x,x'} = \begin{cases} \lambda, \text{ if } x' = x + 1, x \geq -s, \\ \min(s, x + s)\mu, \text{ if } x' = x - 1, x > -s, \\ 0, \text{ otherwise,} \end{cases}$$

which corresponds to arrival and service departure rates.

For a policy in Ω_l, the previous system state definition does not allow for decisions based on the experienced waiting time of a given call. To overcome this difficulty, we decided to explicitly model the wait of the first customer in line (FIL) in the queue as in Koole et al. (2012) and Legros et al. (2017). The approach consists of discretizing the FIL waiting time using successive exponential phases, each with rate γ, and then report the waiting phase in the Markov process. Having large values of γ improves the approximation as it gives a better

representation of the continuously elapsing time. As γ tends to infinity, this approximate setup converges to the original one, which leads to an exact analysis. Again, we denote a state of the system by x, where $x \geq -s$. States with $-s \leq x \leq 0$ correspond to an empty queue and $s + x$ busy agents. States with $x > 0$ correspond to a situation where the FIL is waiting at phase x and all agents are busy. The corresponding transition rate from state x to state x' are denoted by $t_{x,x'}$. As Koole et al. (2012) suggested, the transition probabilities denoted by $q_{x,x-h}$ from a waiting phase x to a waiting phase $x-h$, are $q_{x,x-h} = \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h$ and $q_{x,0} = \left(\frac{\gamma}{\lambda+\gamma}\right)^x$ for $x > 0$ and $0 \leq h < x$. So, for $x, x' \geq -s$, we may write

$$t_{x,x'} = \begin{cases} \lambda, \text{ if } x' = x+1, -s \leq x \leq 0, \\ \gamma, \text{ if } x' = x+1, x > 0, \\ (s+x)\mu, \text{ if } x' = x-1, -s < x \leq 0, \\ s\mu q_{x,x-h}, \text{ if } x' = x-h, x > 0, \text{ and } 0 \leq h \leq x, \\ 0, \text{ otherwise,} \end{cases}$$

which corresponds to arrival, service departure and time elapsed.

Possible actions. If the queue is empty, the possible actions for an agent just after a service completion are either to remain idle or to initiate a class-2 call. For a policy in Ω_e, the possible actions at the arrival of a class-1 call when all agents are busy are either to accept the call in the queue or to outsource it from the system. For a policy in Ω_l, the possible actions after an elapsing of time of the FIL are either to maintain the call in the queue or to outsource it from the system.

The value function formulation. For both policy classes, the maximal event rate is bounded. This renders each system uniformizable. We assume without loss of generality that $\lambda + s\mu = 1$ for a policy in Ω_e, and that $\lambda + s\mu + \gamma = 1$ for a policy in Ω_l such that the rate out of each state is equal to 1. We formulate a 2-step value function, in order to separate transitions and actions. We define the dynamic programming value functions $V_k(x)$, $W_k(x)$ and $U_k(x)$ over $k \geq 0$ steps, depending on the state of the system x, $x \geq -s$. The operators U_k and W_k are decision-making operators that represent the class-1 customer outsourcing decision and the class-2 customer initiation decision respectively. We choose $V_0 = U_0 = W_0 = 0$.

Our optimization problem corresponds to a *constrained* MDP. Constrained MDP's can be solved using various techniques. Here, we use one that introduces the constraint in the objective using a Lagrange multiplier, denoted by L. Under weak conditions, it can be seen that the optimal stationary policy for a certain Lagrange multiplier is optimal for the constrained problem if the value of the constraint under this policy attains exactly the desired proportion of outsourced calls (Altman, 1999). This means that the Lagrange multiplier L controls the proportion of outsourced calls $P_{\overline{S}}$ and should be chosen such that $E(G)$ is maximized.

The costs and rewards involved in $E(G)$ are counted at service initiation or outsourcing epochs. For Ω_e and Ω_l, service initiations occur at λ-transitions from states with vacant servers, and $s\mu$-transitions from states $x > 0$. From states with vacant servers (i.e., for $-s \leq x < 0$), a call starting service does not wait. Therefore, a reward of r_i is counted per served class-i call, $i = 1, 2$. For Ω_e, the waiting time of a class-1

customer who starts service is difficult to estimate in state $x > 0$. However, since customers arrive one by one, the queue length is identical, in distribution, at arrival times and at service initiation epochs. Just after a service initiation from state x, the expected wait of an arriving customer is $\frac{x}{s\mu}$. So, a reward of $r_1\left(1-\omega\frac{x}{s\mu}\right)$ is counted per class-1 call served from state $x > 0$. For Ω_l, the expected duration of a waiting phase is $1/\gamma$. Therefore, a customer who starts service from state $x > 0$ has already waited $\frac{x}{\gamma}$ time units. So, a reward of $r_1\left(1-\omega\frac{x}{\gamma}\right)$ is counted per class-1 call served from state $x > 0$. Finally, the cost L is counted per outsourced call in Ω_e and Ω_l. We choose not to express C_{outs} in the value function because this element is constant and cannot be optimized. Therefore, it does not influence the routing decisions.

For Ω_e, we may then write for $k \geq 0$ and $x \geq -s$,

$$V_{k+1}(x) = \lambda U_k(x) + \min(s, s+x)\mu\left(W_k(x-1) + r_1 \mathbb{1}_{x>0}\left(1-\omega\frac{x}{s\mu}\right)\right) + (1-\lambda-\min(s,s+x)\mu)W_k(x), \tag{3}$$

where the notation $\mathbb{1}_{x\in A}$ is used to express the indicator function of a given subset A, with
$U_k(x) = V_k(x+1) + r_1$ if $-s \leq x < 0$, and $U_k(x) = \max(V_k(x) - L, V_k(x+1))$ if $x \geq 0$,
$W_k(x) = \max(V_k(x), V_k(x+1) + r_2)$ if $-s \leq x < 0$, and $W_k(x) = V_k(x)$ if $x \geq 0$.

For Ω_l, we denote by F the operator on the set of functions f from $\mathbb{Z}$ to $\mathbb{R}$ defined by $F(f(x)) = \sum_{h=0}^{x} q_{x,x-h} f(x-h)$ for $x > 0$, and $F(f(x)) = f(x)$ for $x \leq 0$. This operator is used to simplify the notations. It represents the possible changes in the state of the FIL when either an outsourcing or a service completion occurs. We may thus write, for $k \geq 0$,

$$\begin{aligned} V_{k+1}(x) &= \lambda U_k(x) + (s+x)\mu W_k(x-1) + (1-\lambda-(s+x)\mu)W_k(x), \text{ for } -s \leq x \leq 0, \text{ and,} \\ V_{k+1}(x) &= \gamma U_k(x) + s\mu\left(F(W_k(x)) + r_1\left(1-\omega\frac{x}{\gamma}\right)\right) + (1-\gamma-s\mu)W_k(x), \text{ for } x > 0, \text{ with} \end{aligned} \tag{4}$$

$U_k(x) = V_k(x+1) + r_1$ if $-s \leq x < 0$, and $U_k(x) = \max(F(V_k(x)) - L, V_k(x+1))$ if $x \geq 0$,
$W_k(x) = \max(V_k(x), V_k(x+1) + r_2)$ if $-s \leq x < 0$, and $W_k(x) = V_k(x)$ if $x \geq 0$.

One way of obtaining the long-run average optimal actions is to use the value iteration technique, by recursively evaluating V_k, for $k \geq 0$. As k tends to infinity, the optimal policy converges to the unique average optimal policy. Moreover, the optimal long-run policy is independent of the choice of V_0. The convergence is due to the aperiodic irreducible finite-state Markov chains considered here (e.g., see Theorem 8.5.3 part c of Puterman (1994)). In Theorem 1, through induction on the value function, we prove that the optimal policy for the two policy classes is of threshold type. For Ω_e, we prove that the value function V_k is decreasing and concave. For Ω_l, we instead need to show that $V_k(x+1) - F(V_k(x))$ is decreasing in x for $x \geq 0$. The latter property is referred to as *general concavity*.

Theorem 1. *The optimal policy for outsourcing and reservation within Ω_e and Ω_l is of threshold type.*

Theorem 1 allows us to specify the formulation of the two optimal outsourcing policies and the optimal reservation policy.

- **Reservation threshold policy for Policies π_e^* and π_l^***

We denote by c the threshold of number of agents reserved for class-1 customers, $0 \leq c \leq s$. Consider an idle agent just after a service completion. If the number of idle agents (excluding the idle agent considered) is at least c, then this agent initiates the service of a class-2 customer. Otherwise, she remains idle. In other words, there are c agents that are reserved for class-1 calls, so, there are at least $s - c$ agents working at any time.

- **Call outsourcing policies**

  - *a priori* outsourcing threshold policy (for Policy π_e^*). The decision to allow an arriving class-1 call to join the queue is based on the current number of customers in the queue when all agents are busy. If this number is strictly lower than a certain threshold n ($n \geq 0$) and all agents are busy, then the arriving class-1 customer is allowed to join the queue. Otherwise, it is outsourced by the system.

  - *a posteriori* outsourcing threshold policy (for Policy π_l^*). With outsourcing *a posteriori*, all class-1 customers are allowed to join the queue, regardless of the system state. However, the system does not allow class-1 calls to infinitely stay in the queue. A call waiting in the queue for exactly τ ($\tau \geq 0$) time units is automatically outsourced.

3.2 Performance evaluation

We now evaluate the performance measures which constitute the expected cost, $E(G)$, and the expected waiting time, $E(W)$ for Policy π_e^* and Policy π_l^*. In addition, we evaluate the waiting time distribution of served customers, $P(W_S > t)$, for $t \geq 0$. The latter performance will be considered in Section 4 to compare the two policy types. To express the performance measures, we use similar building blocks as in Zeltyn and Mandelbaum (2005); ϵ, J, J_1, J_H, and $J(t)$. These building blocks were used to express abandonment behavior. In our model, we instead consider outsourcing control which cannot be wholly assimilated with abandonment behavior. This explains why we do not have a common expression for $P(W_S > t)$ for the two policy classes and why $J(0) \neq J$ for Policy π_e^*. The performance measures are given by

$$P_{\overline{S}} = \frac{1 + (\lambda - s\mu)J}{\epsilon + \lambda J}, \quad E(T) = \lambda \frac{\binom{s-1}{c}}{\epsilon + \lambda J}, \quad E(W) = \frac{\lambda J_H}{\epsilon + \lambda J}, \text{ and, } E(W_S) = \frac{s\mu J_1 - J}{\epsilon + s\mu J - 1},$$

where the notation $\binom{n}{k}$ is used to express the binomial coefficient with integer parameters n and k ($0 \leq k \leq n$). Finally, for $t > 0$, we have

$$P(W_S > t) = \frac{\lambda J(t)}{\epsilon + s\mu J - 1}, \text{for Policy } \pi_e^*, \text{ and, } P(W_S > t) = \frac{\lambda J(t) - 1 - (\lambda - s\mu)J}{\epsilon + s\mu J - 1},$$

for Policy π_l^*. For both policy classes, we have

$$\epsilon = \frac{\sum_{x=0}^{c-1} \frac{a^x}{(s-c+x)!}}{\frac{a^{c-1}}{(s-1)!}}.$$

In Table 1, we specify the other building blocks. The derivation of the performance measures follows from a Markov chain analysis. For Policy π_l^*, the approximated model is considered. We next obtain the exact

Table 1: Building blocks

	With Policy π_e^*:	With Policy π_l^*:
J	$\frac{1}{s\mu}\frac{1-\left(\frac{a}{s}\right)^{n+1}}{1-a/s}$	$\frac{1}{s\mu}\frac{1-\frac{a}{s}e^{-\tau(s\mu-\lambda)}}{1-a/s}$
J_1	$\frac{1}{(s\mu)^2}\frac{1-(n+2)\left(\frac{a}{s}\right)^{n+1}+(n+1)\left(\frac{a}{s}\right)^{n+2}}{(1-a/s)^2}$	$\frac{1}{(s\mu)^2}\frac{1-\left(1+\left(1-\frac{a}{s}\right)(1+s\mu\tau)\right)\frac{a}{s}e^{-\tau(s\mu-\lambda)}}{(1-a/s)^2}$
J_H	$\frac{1}{(s\mu)^2}\frac{1-(n+1)(a/s)^n+n(a/s)^{n+1}}{(1-a/s)^2}$	$\frac{1}{(s\mu)^2}\frac{1-(1+\frac{a}{s}\tau(s\mu-\lambda))e^{-\tau(s\mu-\lambda)}}{(1-a/s)^2}$
$J(t)$	$\frac{1}{s\mu}\frac{e^{-s\mu t}}{1-a/s}\sum_{x=0}^{n-1}\frac{(s\mu t)^x((a/s)^x-(a/s)^n)}{x!}$	$\mathbb{1}_{t<\tau}\frac{1}{s\mu}\frac{e^{-t(s\mu-\lambda)}-\frac{a}{s}e^{-\tau(s\mu-\lambda)}}{1-a/s}$

performance measures by letting the elapsing of time rate, γ, tend to infinity. The details of this analysis are omitted. The computation of the optimal thresholds is based on the monotonicity properties of the performance measures as given in Theorem 2. When the second order monotonicity property is not specified, it means that the performance considered is neither convex nor concave. The results of Theorem 2 are also interesting from a queueing perspective. With $c = s$, our model is reduced to a Markovian queue with deterministic reneging under Policy π_l^* -referred to in the queueing literature as the M/M/s+D queue- or to a finite capacity Markovian queue under Policy π_e^* -referred to in the queueing literature as the M/M/s/s+n queue-. The convexity results obtained allow us to retrieve existing results for the M/M/s/s+n queue and to derive new results for the M/M/s+D queue.

Theorem 2. *The following holds:*

- *The expected throughput of class-2 customers, $E(T)$, is decreasing in c and decreasing and convex in n (Policy π_e^*) and in τ (Policy π_l^*),*
- *The proportion of outsourced callers, $P_{\overline{S}}$, is decreasing and convex in c and decreasing and convex in n (Policy π_e^*) and in τ (Policy π_l^*),*
- *The expected waiting time of served class-1 customers, $E(W_S)$, the expected waiting of served and outsourced customers, $E(W)$, and the proportion of served callers who wait more than t, $P(W_S > t)$, are decreasing and convex in c and increasing in n (Policy π_e^*) and in τ (Policy π_l^*).*

Given that $P_{\overline{S}}$ is decreasing in the outsourcing thresholds, the $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$ relation induces a relation between the outsourcing and the reservation thresholds. We obtain

$$n \geq \frac{\ln\left(\overline{P_{\overline{S}}}\cdot\frac{1+(1-a/s)\sum_{x=0}^{c-1}\frac{s!}{(s-c+x)!a^{c-x}}}{1-\frac{a}{s}\left(1-\overline{P_{\overline{S}}}\right)}\right)}{\ln(a/s)}, \text{ for Policy } \pi_e^*, \text{ and } \tau \geq -\frac{\ln\left(\overline{P_{\overline{S}}}\cdot\frac{1+(1-a/s)\sum_{x=0}^{c-1}\frac{s!}{(s-c+x)!a^{c-x}}}{1-\frac{a}{s}\left(1-\overline{P_{\overline{S}}}\right)}\right)}{s\mu-\lambda}, \quad (5)$$

for Policy π_l^*. Inequality (5) will be used in Section 3.3 to initiate the algorithm for the computation of the optimal thresholds.

Special Case: When inbound callers are insensitive to their wait (i.e., $\omega = 0$). Using Theorem 2, Proposition 1 reveals that when callers' purchase willingness is insensitive to their waiting time (i.e., $\omega = 0$),

then reservation should be excluded. In other words, agents should work full time on inbound or outbound calls.

Proposition 1. *When $\omega = 0$, the following holds:*

- *If $\frac{a}{s} \geq 1$, or if $\frac{a}{s} < 1$ and $r_2 > r_1$, then it is optimal to have $c = 0$, and $n = \frac{\ln\left(\frac{\overline{P_{\overline{S}}}}{1-\frac{a}{s}\left(1-\overline{P_{\overline{S}}}\right)}\right)}{\ln(a/s)}$ for Policy π_e^* or $\tau = -\frac{\ln\left(\frac{\overline{P_{\overline{S}}}}{1-\frac{a}{s}\left(1-\overline{P_{\overline{S}}}\right)}\right)}{s\mu-\lambda}$ for Policy π_l^*.*
- *Otherwise, if $\frac{a}{s} < 1$ and $r_2 \leq r_1$, then it is optimal to have $c = 0$ and $n = \tau = \infty$.*

3.3 Computation of the optimal policy

The constraint $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$ indicates that the proportion of outsourced calls should be optimized in the interval $[0, \overline{P_{\overline{S}}}]$. Therefore, in Equations (3) and (4) of Section 3.1, several values for the Lagrange multiplier L should be tested until the maximal expected revenue is reached. This procedure might be long given that for each chosen value of L, we should let k tend to infinity in Equations (3) and (4). Instead, we adopted an algorithmic approach where only a *finite number* of thresholds is tested before reaching their optimal values. For the reservation threshold, c, an exhaustive evaluation is possible since the threshold c can only take $s+1$ values. These values are the integers in the interval $[0, s]$. However, for the outsourcing thresholds, n or τ, an infinite number of values is possible. This renders an exhaustive search inapplicable.

To overcome this difficulty, we formulate an n-terminating problem as in Koçağa and Ward (2010) and Adusumilli and Hasenbein (2010). This consists of expressing the long-run dynamic programming optimality equations for the relative value function, $V^c(x)$, for $x \geq -c$ and the average constant $E(G)^c$ for a given reservation threshold, c, under both policy classes. At this step, we do not consider the constraint $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$. Therefore, we chose $L = 0$. For Policy π_l^*, we consider the approximated model used in Section 3.1. Under both policies, we have

$$V^c(-c) + E(G)^c = \lambda(V^c(-c+1) + r_1) + (s-c)\mu r_2 + (1-\lambda)V^c(-c), \text{ for } x = -c, \tag{6}$$

$$V^c(x) + E(G)^c = \lambda(V^c(x+1) + r_1) + (s+x)\mu V^c(x-1) + (1-\lambda-(s+x)\mu)V^c(x), \text{ for } -c < x < 0,$$

$$V^c(0) + E(G)^c = \lambda \max(V^c(1), V^c(0)) + s\mu V^c(-1) + (1-\lambda-s\mu)V^c(0), \text{ for } x = 0.$$

For $x > 0$, and Policy π_e^*, we have

$$V^c(x) + E(G)^c = \lambda \max(V^c(x+1), V^c(x)) + s\mu\left(V^c(x-1) + r_1\left(1-\omega\frac{x}{s\mu}\right)\right) + (1-\lambda-s\mu)V^c(x).$$

For $x > 0$, and Policy π_l^*, we have

$$V^c(x) + E(G)^c = \gamma \max(V^c(x+1), F(V^c(x))) + s\mu\left(F(V^c(x)) + r_1\left(1-\omega\frac{x}{\gamma}\right)\right) + (1-\gamma-s\mu)V^c(x).$$

We introduce the relative cost difference defined as $\Delta^c(x) = V^c(x) - V^c(x-1)$ for Policy π_e^* and $\Delta^c(x) = V^c(x) - F(V^c(x-1))$ for Policy π_l^*, for $x > -c$. For Policy π_l^*, using the notation $u = \frac{\lambda}{\lambda+\gamma}$, we have

$$V^c(x) - F(V^c(x)) = V^c(x) - \sum_{k=0}^{x-1} u(1-u)^k V^c(x-k) - (1-u)^x V^c(0) = (1-u)\Delta^c(x),$$

for $x > 0$. Subsequently, we rewrite Equation (6) in terms of $\Delta^c(x)$. Under both policies, we may write

$$\begin{aligned}
&E(G)^c - g(x) = \lambda\Delta^c(x+1) - (s+x)\mu\Delta^c(x), \text{ for } -c \le x < 0, \\
&E(G)^c - g(0) = \lambda \max(\Delta^c(1), 0) - s\mu\Delta^c(0), \text{ for } x = 0, \\
&E(G)^c - g(x) = \lambda \max(\Delta^c(x+1), 0) - s\mu\Delta^c(x), \text{ for Policy } \pi_e^*, \text{ and } x > 0, \\
&E(G)^c - g(x) = \gamma \max(\Delta^c(x+1), 0) - (s\mu+\gamma)(1-u)\Delta^c(x), \text{ for Policy } \pi_l^*, \text{ and } x > 0,
\end{aligned} \tag{7}$$

with $\Delta^c(-c) = 0$, and where $g(x)$ is the reward function defined as

$$g(x) = \begin{cases}
\lambda r_1 + (s-c)\mu r_2, \text{ for } x = -c, \\
\lambda r_1, \text{ for } -c < x < 0, \\
0, \text{ for } x = 0, \\
s\mu r_1 \left(1 - \omega \dfrac{x}{s\mu}\right) \text{ for Policy } \pi_e^*, \text{ and } s\mu r_1 \left(1 - \omega \dfrac{x}{\gamma}\right) \text{ for Policy } \pi_l^*, \text{ for } x > 0.
\end{cases} \tag{8}$$

Theorem 1 proves that the optimal policy for call outsourcing is of threshold type. Therefore, the solution of Equation (7) is given by the relative difference $\Delta^{c,n}(x)$, for $-c \le x \le n$, and the expected revenue $E(G)^{c,n}$ both depending on the reservation threshold, c, and on the outsourcing threshold, n, such that for both policies we have

$$\begin{aligned}
&E(G)^{c,n} - g(x) = \lambda\Delta^{c,n}(x+1) - (s+x)\mu\Delta^{c,n}(x), \text{ for } -c \le x \le 0, \\
&E(G)^{c,n} - g(x) = \lambda\Delta^{c,n}(x+1) - s\mu\Delta^{c,n}(x), \text{ for Policy } \pi_e^*, \text{ and } 0 < x \le n, \\
&E(G)^{c,n} - g(x) = \gamma\Delta^{c,n}(x+1) - (s\mu+\gamma)(1-u)\Delta^{c,n}(x), \text{ for Policy } \pi_l^*, \text{ and } 0 < x \le n,
\end{aligned} \tag{9}$$

where $\Delta^{c,n}(-c) = \Delta^{c,n}(n+1) = 0$. Recall that for Policy π_l^*, the outsourcing threshold is a positive real, τ. In Equation (9), we approximate the deterministic duration, τ, by an Erlang distribution with n phases and rate γ per phase. By relating n and τ via $\frac{n}{\gamma} = \tau$, this Erlang distribution converges to the deterministic one as n and γ tend to infinity. In this way, the same notation, n, can be used for both policies.

The relations in (9) define a system of linear equations which can be solved explicitly. Using an induction step, we can show after some algebra that

$$\Delta^{c,n}(-c+x) = \left(\frac{E(G)^{c,n}}{\lambda} - r_1\right) \sum_{i=0}^{x-1} \frac{a^{-i}(s-c+x-1)!}{(s-c+x-1-i)!} - r_2 \frac{a^{-x}(s-c+x-1)!}{(s-c-1)!},$$

for $1 \le x \le c$, for Policy π_e^* and Policy π_l^*. For Policy π_e^*, for $x > 0$, we obtain

$$\begin{aligned}
\Delta^{c,n}(x) = {} & \frac{E(G)^{c,n}}{\lambda} \left(\frac{a}{s}\right)^{1-x} \left(\sum_{i=0}^{c} \frac{a^{-i}s!}{(s-i)!} + \frac{a}{s} \frac{1 - \left(\frac{a}{s}\right)^{x-1}}{1 - \frac{a}{s}}\right) - r_2 \left(\frac{a}{s}\right)^{1-x} \frac{a^{-(c+1)}s!}{(s-c-1)!} \\
& - r_1 \left(\frac{a}{s}\right)^{1-x} \left(\sum_{i=1}^{c} \frac{a^{-i}s!}{(s-i)!} + \frac{1 - \left(\frac{a}{s}\right)^{x-1}}{1 - \frac{a}{s}} - \frac{\omega}{s\mu} \frac{1 - x\left(\frac{a}{s}\right)^{x-1} + (x-1)\left(\frac{a}{s}\right)^{x}}{\left(1 - \frac{a}{s}\right)^2}\right).
\end{aligned} \tag{10}$$

For Policy π_l^*, we introduce the notation $a_\gamma = s\frac{\lambda+\gamma}{s\mu+\gamma}$. For $x > 0$, we get

$$\Delta^{c,n}(x) = \frac{E(G)^{c,n}}{\lambda}\left(\frac{a_\gamma}{s}\right)^{1-x}\left(\sum_{i=0}^{c}\frac{a^{-i}s!}{(s-i)!} + \left(1+\frac{\lambda}{\gamma}\right)\frac{a}{s}\frac{1-\left(\frac{a_\gamma}{s}\right)^{x-1}}{1-\frac{a}{s}}\right) - r_2\left(\frac{a_\gamma}{s}\right)^{1-x}\frac{a^{-(c+1)}s!}{(s-c-1)!} \tag{11}$$
$$- r_1\left(\frac{a_\gamma}{s}\right)^{1-x}\left(\sum_{i=1}^{c}\frac{a^{-i}s!}{(s-i)!} + \frac{s\mu}{\gamma}\frac{a_\gamma}{s}\frac{1-\left(\frac{a_\gamma}{s}\right)^{x-1}}{1-\frac{a_\gamma}{s}} - \frac{\omega}{\gamma}\frac{a_\gamma}{s}\frac{1-x\left(\frac{a_\gamma}{s}\right)^{x-1}+(x-1)\left(\frac{a_\gamma}{s}\right)^{x}}{\left(1-\frac{a_\gamma}{s}\right)^2}\right).$$

The expected revenue, $E(G)^{c,n}$, can be obtained by solving $\Delta^{c,n}(n+1) = 0$. This allows us to retrieve the expression of $E(G)$ in Section 3.2 directly for Policy π_e^* and after letting γ and n tend to infinity for Policy π_l^*.

Using the result of Lemma 1, Theorem 3 proves that the first local maximum of $E(G)^{c,n}$ found by increasing n is the optimal outsourcing threshold.

Lemma 1. *If $E(G)^{c,n_1} \geq E(G)^{c,n_2}$ for $n_1, n_2 \in \mathbb{N}$, then*

$$\Delta^{c,n_1}(x) \geq \Delta^{c,n_2}(x), \text{ for } -c+1 \leq x \leq \min(n_1,n_2)+1. \tag{12}$$

Theorem 3. *If there exists a solution to Equation (9) with $E(G)^{c,m} > E(G)^{c,k}$, for $0 \leq k \leq m-1$ and $E(G)^{c,m+1} < E(G)^{c,m}$, then for all $n > m$, we have $E(G)^{c,m} \geq E(G)^{c,n}$.*

We may encounter a situation where $E(G)^{c,n}$ is increasing in n. This means that it is optimal to serve all inbound calls in-house. In this case, in Proposition 2, we provide a stopping criterion for the search for the optimal outsourcing threshold.

Proposition 2. *If $E(G)^{c,n}$ is increasing in n, then $E(G)^{c,\infty} - E(G)^{c,n} \leq \lambda\Delta^{c,n}(n)$, for Policy π_e^*, and $E(G)^{c,\infty} - E(G)^{c,n} \leq \gamma\Delta^{c,n}(n)$, for Policy π_l^*.*

We are now in a position to establish an algorithm to compute the optimal outsourcing threshold. Inequality (5) allows us to determine, for each reservation threshold c, the lowest possible outsourcing threshold such that the constraint $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$ is satisfied. Moreover, Theorem 2 proves that $P_{\overline{S}}$ is decreasing in the outsourcing thresholds. Therefore, by increasing n, the constraint $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$ remains satisfied. While increasing the outsourcing threshold, the result of Theorem 3 indicates that the first local maximum for the expected revenue is also the global one. In the increasing case, Proposition 2 provides a stopping criterion for the search of the optimal outsourcing threshold. Therefore, for each c, we can determine the optimal outsourcing threshold, n_c, after a finite number of iteration. The optimal reservation threshold, c, is then $c^* = \underset{c=0,1\cdots,s}{\arg\max}\, E(G)^{c,n_c}$.

The algorithm is as follows:

Algorithm 1: Computation of the optimal outsourcing threshold for reservation threshold c.

1. *Initialisation.* Set n_c as the lowest integer such that Inequality (5) is respected and compute $E(G)^{c,n_c}$, $E(G)^{c,\infty}$, and $\Delta^{c,n_c}(n_c)$ using (10) or (11).

2. *Iteration step:* Increase n_c by one and compute $E(G)^{c,n_c}$ and $\Delta^{c,n_c}(n_c)$ using (10) or (11).

If $E(G)^{c,n_c} < E(G)^{c,n_c-1}$, then the outsourcing threshold $n_c - 1$ is optimal.

If $E(G)^{c,n_c} \geq E(G)^{c,n_c-1}$, then

- If $E(G)^{c,\infty} - E(G)^{c,n} \leq \lambda\Delta^{c,n}(n)$ for Policy π_e^*, or $E(G)^{c,\infty} - E(G)^{c,n} \leq \gamma\Delta^{c,n}(n)$ for Policy π_l^*, then it is optimal not to outsource any customer (i.e., $n_c = \infty$ is optimal).
- Otherwise, go back to the Iteration step.

4 Comparison between outsourcing policies

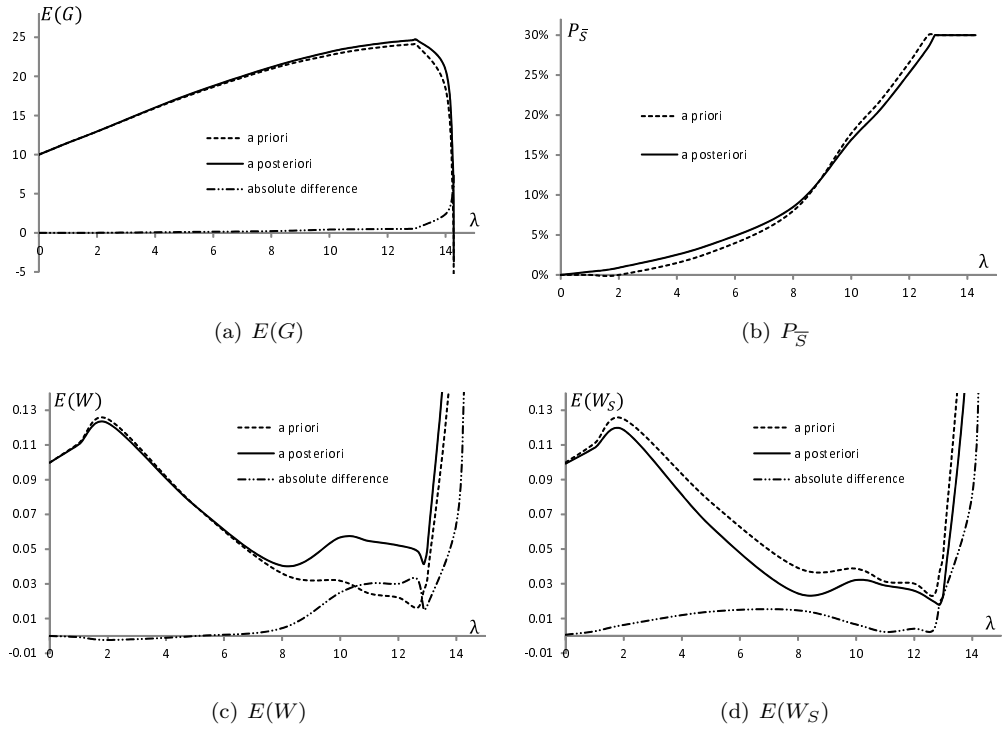
The main result is that Policy π_l^* outperforms Policy π_e^* for Problem (2). This result is proven in Theorem 4. The first point of Theorem 4 is given to qualify our main result. Although Policy π_l^* outperforms Policy π_e^* from a revenue maximizer perspective, the improvement is detrimental to the quality of service measured by $E(W)$ of both served and outsourced customers when the two policy classes have the same reservation threshold and the same proportion of outsourced calls.

Theorem 4.

1. *For a given reservation threshold and a given proportion of outsourced calls, the random variable W_S is highest for Policy π_e^* under the usual stochastic ordering and the expected waiting time of all customers (served in-house or outsourced), $E(W)$, is lowest for Policy π_e^*.*
2. *$E(G)$ is maximized for Policy π_l^*.*

In the following illustrations, we provide some numerical experiments to compare Policy π_e^* and Policy π_l^*. The aim is to determine in which contexts the difference between the two policy classes may be significant.

Effect of the congestion. In Figure 1, we evaluate the two policies in terms of expected revenue, expected waiting time, and proportion of outsourced calls as functions of the class-1 arrival rate. As the workload increases, the difference between the two policies also increases (Figure 1(a)). This can be explained by the increase in the proportion of outsourced calls (Figure 1(b)). Figure 1(b) also reveals that the proportion of outsourced calls is very close under the two policies. This information, taken together with the observation of very close reservation thresholds under the two policies, indicates that the conditions of the first point of Theorem 4 are close to be respected in most cases. This validates the idea that Policy π_l^* is detrimental to the service quality measured by $E(W)$. Note that counterexamples can be found with very low arrival rates (see Figure 1(c) for $\lambda = 2$). As shown in Figures 1(c) and 1(d), the difference between the two policies for $E(W)$ and $E(W_S)$ is not monotonous as a function of the arrival rate. We should recall that the optimal control parameters are chosen to optimize $E(G)$ and not $E(W)$. This explains the irregular behavior of $E(W)$ as a function of λ. In some situations, the control parameters are chosen to encourage the service of class-2 calls, with longer waits for class-1 customers, while in other situations, class-1 calls are given shorter wait times, and the service initiation of class-2 calls is restricted. We observe however that the difference between the two policies tends to increase with workload in high workload situations.



(a) $E(G)$ (b) $P_{\overline{S}}$

(c) $E(W)$ (d) $E(W_S)$

Figure 1: Comparison between the two policy classes ($s = 10$, $\mu = 1$, $r_1 = 3$, $r_2 = 1$, $\omega = 1$, $\frac{C_{outs}}{\lambda \overline{P_S}} = 1/2$, $\overline{P_S} = 30\%$)

Effect of the call center size. Table 2 compares the expected revenue, $E(G)$, and the expected wait, $E(W)$, for the two policy classes and different call center sizes. We chose $a/s = 0.8, 1$ and 1.2 to reflect different congestion situations. Columns 5 and 6 give the relative difference in revenue, RD_G, defined as $RD_G = \frac{E(G)_{\text{Policy } \pi_l^*} - E(G)_{\text{Policy } \pi_e^*}}{|E(G)_{\text{Policy } \pi_e^*}|}$, and the absolute difference in revenue, AD_G, defined as $AD_G = E(G)_{\text{Policy } \pi_l^*} - E(G)_{\text{Policy } \pi_e^*}$. Columns 7 and 8 provide the expected revenue generated per agent and per time unit. The last three columns specify the expected wait and the absolute difference in service quality, AD_W, defined as $AD_W = E(W)_{\text{Policy } \pi_l^*} - E(W)_{\text{Policy } \pi_e^*}$. We also specify the cost per outsourced call if the constraint in the contract is saturated; $\frac{C_{\text{outs}}}{\lambda \overline{P_S}}$. The table reveals that the difference in revenue between the two outsourcing

Table 2: Performance comparison ($\mu = 1$, $r_1 = 3$, $r_2 = 1$, $\omega = 1$, $\frac{C_{outs}}{\lambda \overline{P_S}} = 1/2$, $\overline{P_S} = 20\%$)

Parameters		$E(G)$				$\frac{E(G)}{s}$		$E(W)$		
a/s	s	Policy π_l^*	Policy π_e^*	RD_G	AD_G	Policy π_l^*	Policy π_e^*	Policy π_l^*	Policy π_e^*	AD_W
0.8	1	0.94	0.48	94.178%	0.454	0.94	0.48	0.741	0.566	0.175
1	1	-0.39	-1.28	69.850%	0.897	-0.39	-1.28	1.483	1.195	0.288
1.2	1	-9.25	-11.74	21.218%	2.491	-9.25	-11.74	4.766	4.027	0.739
0.8	10	21.67	21.39	1.339%	0.286	2.17	2.14	0.058	0.039	0.019
1	10	23.75	23.21	2.335%	0.542	2.38	2.32	0.048	0.025	0.024
1.2	10	18.04	15.83	13.943%	2.207	1.80	1.58	0.400	0.335	0.065
0.8	50	120.39	120.24	0.124%	0.149	2.41	2.40	0.021	0.020	0.001
1	50	132.23	131.77	0.347%	0.458	2.64	2.64	0.022	0.018	0.004
1.2	50	138.19	136.62	1.144%	1.563	2.76	2.73	0.014	0.008	0.007
0.8	200	497.79	497.78	0.001%	0.004	2.49	2.49	0.000	0.000	0.000
1	200	553.93	553.09	0.153%	0.844	2.77	2.77	0.008	0.007	0.001
1.2	200	567.13	565.71	0.251%	1.421	2.84	2.83	0.006	0.005	0.002
0.8	400	1001.79	1001.79	0.000%	0.000	2.50	2.50	0.000	0.000	0.000
1	400	1122.59	1122.13	0.042%	0.468	2.81	2.81	0.002	0.001	0.000
1.2	400	1141.73	1140.47	0.110%	1.260	2.85	2.85	0.004	0.002	0.001

policies is significant (i) in small call centers, (ii) in congested situations, and (iii) when the proportion of

outsourced customers chosen is high. In large call centers, the effect of the service time variability is reduced, and agents have greater efficiency as shown in Columns 7 and 8. Customers' wait is thus better controlled, and the relative improvement obtained by choosing a good routing strategy is reduced. As shown in the last three columns, the difference in service quality is also the most substantial in small call centers. In large call centers, the control parameters are adjusted in a way which cancels the wait.

This analysis allows us to specify the domain of applicability of our study. With more than 50 agents present, the two policies are virtually the same. The improvement which can be obtained by implementing Policy π_l^* instead of Policy π_e^* is marginal with less than 1% difference. This means that the results of our study mostly reflect small call centers or call centers organized in small independent teams. Examples of small-size call centers organization can be found for helpdesks of very specialized services, where agents might need special tools to solve the client's problems. In banks also, the management of large accounts requires small teams of specially trained agents. In general, in the Business-to-Business environment, call centers are usually small as compared to the Business-to-Consumer sector (Chevalier and Van den Schrieck, 2008). For management reasons, some large call centers choose to be organized in smaller teams with identical skills. Although the beneficial pooling effect is reduced in smaller systems, the human resource management can be performed in a much better way. Agents' motivation and responsibility would increase. For instance Bouygues Telecom decided to adopt a small-team organization. For this call center, the number of agents simultaneously present is in the order of 1000 and the corresponding number of agents present in each team would be ranging from 20 to 50 (Jouini et al., 2008). Our study provides a valuable decision-support tool for managing outsourcing and reservation decisions for this type of environment.

Routing solutions for extreme workload situations. We now focus further on extreme workload cases. This analysis may help to explain routing practices commonly adopted in call centers. In Table 3, using Taylor expansions, we provide equivalent expressions of the performance measures when a is in the neighborhood of ∞ and in the neighborhood of 0. Let us start with *high workload* situations. In both policy classes, $P_{\overline{S}}$ is

Table 3: Equivalent expressions of the performance measures

	a is in the neighborhood of ∞		a is in the neighborhood of 0	
	Policy π_e^*	Policy π_l^*	Policy π_e^*	Policy π_l^*
$E(T)$	$\frac{\mu(s-c)s!s^n}{(s-c)!}a^{-(c+n)}$	$\frac{\mu(s-c)s!}{(s-c)!}a^{-c}e^{-\tau(s\mu-\lambda)}$	$\mu(s-c)$	
$P_{\overline{S}}$	$\left[\sum_{x=0}^{c}\frac{s!}{(s-c+x)!a^{c-x}}\right]^{-1}$		$\frac{(s-c)!}{s!s^n}a^{c+n}$	$\frac{(s-c)!}{s!}a^c e^{-\tau s\mu}$
$E(W_S)$	$\frac{n}{s\mu}$	τ	$\frac{1}{s\mu}\frac{(s-c)!a^c}{s!}$	

insensitive to the outsourcing threshold. Moreover, $E(T)$ is decreasing and $E(W_S)$ is increasing in n (Policy π_e^*) and in τ (Policy π_l^*). Therefore, the outsourcing thresholds should be chosen as low as possible. For this purpose, the constraint for the proportion of outsourced calls should be saturated (i.e., $P_{\overline{S}} = \overline{P_{\overline{S}}}$) as observed in Figure 1(b).

We now consider *low workload* situations. For both policies, $E(T)$ and all the performance measures related to the waiting time are only controlled by c. The only measure that depends on the outsourcing parameters

is $P_{\overline{S}}$. Since $P_{\overline{S}}$ is decreasing in n and in τ, $n = \infty$ and $\tau = \infty$ are optimal (no outsourcing). In Proposition 3, we show that either $c = 0$ or $c = 1$ is optimal.

Proposition 3. *For Policy π_e^* or Policy π_l^* in low workload situations, $n = \infty$ or $\tau = \infty$ is optimal and if $\mu r_2 \geq \frac{a}{s} r_1 \omega$, then $c = 0$ is optimal. Otherwise, $c = 1$ is optimal.*

Therefore, in low workload situations, it is not optimal to outsource calls and reservation should be limited to one agent at most. This result is intuitive; with too many agents there is no need to outsource and reservation should be limited. The reason why $c = 0$ is not necessarily optimal is because with $c = 0$ all class-1 callers have to wait. So, even in a low workload situation, if the service times are long it might be beneficial to have at least one idle agent to avoid waiting.

This analysis of extreme workload situations may partially confirm some common intuitions in call center management. Reservation and outsourcing do not seem to meet the same environmental conditions. Initiating outbound calls is generally considered by managers as a way to use overstaffing capacity. With too many resources, outsourcing no longer appears useful. Instead, outsourcing is used to better manage congested situations when the call center's resources cannot handle the flow of arriving customers.

5 Robustness of the analysis

This section develops different natural extensions of the initial model. Section 5.1 considers the feature of abandonment. Section 5.2 explores the consequences of the wait having a positive impact on purchase behavior in a context of abandonment. Section 5.3 evaluates the impact of having different service rates for class-1 and class-2 customers. The idea is to determine whether the conclusion of Section 4 is still valid in these different settings.

5.1 Impact of abandonment

We now add the abandonment feature to the model. We assume that the patience of each customer in the queue is exponentially distributed with rate β. This changes the MDP formulation of Section 3.1. For Ω_e with abandonment, the total event rate, $\lambda + s\mu + x\beta$, is an unbounded function of the system state. Therefore, uniformization does not apply for the original model. To overcome this difficulty, we truncated the system with parameter N, such that the maximal event rate, $\lambda + s\mu + N\beta$, is bounded. This parameter should be chosen as high as possible so that any further increase of N does not impact the policy obtained and the expected revenue, $E(G)$. As in Section 3.1, we assume that $\lambda + s\mu + N\beta = 1$. Equation (3) becomes

$$\begin{aligned} V_{k+1}(x) =& \lambda U_k(x) + \min(s, s+x)\mu \left(W_k(x-1) + r_1 \mathbb{1}_{x>0} \left(1 - \omega \frac{x}{s\mu} \right) \right) + x\beta \mathbb{1}_{x>0} W_k(x-1) \\ &+ (1 - \lambda - \min(s, s+x)\mu - x\beta \mathbb{1}_{x>0}) W_k(x), \end{aligned} \tag{13}$$

for $k \geq 0$ and $-s \leq x \leq N$, where the operators U_k and W_k are defined as in Section 3.1 for $x < N$. We chose $W_k(N) = V_k(N)$ and $U_k(N) = V_k(N) - L$, such that a rejection from state N is seen as an outsourcing decision.

For Ω_l with abandonment, we used the approximation model developed in a previous contribution (Legros et al., 2017). The idea is to approximate the abandonment distribution by a homogeneous Coxian distribution evolving with rate γ. The purpose of this method is to have a uniformizable MDP with decisions based on the experienced wait. From Theorem 2 in Section 4 of Legros et al. (2017), the transition probabilities from waiting phase $x > 0$ to a lower waiting phase, $x - h$, for $0 \leq h \leq x$, are given by

$$q_{x,0} = \prod_{k=1}^{x} \left(1 + \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^k\right)^{-1}, \text{ and, } q_{x,x-h} = \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^{x-h} \prod_{k=x-h}^{x} \left(1 + \frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\beta}\right)^k\right)^{-1}.$$

Therefore, for $k \geq 0$, and $x > 0$, the second line of Equation (4) is changed to

$$V_{k+1}(x) = \gamma\frac{\gamma}{\gamma+\beta}U_k(x) + \gamma\frac{\beta}{\gamma+\beta}F(W_k(x)) + s\mu\left(F(W_k(x)) + r_1\left(1 - \omega\frac{x}{\gamma}\right)\right) + (1 - \gamma - s\mu)W_k(x). \quad (14)$$

Using the value iteration technique, we find that the *long-run* optimal policy (i.e., as k tends to infinity) for outsourcing and reservation is of threshold type as defined in Section 3.1. However, the value iteration technique does not allow us to prove this result. Contrary to the case without abandonment, the monotonicity properties of the value function which define a threshold policy are not valid for each k. For Ω_e, the concavity property at $x = 0$ can be broken for some k if $\beta > \mu$. Note that the condition $\mu \geq \beta$ was also found to be a limitation for proving other second order monotonicity properties in the system parameters for the M/M/s+M queue (e.g., see Theorem 3 in Armony et al. (2009)). For Ω_l, the difficulty lies in the transition probabilities, $q_{x,x-h}$, for $0 \leq h \leq x$, and $x > 0$. Without abandonment, in the proof of Theorem 1, we used the property $q_{x+1,x+1-h} = q_{x,x-h}$, for $0 \leq h < x$, to show the propagation of the *general concavity property* in x. This equality is broken with abandonment which prevents proving the induction step.

Interestingly, the $n-$terminating approach in Section 3.2 for computing the optimal outsourcing thresholds is valid. More precisely, the results in Lemma 1, Theorem 3 and Proposition 2 can be extended to the case with abandonment. This can be seen by rewriting the relative difference, $\Delta^c(x)$ in Equation (7) and $\Delta^{c,n}(x)$ in Equation (9) with abandonment, for $x > 0$. For Policy π_e^*, we obtain,

$$\begin{aligned} E(G)^c - g(x) &= \lambda \max(\Delta^c(x+1), 0) - (s\mu + x\beta)\Delta^c(x), \text{ for } 0 < x < N, \\ E(G)^{c,n} - g(x) &= \lambda\Delta^{c,n}(x+1) - (s\mu + x\beta)\Delta^{c,n}(x), \text{ for } 0 < x \leq n \leq N, \end{aligned} \quad (15)$$

with $\Delta^{c,n}(n+1) = 0$. For Ω_l, for $x > 0$, note that $q_{x,k} = (1 - q_{x,x})q_{x-1,k}$, for $x > 1$ and $0 \leq k \leq x - 1$. Thus, we have $V(x) - F(V(x)) = (1 - q_{x,x})(V(x) - F(V(x-1)))$. Therefore, Equations (7) and (9) can be rewritten as

$$\begin{aligned} E(G)^c - g(x) &= \gamma\frac{\gamma}{\gamma+\beta}\max(\Delta^c(x+1), 0) - (1 - q_{x,x})\left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\Delta^c(x), \text{ for, } x > 0, \\ E(G)^{c,n} - g(x) &= \gamma\frac{\gamma}{\gamma+\beta}\Delta^{c,n}(x+1) - (1 - q_{x,x})\left(s\mu + \gamma\frac{\beta}{\gamma+\beta}\right)\Delta^{c,n}(x), \text{ for } 0 < x \leq n, \end{aligned} \quad (16)$$

with $\Delta^{c,n}(n+1) = 0$. Equations (15) and (16) can be used to extend the proofs of Section 3.2 to the case with abandonment. The only change in the proofs is the definition of the modified problem with threshold

level m for Theorem 3 and Proposition 2. Instead of assuming that the transition rates are identical between the original and the modified problem, we chose to have the transition rates constant in the modified problem for $x > m$ and equal to their value in the original problem at $x = m + 1$.

Consequently, Algorithm 1 can be used to obtain the optimal reservation and outsourcing thresholds by solving Equations (15) and (16) numerically. There remains to provide the performance measures for initiating the algorithm and facilitating the comparison between the two policy classes. For Policy π_e^*, a Markov chain analysis can lead to the performance measures. For Policy π_l^*, the combination of abandonment and outsourcing can be seen as a global reneging behavior, where the reneging behavior is the minimum between a deterministic threshold τ and an exponential duration with parameter β. This allows us to adjust some of the results of Zeltyn and Mandelbaum (2005) to our model. As for the case without abandonment, the details of the performance measures' computation are omitted. As in Section 3.3, we express the performance measures as functions of certain building blocks. For both policy classes, we have

$$E(T) = \lambda \frac{\frac{\binom{s-1}{c}}{a^c/c!}}{\epsilon + \lambda J}, \quad E(W) = \frac{\lambda J_H}{\epsilon + \lambda J}, \quad \text{and, } P_A = \beta \frac{\lambda J_H}{\epsilon + \lambda J}.$$

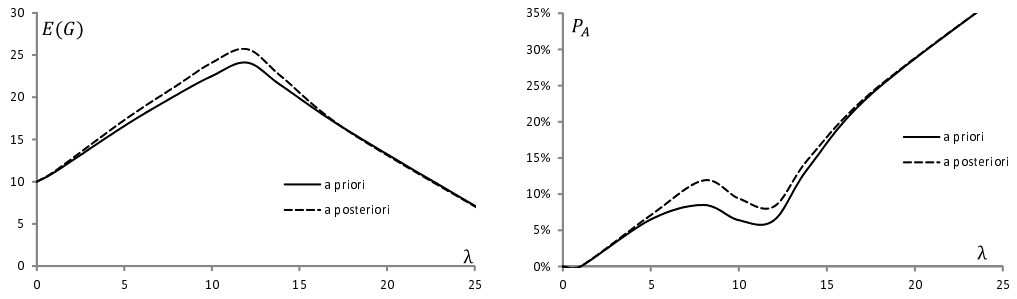
For Policy π_e^*, we have $P_{\overline{S}} = \frac{\lambda I}{\epsilon + \lambda J}$, and $E(W_S) = \frac{\lambda J_1}{\epsilon + \lambda(J-I) - \lambda\beta J_H}$. For Policy π_l^*, we have $P_{\overline{S}} = \frac{1 + (\lambda - s\mu)J - \lambda\beta J_H}{\epsilon + \lambda J}$, and $E(W_S) = \frac{s\mu J_1 - J}{\epsilon + s\mu J - 1}$. The building block ϵ is identical to that of Section 3.2. In Table 4, we specify the other building blocks with abandonment. The building block J_1 cannot be expressed explicitly. Consequently,

Table 4: Building blocks

	Policy π_e^*	Policy π_l^*
I	$\frac{\lambda^n}{\prod_{i=0}^{n}(s\mu + i\beta)}$	$-$
J	$\sum_{k=0}^{n} \frac{\lambda^k}{\prod_{i=0}^{k}(s\mu + i\beta)}$	$\frac{1}{s\mu} + \sum_{k=1}^{\infty} \frac{\lambda^k \left(1 - e^{\frac{\lambda}{\beta}(1 - e^{-\beta\tau}) - (s\mu + k\beta)\tau}\right)}{\prod_{i=0}^{k}(s\mu + i\beta)}$
J_1	$\frac{1}{s\mu + \beta} \sum_{k=0}^{n-1} \frac{(k+1)\left(\frac{\lambda s\mu}{s\mu + \beta}\right)^k}{\prod_{i=0}^{k}(s\mu + i\beta)}$	$\int_0^{\infty} x e^{\frac{\lambda}{\beta}(1 - e^{-\beta \min(x,\tau)}) - s\mu x} \mathrm{d}x$
J_H	$\sum_{k=1}^{n} \frac{k\lambda^{k-1}}{\prod_{i=0}^{k}(s\mu + i\beta)}$	$\frac{1}{\beta}\left[J - e^{-(s\mu+\beta)\tau + \frac{\lambda}{\beta}(1 - e^{-\beta\tau})}\left(\frac{1}{s\mu} + \sum_{k=0}^{\infty} \frac{\lambda^k (1 - e^{-k\beta\tau})}{\prod_{i=1}^{k+1}(s\mu + i\beta)}\right)\right]$

the expected revenue which involves J_1 must be calculated numerically and the comparison between the two policy classes can only be made numerically.

Figure 2 compares the two policy classes. This confirms the result of Section 4 which states that Policy π_l^* outperforms Policy π_e^* in terms of revenue, but is detrimental to service quality (measured here by the percentage of abandonment). Nevertheless, the difference between the two policies is reduced with highly impatient customers. As in Figure 1, we observe that the difference between the two policies increases with the arrival rate. However, this observation is valid only up to a certain arrival rate. When the arrival rate is very high compared to service capacity, the difference between the two policies decreases with the arrival rate, because the efficiency-driven regime is reached (Whitt, 2004). For this regime, the threshold c is irrelevant as agents have no opportunity to initiate a class-2 call since the queue is never empty. Under both policy classes,



(a) $E(G)$ (b) P_A

Figure 2: Comparison between the two policy classes ($s = 10$, $\mu = 1$, $r_1 = 3$, $r_2 = 1$, $\frac{C_{outs}}{\lambda \overline{P_{\overline{S}}}} = 1/2$, $\omega = 1$, $\overline{P_{\overline{S}}} = 20\%$, $\beta = 1$)

the proportion of served customers is equal to $\frac{s\mu}{\lambda}$ and the expected waiting time of served customers is $\frac{\ln\left(\frac{\lambda}{s\mu}\right)}{\beta}$. This renders the expected revenue identical under both policies.

5.2 When the wait has a positive impact

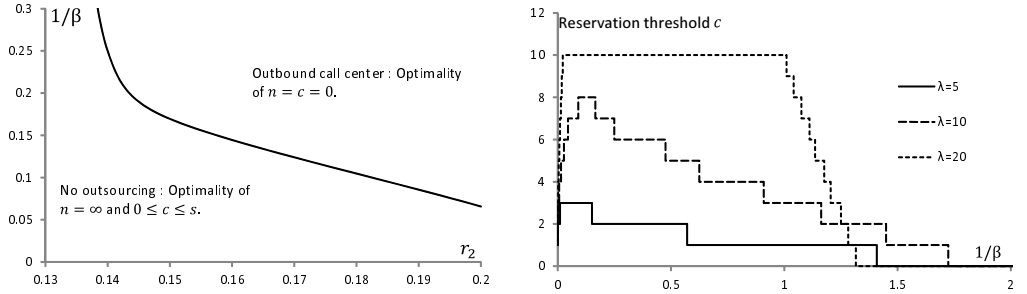
We also explored the case where waiting has a positive impact on purchase behavior. For this purpose, we assume that $\omega < 0$. In this way, the longer customers wait before being served, the more likely they are to accept the purchase offer. Without abandonment, the manager should choose an understaffing level such that $\lambda > s\mu$ and should not outsource any call. In this way, the wait and the expected revenue would be infinite. This situation is, however, unlikely to happen since callers would not wait infinitely. With abandonment, letting customers wait a long time increases the revenue per served caller but reduces the number of callers who accept to stay in the queue. Therefore, allowing customers to wait a long time is not necessarily beneficial. Using the tools developed for the case $\omega > 0$ (i.e., negative impact of the wait), we investigate how reservation and outsourcing could be implemented with $\omega < 0$.

The optimal policy with $\omega < 0$ can be obtained by recursively evaluating V_k using Equations (13) and (14). Without the constraint $P_{\overline{S}} \leq \overline{P_{\overline{S}}}$, regardless of the contract cost C_{outs}, we observe that extreme decisions should be taken for outsourcing; either all calls should be served in-house (i.e., $n = \tau = \infty$ is optimal) or all calls should be outsourced (i.e., $n = \tau = 0$, and $c = 0$). This result is proven in Proposition 4 when assuming a threshold reservation policy. This means that it is never optimal to have a non-extreme proportion of outsourced calls as was the case with $\omega > 0$. Therefore, either the outsourcer should serve all inbound callers ($P_{\overline{S}} = 100\%$) and the call center becomes a specialized outbound contact center, or the call center should not implement any outsourcing strategy ($P_{\overline{S}} = 0\%$). In both cases, the two policy classes for outsourcing are identical. When all calls are served in-house, the outbound call initiation follows a threshold policy as in the case $\omega > 0$. It is not possible to prove this result by induction on V_k however. The reason is that $V_k(x)$ is no longer decreasing and concave for $x \leq 0$, or for any combination of the system parameters.

Proposition 4. *For a given reservation threshold c, it is either optimal to have $n = \tau = 0$ or $n = \tau = \infty$.*

Figure 3 presents the optimal policy, computed with Equations (13) and (14), for different combinations of the system parameters. Figure 3(a) presents the preference zones -separated by the curve- for outsourcing

all inbound calls or serving all of them in-house as functions of the reward for initiating an outbound call, r_2, and customers' expected patience, $1/\beta$. Given that the wait has a positive effect on revenue with $\omega < 0$, the only motivation for having $n = \tau = 0$ is to provide some service capacity for outbound calls. Therefore, if $n = \tau = 0$ is optimal, it means that outbound calls are significantly more valuable than inbound ones. In order to maximize the time spent on outbound calls, it makes sense to also choose $c = 0$. As observed in Figure 3(a), the motivation for choosing this strategy increases with the reward for serving an outbound call, r_2, and with the expected patience, $1/\beta$. When it is optimal to be a specialized outbound call center, inbound callers are seen as an obstacle to achieving high expected revenue. The reason is that these callers keep some agents busy who could instead be initiating more valuable outbound calls. If the patience of inbound callers increases, then inbound callers may stay longer in the system and agents could potentially be busier with them. This strengthens the motivation to outsource all of them and explains the impact of patience.



(a) Preference zone ($\lambda = 10$)

(b) Optimal reservation threshold ($r_2 = 0.01$, $n = \tau = \infty$)

Figure 3: Optimal policy ($s = 10$, $\mu = 1$, $r_1 = 0.1$, $\omega = -0.01$, $C_{outs} = 1/10$)

Figure 3(b) presents the way the optimal reservation threshold should be chosen as a function of expected patience, $1/\beta$, for different values of the arrival rate, λ, in a situation where it is optimal to have $n = \tau = \infty$ (i.e., no outsourcing). Reservation tends to increase with the arrival rate in such a way as to provide sufficient idle agents for inbound callers. When inbound callers are very impatient (i.e., for low values of $1/\beta$), the wait plays a negligible role in the call center's revenue since callers refuse to wait. Therefore, an increase in customers' patience is seen as an increase in demand for inbound calls. One way to respond to this demand is to increase the reservation threshold, c. When callers are more patient, the wait can be used to increase the call center's revenue. The more patient customers are, the more profitable it is to let them wait. One way to increase the wait is to reduce the reservation threshold, c. This explains the non-monotonous evolution of c as a function of $1/\beta$.

5.3 Analysis with different service rates

We now investigate the effect of having different service rates with class-1 and class-2 calls. We denote the service rate of class-i calls by μ_i, for $i = 1, 2$. As in the case of equal service rates, if $1 - s\mu_1/\lambda \geq \overline{P_S}$ (unstable situation), then the optimization problem has no solution. The optimal policy with different service rates is likely to be a state-dependent threshold policy where the threshold depends on the number of class-1 and class-2 calls in the system. From a practical point of view, a state-dependent threshold policy can be difficult

to implement using call center software. Therefore, we prefer to study the simpler threshold policies considered in the sections above. This choice is partially supported by the observation of Bhulai and Koole (2003) who showed that a threshold policy is close to optimal in a similar reservation-related queueing model.

First, we evaluate the system performance under the two policy classes. Despite the fairly simple policies under consideration, explicit analysis of the Markov chain is very involved. Alternatively, we propose to use the value iteration technique to compute the performance measures. This approach is consistent with the MDP approach of Section 3. The idea under both policy classes is to recursively define a value function, denoted by V_k, on a 2-dimensional aperiodic irreducible finite state Markov chain. As k tends to infinity, the $V_{k+1} - V_k$ difference tends to the sought metric.

Policy π_e^*. A state of the system is defined by the couple (x, y) where x is the number of calls (class-1 + class-2) in the system and y is the number of class-2 customers in service, for $s - c \leq x \leq s + n$ and $0 \leq y \leq s - c$. We have

$$\begin{aligned} V_{k+1}(x,y) = {} & c_1(x-s)^+ + \frac{\lambda}{\lambda + s\max(\mu_1,\mu_2)} \left[\mathbb{1}_{x<s+n} V_k(x+1,y) + \mathbb{1}_{x=s+n}(V_k(x,y)+c_2)\right] \\ & + \frac{\min(x-y, s-y)\mu_1}{\lambda + s\max(\mu_1,\mu_2)} \left[\mathbb{1}_{x>s-c} V_k(x-1,y) + \mathbb{1}_{x=s-c} V_k(x,y+1)\right] \\ & + \frac{y\mu_2}{\lambda + s\max(\mu_1,\mu_2)} \left[c_3 + \mathbb{1}_{x>s-c} V_k(x-1,y-1) + \mathbb{1}_{x=s-c} V_k(x,y)\right] \\ & + \left(1 - \frac{\lambda + \min(x-y, s-y)\mu_1 + y\mu_2}{\lambda + s\max(\mu_1,\mu_2)}\right) V_k(x,y), \end{aligned} \tag{17}$$

with $V_0(x, y) = 0$, for $s - c \leq x \leq s + n$ and $0 \leq y \leq s - c$, and where the cost parameters c_1, c_2, and c_3 are chosen in order to derive the performance measures. With $c_1 = \frac{1}{\lambda}$ and $c_2 = c_2 = 0$, we obtain the expected waiting time $E(W)$. With $c_2 = \frac{\lambda + s\max(\mu_1,\mu_2)}{\lambda}$ and $c_1 = c_3 = 0$, we obtain the proportion of outsourced customers, $P_{\overline{S}}$. Using $E(W) = (1 - P_{\overline{S}})E(W_S)$, the expected waiting time of served customers can also be determined. Finally, with $c_3 = \lambda + s\max(\mu_1, \mu_2)$ and $c_1 = c_2 = 0$, we obtain the throughput of served class-2 customers. Therefore, the expected revenue and the service quality can be fully determined.

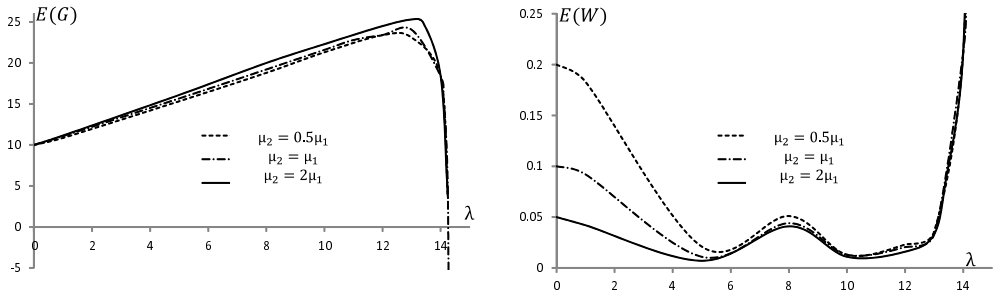
Policy π_l^*. We use the approximated model defined in Section 3. The maximal number of waiting phases is denoted by n, with $\frac{n}{\gamma} = \tau$. A state of the system is defined by the couple (x, y) where $s + x$ is the number of calls (class-1 + class-2) in the system if $x \leq 0$ or x is the waiting phase of the FIL if $x > 0$ and y is the number of class-2 customers in service, for $-c \leq x \leq n$ and $0 \leq y \leq s - c$. We extend the definition of the operator F to the set of functions f from $\mathbb{Z}^2$ to $\mathbb{R}$ by $F(f(x, y)) = \sum_{h=0}^{x} q_{x,x-h} f(x-h, y)$ for $x > 0$, and $F(f(x, y)) = f(x, y)$

for $x \leq 0$. We have

$$
\begin{aligned}
V_{k+1}(x,y) = c_1 x^+ &+ \frac{\lambda}{\lambda+\gamma+s\max(\mu_1,\mu_2)}\left[\mathbb{1}_{-c\leq x\leq 0}V_k(x+1,y)+\mathbb{1}_{0<x\leq n}V_k(x,y)\right] \\
&+ \frac{\gamma}{\lambda+\gamma+s\max(\mu_1,\mu_2)}\left[\mathbb{1}_{-c\leq x\leq 0}V_k(x,y)+\mathbb{1}_{0<x<n}V_k(x+1,y)+\mathbb{1}_{x=n}(F(V_k(x,y))+c_2)\right] \\
&+ \frac{\min(x+s-y,s-y)\mu_1}{\lambda+\gamma+s\max(\mu_1,\mu_2)}\left[\mathbb{1}_{x=-c}V_k(x,y+1)+\mathbb{1}_{-c<x\leq 0}V_k(x-1,y)+\mathbb{1}_{0<x\leq n}F(V_k(x,y))\right] \\
&+ \frac{y\mu_2}{\lambda+\gamma+s\max(\mu_1,\mu_2)}\left[c_3+\mathbb{1}_{x=-c}V_k(x,y)+\mathbb{1}_{-c<x\leq 0}V_k(x-1,y-1)+\mathbb{1}_{0<x\leq n}F(V_k(x,y-1))\right] \\
&+ \left(1-\frac{\lambda+\gamma+\min(x+s-y,s-y)\mu_1+y\mu_2}{\lambda+\gamma+s\max(\mu_1,\mu_2)}\right)V_k(x,y),
\end{aligned}
\tag{18}
$$

with $V_0(x,y)=0$, for $-c \leq x \leq n$ and $0 \leq y \leq s-c$. With $c_1 = \frac{s\mu}{\lambda\gamma}$ and $c_2 = c_3 = 0$, we obtain the expected waiting time of served customers, $E(W_S)$. With $c_2 = \frac{\lambda+\theta+s\max(\mu_1,\mu_2)}{\lambda}$ and $c_1 = c_3 = 0$, we obtain the proportion of outsourced customers, $P_{\overline{S}}$. Using $E(W) = (1-P_{\overline{S}})E(W_S) + P_{\overline{S}}\frac{n}{\gamma}$, the expected waiting time of served and outsourced customers can also be determined. Finally, with $c_3 = \lambda+\gamma+s\max(\mu_1,\mu_2)$ and $c_1 = c_2 = 0$, we obtain the throughput of served class-2 customers.

Under both policy classes, the optimal thresholds can be computed using Algorithm 1. Before comparing the two policy classes, in Figure 4 we evaluate the impact of the service rate of class-2 calls on the expected revenue and on the quality of service under Policy π_e^*. Note that similar observations could be made for Policy π_l^*. In the different examples ($\mu_2 = 0.5, 1$, and 2), we selected the product $r_2 \times \mu_2 = 1$, such that the revenue rate of an agent working on a class-2 call is maintained as constant. As expected, we can observe that the



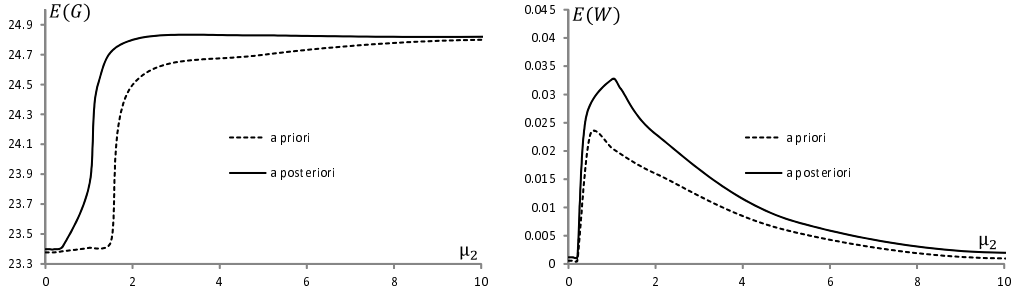
(a) $E(G)$ (b) $E(W)$

Figure 4: Impact of the service rate of class-2 customers ($s = 10$, $\mu_1 = 1$, $r_1 = 3$, $r_2 \times \mu_2 = 1$, $\omega = 1$, $\frac{C_{outs}}{\lambda \overline{P_{\overline{S}}}} = 1/2$, $\overline{P_{\overline{S}}} = 30\%$, *a priori* policy)

expected revenue increases and the expected waiting time decreases with μ_2. When the class-2 calls service is short, there are more opportunities either to initiate more class-2 calls or to serve class-1 calls with a shorter waiting time. Nevertheless, the expected revenue is not highly sensitive to μ_2 (Figure 4(a)) and the sensitivity of the expected waiting time to μ_2 tends to decrease with the arrival rate. When the arrival rate is low, the revenue is mostly driven by the class-2 calls service. The expected waiting time of served class-1 calls has only a little effect on the revenue. This explains why the revenue is also virtually insensitive to μ_2 in this case. As λ increases, the effect of the waiting time on the revenue increases but fewer class-2 calls are initiated. This reduces the effect of μ_2 on $E(W)$ and $E(W_S)$. Moreover, when the arrival rate is very high, then $c = s$ is optimal. Therefore, the call center only treats class-1 calls. This cancels out the effect of the class-2 calls

service rate.

We are also interested in the impact of μ_2 in the comparison between the two policy classes. In Figure 5,



(a) $E(G)$ (b) $E(W)$

Figure 5: Impact of the service rate of class-2 calls ($s = 10$, $\lambda = 12$, $\mu_1 = 1$, $r_1 = 3$, $r_2 \times \mu_2 = 1$, $\omega = 1$, $\frac{C_{outs}}{\lambda \overline{P_S}} = 1/2$, $\overline{P_S} = 30\%$)

we give the expected revenue and the expected waiting time as functions of μ_2. We choose $\lambda = 12$ in order to consider a zone where the service rate μ_2 sufficiently impacts the expected revenue (see Figure 4(a)). As expected from Figure 4(a), the expected revenue increases with μ_2. Contrary to what was observed in Figure 4(b), the expected waiting time can increase in μ_2. This is due to the choice of threshold parameters which incentivizes initiation of class-2 calls detrimentally to the service of class-1 customers. The difference between the two policy classes in terms of expected revenue (Figure 5(a)) and expected waiting time (Figure 5(b)) is maximal when μ_2 is close to μ_1. When μ_2 tends to zero, the expected service time of a class-2 call tends to infinity. Initiating a class-2 call would thus block an agent. It is therefore optimal not to initiate any class-2 call (i.e., $c = s$ is optimal). The waiting time of class-1 calls is therefore low, which also renders the difference between the two policy classes low. As μ_2 increases, initiating class-2 calls becomes more interesting. The choice is thus to decrease the reservation threshold, which in turns leads to higher waiting times for class-1 calls and to a higher difference between the two policy classes. When μ_2 is high, it becomes optimal not to reserve any agents for class-1 calls (i.e., $c = 0$ is optimal). Therefore, the effect of increasing μ_2 is to reduce the waiting time of class-1 calls, which in turn reduces the difference between the two outsourcing policies.

6 Conclusion

The practice of initiating or outsourcing calls in contact centers is becoming increasingly prevalent. These two levels of decisions allow managers to meet service quality and revenue targets. However, to our knowledge, no papers have addressed the control problem of outsourcing and reservation within a single framework. To this end, we considered a call center with inbound and outbound calls in a cross-selling context. One distinguishing feature of our model was that the propensity of inbound callers to buy was related to their waiting experience, based on the understanding that it may be detrimental to keep customers too long in the system. One solution to limit system congestion is to outsource part of the inbound calls. To maximize the call center's revenue, we considered the impact of call outsourcing following a wait (*a posteriori* outsourcing) as against outsourcing upon arrival (*a priori* outsourcing).

Using a Markov decision process approach, we proved the optimality of a reservation and outsourcing

threshold policy. By studying the relative value function under the optimal policy, we showed that the optimal outsourcing threshold could be computed within a finite number of iterations. Next, we derived closed-form expressions of the performance measures under both policy classes and proved the first and second order monotonicity results in the control parameters. Our main finding was that postponing an outsourcing decision improves the call center's revenue by better serving in-house customers, albeit detrimentally to outsourced ones. We believe that this result can be extended to other cases where the main focus is on served customers. We showed that the benefits of implementing an *a posteriori* policy were most significant in small congested call centers with relatively patient customers, similar expected service time for inbound and outbound calls, and when the wait has a negative effect on customers' purchase behavior.

Our analysis can be extended in several other directions to better model the operational complexity of call centers, as well as that of customer behavior. One important extension from a practical viewpoint is to allow for non-stationary arrivals or arrivals which depend on the previous customers' experiences with the call center to account for retention or acquisition phenomena. The call centers' complexity may include multiple pools of agents, different channels (chats, emails), as well as more complex service requirements. Also, from a practical perspective, separate pools of agents often handle inbound calls or initiate outbound calls. This paper provides a deeper understanding of the benefits of cross-training agents to perform each of these tasks. At methodological level, we believe that the idea of constructing a time-based decision-making policy is general enough to apply to other operations management issues. For instance, in situations where the holding cost function is non-linear, it could be interesting to develop such policies as opposed to quantity-based ones.

References

Adusumilli, K. and Hasenbein, J. (2010). Dynamic admission and service rate control of a queue. *Queueing Systems*, 66(2):131–154.

Akşin, Z., Ata, B., Emadi, S., and Su, C. (2016). Impact of delay announcements in call centers: An empirical approach. *Operations Research*, 65(1):242–265.

Akşin, Z., Ata, B., Emadi, S. M., and Su, C.-L. (2013). Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746.

Akşin, Z., De Véricourt, F., and Karaesmen, F. (2008). Call center outsourcing contract analysis and choice. *Management Science*, 54(2):354–368.

Aksin, Z. and Harker, P. (1999). To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers. *Journal of Service Research*, 2(1):19–33.

Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

Armony, M. and Gurvich, I. (2010). When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing & Service Operations Management*, 12(3):470–488.

Armony, M., Plambeck, E., and Seshadri, S. (2009). Sensitivity of optimal capacity to customer impatience in an unobservable M/M/s queue (why you shouldn't shout at the DMV). *Manufacturing & Service Operations Management*, 11(1):19–32.

Bassamboo, A., Harrison, J., and Zeevi, A. (2006). Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research*, 54(3):419–435.

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

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Byers, R. and So, K. (2007). Note- A mathematical model for evaluating cross-sales policies in telephone service centers. *Manufacturing & Service Operations Management*, 9(1):1–8.

Chevalier, P. and Van den Schrieck, J. (2008). Optimizing the staffing and routing of small-size hierarchical call centers. *Production and Operations Management*, 17(3):306–319.

Cui, S., Veeraraghavan, S., Wang, J., and Zhang, Y. (2018). In-queue observation and abandonment. *Available at SSRN 3290868*.

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.

Emadi, S. and Swaminathan, J. (2017). Impact of callers' history on abandonment: Model and implications. Technical report, Working paper.

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

Gans, N. and Zhou, Y. (2007). Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management*, 9(1):33–50.

Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97.

Güneş, E., Akşin, O., Örmeci, E., and Özden, S. (2010). Modeling customer reactions to sales attempts: If cross-selling backfires. *Journal of Service Research*, 13(2):168–183.

Güneş, E. and Akşin, Z. (2004). Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing & Service Operations Management*, 6(4):338–357.

Gurvich, I., Armony, M., and Maglaras, C. (2009). Cross-selling in a call center with a heterogeneous customer population. *Operations research*, 57(2):299–313.

Gurvich, I. and Perry, O. (2012). Overflow networks: Approximations and implications to call center outsourcing. *Operations research*, 60(4):996–1009.

Harrison, J. and Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1):20–36.

Hasija, S., Pinker, E., and Shumsky, R. (2008). Call center outsourcing contracts under information asymmetry. *Management Science*, 54(4):793–807.

Jennings, O., Mandelbaum, A., Massey, W., and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394.

Jouini, O., Dallery, Y., and Nait-Abdallah, R. (2008). Analysis of the impact of team-based organizations in call center management. *Management Science*, 54(2):400–414.

Koçağa, Y. and Ward, A. (2010). Admission control for a multi-server queue with abandonment. *Queueing Systems*, 65(3):275–323.

Koçağa, Y. L., Armony, M., and Ward, A. R. (2015). Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, 24(7):1101–1117.

Koole, G. (2013). *Call center optimization*. Lulu. com.

Koole, G., Nielsen, B. F., and Nielsen, T. B. (2012). First in line waiting times as a tool for analysing queueing systems. *Operations research*, 60(5):1258–1266.

Legros, B., Jouini, O., and Koole, G. (2017). A uniformization approach for the dynamic control of queueing systems with abandonments. *Operations Research*, 66(1):200–209.

Lerzan, Ö. and Akşin, Z. (2010). Revenue management through dynamic cross selling in call centers. *Production and Operations Management*, 19(6):742–756.

Lu, Y., Musalem, A., Olivares, M., and Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science*, 59(8):1743–1763.

Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: the Palm/Erlang-a queue, with applications to call centers. In *Advances in services innovations*, pages 17–45. Springer.

Mandelbaum, A. and Zeltyn, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205.

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

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

Ren, Z. and Zhou, Y. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383.

Robbins, T., Medeiros, D., and Harrison, T. (2010). Does the Erlang C model fit in real call centers? In *Proceedings of the 2010 Winter Simulation Conference*, pages 2853–2864. IEEE.

Schrieck, J., Akşin, Z., and Chevalier, P. (2014). Peakedness-based staffing for call center outsourcing. *Production and Operations Management*, 23(3):504–524.

Ulku, S., Hydock, C., and Cui, S. (2017). Making the wait worthwhile: Experiments on the effect of queueing on consumption.

Veeraraghavan, S., Xiao, L., and Zhang, H. (2018). Impatience and learning in queues. *Under Review*.

Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461.

Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102.

Yao, J. (2016). *Asymptotic Analysis of Service Systems with Congestion-Sensitive Customers*. PhD thesis, Columbia University.

Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue. *Queueing Systems*, 51(3-4):361–402.