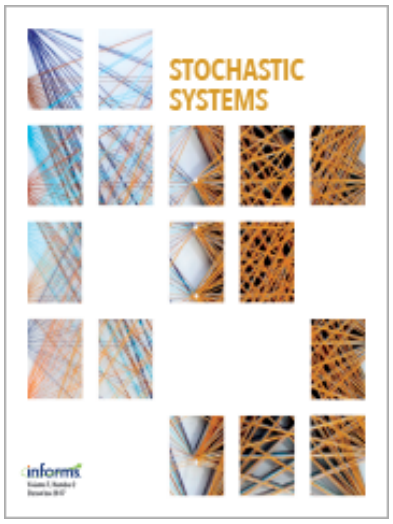


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Practice-Oriented Overview of Call Center Workforce Planning

Ger M. Koole, Siqiao Li

To cite this article:
Ger M. Koole, Siqiao Li (2023) A Practice-Oriented Overview of Call Center Workforce Planning. Stochastic Systems 13(4):479-495. https://doi.org/10.1287/stsy.2021.0008

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 13, No. 4, December 2023, pp. 479–495
ISSN 1946-5238 (online)

A Practice-Oriented Overview of Call Center Workforce Planning

Ger M. Koole,[a] Siqiao Li[a,b,*]

[a] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam 1081 HV, Netherlands; [b] CCmath B.V., 1181 GV Amstelveen, Netherlands

*Corresponding author

Contact: ger.koole@vu.nl, https://orcid.org/0000-0003-1776-8369 (GMK); siqiao@ccmath.com, https://orcid.org/0000-0002-4576-8342 (SL)

Received: May 10, 2021
Revised: October 12, 2022
Accepted: May 23, 2023
Published Online in Articles in Advance: July 4, 2023

https://doi.org/10.1287/stsy.2021.0008

Abstract. We give an overview of the practice and science of call center workforce planning, in which we evaluate the commonly used methods by their quality and the theory by its applicability. As such, this paper is useful for developers and consultants interested in the background and advanced methodology of workforce management and for researchers interested in practically relevant science.

Keywords: call centers • queueing models • workforce planning • forecasting

1. Introduction

Call centers are an important part of our global economy. Numbers vary, but the global market is estimated at around US$500 billion, whereas in the United States alone, more than a million people work in call centers. A simple extrapolation leads to more than 50 million call center agents worldwide. The industry is still expected to grow strongly to more than 700 billion by 2030 (Global Industry Analysts, Inc. 2023), although it should be noted that long-term forecasts are known to be highly unreliable and can be heavily influenced by new and existing technology, such as ChatGPT.

Call centers are also a fascinating area for stochastic modeling. In manufacturing, most production is being done before the demand occurs, and then the product is lying on a shelf in a shop or a distribution center waiting for customer demand. In (nonurgent) healthcare, production is smoothed in time to meet capacity: a patient makes an appointment with a healthcare provider at a moment that suits, above all, the provider. In aviation and hospitality, demand is pushed by financial incentives toward low-demand time slots. Inbound call centers have in common with emergency healthcare that demand has to be met almost instantaneously by supply. And, whereas a hospital has at least 15 minutes to prepare for the arrival of a trauma patient, a call center often has to answer a call within 20 seconds. And it can be life-saving as in the case of an emergency call center.

To be able to deliver this type of service, planners have to deal with fluctuations, unforeseen (such as the variability of the Poisson process or illness of employees, often called agents) and foreseen (such as intraday and intraweek seasonality in demand). Call centers cannot react instantaneously to all fluctuations and, therefore, have to schedule overcapacity as well. Designing and planning the call center in such a way that the optimal combination of flexibility and overcapacity is scheduled is the essence of call center workforce planning. This overview focuses on the practice of workforce planning or workforce management (WFM) as it is commonly called in practice.

To be able to schedule effectively, different decisions at different time scales are required. As a framework for this overview, we use these different decision processes. The four central planning processes are

- The long-term budget planning process, which is input for the corporate management that sets the financial boundaries.
- The tactical capacity planning process, in which decisions concerning the agent pool are made—mostly the hiring of new agents and the training of new skills.
- The short-term operational planning process, which starts with deciding what volume goes to the external partners and what is handled internally and then consists of agent scheduling (which needs to be communicated a few weeks in advance).

- Finally, intraday management, adaptations to the schedule and task assignments, which is done at the day itself.

The processes are shown in Figure 1. Here, "X" refers to the day of execution; thus, for example, "X + 1Q" means one quarter before the day of execution. BPO stands for business process outsourcer, a partner specialized in delivering call center services. To allow BPOs to schedule the required amount of agents, forecasts and/or required staffing levels are communicated at multiple moments in time, often starting a few months in advance.

Depending on the particular call center, the situation might be slightly different, and smaller call centers with stable volumes might not execute the long-term steps explicitly. Note also that this scheme is biased toward the European situation with its strict labor laws forcing call centers to schedule carefully and publish schedules well in advance.

As can be seen from the figure, every process starts with forecasting. An exception is intraday management: the forecast is rarely updated after the agent schedule is made although that would likely result in increased accuracy. The forecast is input to a planning step, the content and output of which depend on the process. Every subsequent process depends on the output of the previous process: you cannot hire the required agents if the budget is not reserved in the budget planning process, etc.

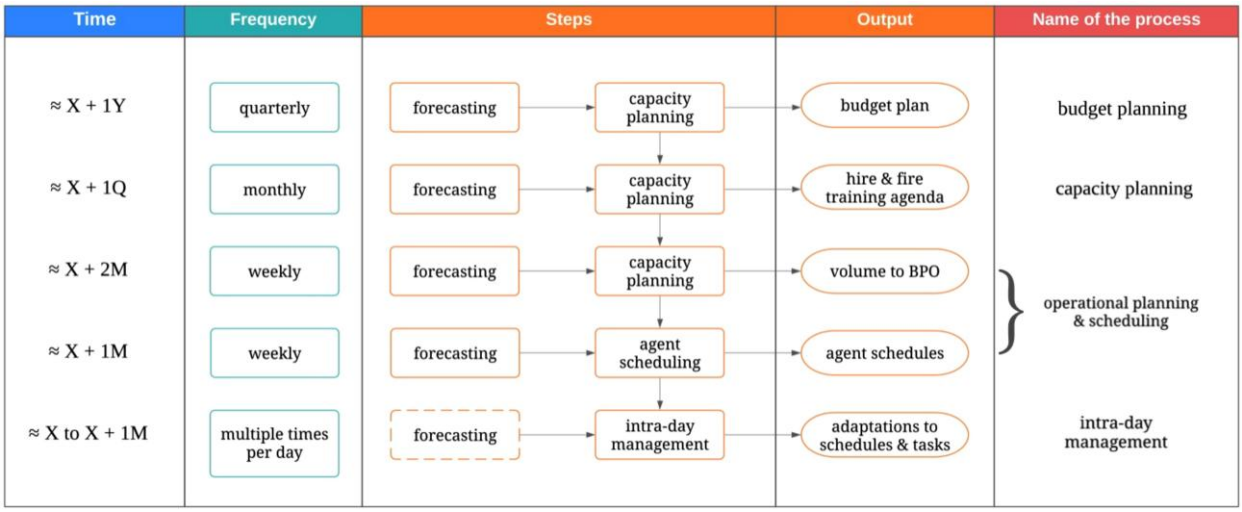
Next to the processes from Figure 1, call centers have two more less explicit processes: a long-term, ad hoc process that is about setting up and improving the overall design of the call center: the shift structure, the way forecasts are made, opening times, channels that are offered, etc. We call this the design of the call center. Finally, there is the real-time routing, the assignment of customer contacts to agents, which has a big impact on the performance. This is automated and part of the telephony/omnichannel switch. Although this could also benefit from updated forecasts and other real-time information, this is rarely done. Routing and intraday management can be seen as complementary; one is manual, whereas the other is automated. We discuss them together.

Note also that we left out the connections with other departments. Forecasting, for example, takes input from marketing and sales to obtain the dates of marketing campaigns and sales forecasts, and the budget plan is used in negotiations with higher management to set the final budget. Furthermore, the processes are not always as linear and unidirectional as they seem: there might be interaction between forecasting, scheduling, and marketing about the feasibility and the schedule of marketing campaigns; capacity planning might lead to adaptations to the budget, etc.

In the next sections, we discuss one by one the different WFM processes. The processes are connected in two ways: up (short to long term) and down (long to short term). Longer term processes determine the amount and types of resources that can be used by shorter term processes. That is, the available budget constrains the capacity that can be hired, the available agent pool constrains the number of agents that can be scheduled, and intramanagement is bound by the number of agents that is scheduled. This is reflected by the downward arrows in Figure 1.

Vice versa, the short-term processes and also the design determine how the resources are used, which model should be used to determine the required capacity, and with that also how efficiently resources are used. Thus, the scheduling method depends on the design and the routing. In its turn, the capacity and also the budget depend on

Figure 1. The WFM Processes



all shorter term processes although often simple methods based on ratios are used: simply said, if the forecast is 20% higher than last year, then the call center requires 20% more people and budget. Therefore, we start with discussing design and then work our way up Figure 1 from real-time management to scheduling to capacity planning to budget planning. Note that this is also the way in which WFM gets more mature as a call center grows: the starting moment for a call center can be well-defined as the moment that a number of employees gets a joint number, thus, when routing is put into place. When the call center grows, first, agent scheduling becomes an explicit process and, then, later on, capacity and budget planning.

Papers and other relevant sources of information are cited when appropriate. We do not try to be complete with respect to the literature; we focus on what we consider to be most relevant to practice. Papers inspired by call centers but of little use to its operations are left out.

We also discuss, at the end of each section, research opportunities. Despite the huge call center literature, there are still quite a number of open questions. Often they involve the combination of multiple steps in the planning process, for example, the amount of flexibility needed in the planning process to deal with workload and workforce forecasting errors. There are also opportunities to use modern techniques from artificial intelligence (AI). For example, replacing the Erlang formulas by black-box machine learning (ML) methods is currently a hot topic. Expanding these methods to multiskilled situations would be very interesting and practically relevant.

We end this introduction by citing some general call center WFM references. The following are academic overviews: Gans et al. (2003), Avramidis and L'Ecuyer (2005), and Akşin et al. (2007). Practitioner-oriented books are Cleveland and Mayben (1997) and Koole (2013).

2. Design

With the design of a call center, we refer to decisions related to opening hours, communication channels offered, if and how external partners are used, how multiskilled or multichannel agents are employed, etc. A good design makes it possible to find the right trade-off between the interests of all stakeholders: customers, employees, and management. For example, prolonged opening hours lead to better customer service, lower agent satisfaction, and higher costs. Good customer service is usually defined by service level agreements (SLAs), which are design decisions. Employee satisfaction is represented, for example, in the types of possible shifts and fairness among agents. The interests of management are represented by the financial side.

An important design decision is which communication channels to offer to the customer. Although inbound calls are often the most prominent channel, we also see companies that only offer chat as a means for communication in a hope to lower communication costs. Because chat agents can handle multiple chats in parallel, these call centers expect that the overall time spent on a contact is lower. To give credit to the different channels, the term "contact center" has been introduced. Few people, however, use it; thus, a call center is, most of the time, a contact center mixing contact from different channels. A notable exception are call centers dedicated to outbound marketing campaigns. Through predictive dialing, they deal with fluctuations in the fraction of calls that are answered and the speed at which this is done. Quite a number of patents for algorithms can be found on Google Scholar on this subject.

The service level (SL) is just one aspect of the quality of a call. In fact, some people claim that "the best service is no service" (Price and Jaffe 2008). Indeed, we see a tendency for offering automated customer service by using, for example, AI in chatbots and call avoidance by, for example, improved product design and websites. There is evidence that making calling unnecessary is the best customer service, and if people call, avoid that they have to make another call later on (Dixon et al. 2013). Avoiding calls is also cheaper, and as most call centers are seen as cost centers, there is a strong incentive to reduce costs. However, there is no evidence that the call center market is shrinking, in fact, on the contrary (Mazareanu 2019). A possible explanation is the popularity of shared service centers, which operate effectively as call centers (e.g., the human resources department at our university). As such, we see a tendency across industries from decentralized service to centralized service (operated as a call center, potentially offshored to a country with lower wages) to self-service.

To evaluate the impact of design decisions, techniques similar to those used for capacity management and budget planning can be used. We refer to the relevant sections. We continue this section with some general guiding principles on how the WFM processes should be designed.

The first is about when to make decisions. Decisions that limit flexibility should be made as late as possible. That way we can better deal with fluctuations because, for all types of fluctuations, it holds that, over time, more information becomes available; that is, the variability of the unknown variable decreases over time. For example, take a multiskilled call center. During the scheduling phase, skills can be assigned to agents, blocking them for other skills unless traffic management changes the schedule. Letting skill-based routing (SBR) do the assignment is much more efficient: even a fixed assignment at the last moment is better because the latest forecast can be used and availability

is fully known; you know, for example, who is ill. A reassignment could be part of intraday management, but why then schedule in the first place?

An often-heard objection against SBR is that agents have to change skill (e.g., move from one language to another) frequently, which can be annoying. Similar objections hold against blending multiple channels, especially when email handling is interrupted for inbound calls. Good routing, however, can avoid that: in certain systems, you can limit the number of times that an email might be interrupted by a call, and one can think of similar solutions for SBR.

When the decision is related to something that influences employee satisfaction, then making decisions later might be more efficient but, at the same time, decrease employee satisfaction, which negatively impacts the performance of the call center. However, flexibility is not always required at the maximum level; asking only a fraction of the employees to be flexible might give you the majority of the advantages of flexibility, which is the next guiding principle.

A little flexibility goes a long way. Wallace and Whitt (2005) observe this for the number of multiskilled agents in an SBR setting, but this holds in general: a few agents with part-time shifts, a few back-office agents who can help in the front office (the call center), etc., can help to obtain the biggest part of the advantages of flexibility. Another way to state it is that flexibility shows decreasing returns. Note that this observation can be mathematically proven in the case of costs linear in the level of flexibility and a convex feasible region.

From this, it also follows that you can better have a bit of multiple types of flexibility than a large amount of one type of flexibility. However, using all these forms of flexibility together in the smartest way possible is a challenging task that requires appropriate tooling. Nobody can immediately grasp the consequences on all skills of one agent less, especially on the ones the agent does not have. That brings us to the final guiding principle.

Automated decision making is preferred over manual. Few decisions are fully manual or automated; for most decisions, there is some tool (which can be a spreadsheet) that supports the decision. The better the tooling, the less human interference is required. We argue that a higher degree of automation is usually better: advanced knowledge in the form of algorithms can be implemented in software, knowledge that most planners never obtain. An important constraint is that the outcomes should be transparent for the planner to be able to explain the outcomes and interact with them. As an example, take an agent who wants the afternoon off. Usually an intraday manager looks at the current SL and makes a decision on the basis of that. It would be much better to have an SL prediction for the rest of the day on which to base the decision, but that requires advanced tooling. From there, it is a small step to an automated system that compares the predicted SL with the SLA and makes a decision on that basis. Clearly, this makes the call center more efficient and also makes the WFM team smaller, leading to additional cost savings, which, by itself, are usually higher than the costs of the software.

There are different examples in the literature of designs that are well-built, also from the point of view of WFM. We cite a few of them. Jouini et al. (2008) study a multiskilled situation with different teams in which every team has its own customer base. To obtain flexibility and, with that, economies of scale, unidentified callers are routed to the least occupied team. This combines motivational incentives, such as being able to compare teams on quality of service (QoS) and upsell, with WFM aspects. Legros et al. (2015a) propose a new SBR architecture for the situation that every agent has two skills. Saltzman and Mehrotra (2001) study through simulation the introduction of a new paid service, including its impact on other services. The efficiency of the proposed architecture is compared with chaining using simulation. Akşin et al. (2008) discuss the optimal capacity levels under two different types of outsourcing contracts: volume- and capacity-based. They observe that no contract type is universally preferred, and both the operating environments and cost–revenue structures matter. Gans and Zhou (2007), Hasija et al. (2008), and Gurvich and Perry (2012) study overflow rules in the context of outsourcing. Some call centers offer a call-back option to smooth the arrival traffic, whereby customers may register a request when all agents are busy. Later, the system calls them back within a prespecified time slot. In this way, waiting inbound turns to the outbound task at scheduled moments. The analysis of such systems is conducted in Armony and Maglaras (2004a, b), Hathaway et al. (2020), Legros et al. (2016, 2017).

As many design decisions have consequences for the required workforce, a WFM specialist should be consulted when making these decisions. Jack et al. (2006) emphasize the importance of the workforce to the call center. Evaluating these decisions is not an easy task: to understand the consequences for all stakeholders, capacity planning should be conducted for different scenarios for, for example, different opening times and channels. See Section 5 for more details and scientific challenges. In practice WFM consultants are rarely consulted proactively when making design decisions; they are usually consulted afterward, when targets are not met and management has the idea that better WFM might be the solution.

3. Routing and Intraday Management

Routing in call centers is usually static: the rules are entered once in the telephony switch or automatic call distributor (ACD), and they do not depend on current service or staffing levels. Typically, routing is arranged through

priorities of agents for certain types of calls, which can be different per agent: agents usually have primary and secondary skills. SLAs can be different for all skills and channels, and certain SLAs are more important to be met than others. When multiple agents with the same priority can handle a call, then usually the one with the longest idle time since the last call is selected. (Note that this rule opens the possibility for the agent to trick the system: by going on a one-second break, the agent has again the shortest idle time.) When an agent becomes idle, the agent is often assigned to the longest waiting call among the highest priority calls. Nowadays, more sophisticated routing rules are supported by ACDs, such as Genesys. Examples are threshold policies based on the queue size or customer waiting time. Customer satisfaction can also be considered by assigning calls to agents who have the best resolution rates. Although this gives many possibilities for routing and many parameters to be set, there is no guarantee whatsoever that the best possible performance is achieved. For this reason, intraday managers often change the priorities of agents during the day. Unfortunately, they are not supported by software, and they cannot oversee all implications of their actions, which are, therefore, often highly suboptimal. Systems, such as Avaya Business Advocate (Avaya 2011) try to improve such situations. But experiences are mixed because of a lack of understanding and control by the user. Ideally, instead of letting intraday managers make last-minute adjustments to the system, the routing rules in the ACD should adapt to the current situation automatically with, as a goal, meeting the SLAs in an efficient way and being fair to the agents. Fairness in this context means that agents have comparable workloads.

Many routing algorithms are proposed in the literature but mainly for heavy-traffic regimes with fixed staffing, such as Armony and Ward (2010), Atar (2005), Atar et al. (2010), Mandelbaum and Stolyar (2004), Milner and Olsen (2008), and Ward and Armony (2013). Few studies on SBR exist that take various service levels and also fairness between agents into account. Notable exceptions are Chan et al. (2014a) and Li and Koole (2020), both using simulation. Chan et al. (2014a) consider a policy that depends, through weights, on the service and occupancy levels. The weights that give optimal stationary performance are obtained requiring full knowledge of arrival rates and staffing levels. Li and Koole (2020) is also based on weights but introduces a heuristic to obtain the best performance by the end of a day without explicitly using the system's parameters but the service level up to that moment, which is the usual performance measure in practice. Worth mentioning is also Mehrotra et al. (2012), which also takes agent proficiencies into account.

Routing between channels is called blending. Most of the studies consider blending inbound and outbound calls. Bhulai and Koole (2003) and Gans and Zhou (2003) both show that a non–work-conserving policy is optimal: some agents should be kept free for inbound calls even though outbound calls are waiting to be handled. Otherwise, the SL on inbound is too low. This greatly improves the efficiency compared with separate agent groups, and it is robust to changes in parameters, such as the arrival rate. Other threshold policies can be found in Deslauriers et al. (2007), Legros (2017), and Pang and Perry (2015), to name a few. Legros et al. (2015b) develops a threshold policy that adaptively adjusts the number of agents reserved for inbound calls to achieve the SLA of inbound calls as well as maximize the throughput of emails.

According to our knowledge, no papers yet deal with the blending of synchronous channels, such as inbound and chat. Whereas a few papers deal with the routing of chats (Tezcan and Zhang 2014, Legros and Jouini 2019), they both consider a single chat type and identical agents. Tezcan and Zhang (2014) give a routing rule that minimizes the abandonment rate and the staffing level in the long run. Legros and Jouini (2019) consider that customers can also abandon during the service because of long handling times and propose a routing policy that allows agents not to work up to the maximum number of chats even when the queue is not empty. Further relevant references include Cui and Tezcan (2016) and Luo and Zhang (2013).

One may notice that the literature mentioned so far focuses merely on call center efficiency. Their targets are set to minimize the speed of response or abandonments, maximize SL, and so on. Quality metrics, such as call resolution, customer satisfaction, and agent preference, are barely taken into account. One of the reasons is that the relevant data cannot be easily retrieved from the call center system, requiring additional processing steps. Some exceptions are Ghareeb et al. (2016) and Zhan and Ward (2014).

Intraday management is changes made to the deployment of agents during the day of execution (or just before). These changes can be to the activities they do. Sometimes this is motivated by the SL: agent priorities can be changed or, for example, meetings can be canceled to improve the SL or even scheduled at the last moment when many agents are idle. The changes in activity can also have other motivations, such as the urgent need to schedule a meeting. At all times, the consequences to the SL should be taken into account.

Next to changes in activity, intraday management deals with changes in working hours. This starts as soon as the schedule is published by the planners when, for example, agents request schedule changes for personal reasons or when the forecast has changed significantly and more or fewer agents are needed. This continues throughout the day of execution; many call centers have a flexible workforce layer through which they can upscale or downscale on short notice even during the day itself. The management of this is often not based on SL predictions, and only a few papers address this type of issue. An exception is Roubos et al. (2017) in which the staffing levels are adapted

during the day in an optimal way as to obtain the required SL by the end of the day. Reforecasting using the most recent actuals is crucial when evaluating schedule changes. See, for example, Shen and Huang (2008) and references in Ibrahim et al. (2016b).

As for research opportunities, the blending of synchronous channels, such as inbound and chat is little studied. There is definitely room for more elaborate studies on adaptive routing. Finally, we see a move toward chat with many different skills: many companies allow you to start a chat or call from the website, and then every web page becomes a separate skill. How should we route? Can the ACD learn at which skills an agent is good? How do we implement the exploration–exploitation trade-off for such a system? Robust scheduling—schedules that allow for (the right amount of) changes—might also be an interesting area for research.

4. Operational Planning and Scheduling

In the next three sections, we consider the operational planning process: first forecasting, then staffing, and finally scheduling.

4.1. Operational Forecasting

Operational call center forecasting is concerned with the prediction of call volumes at the interval level, usually per quarter of an hour, for every queue or skill separately, a few weeks in advance. The process is separated into first making forecasts at the daily level and then making intraday patterns to distribute the daily forecasts over the day.

The day-level forecast should take into account all factors that influence call volume: long-term trends; intrayear, intraweek, and intraday seasonality; and events such as holidays, marketing actions, and IT problems of products and call center systems. An important task of a forecaster is to explain what the forecaster predicts; thus, it is important that the forecasting method is transparent such that the forecaster can say something such as, "Next week on Monday, we have 2,000 calls more than last week. There is a marketing campaign with an expected impact of 3,500, but the base level is 1,500 lower because of the holidays." Managers do not allow decisions to be made on forecasts purely based on black-box forecasting. They want the reasons behind a prediction. Note that, in contrast with the linear processes described in Figure 1, there might be interactions between forecasting and planning as well. For instance, events such as marketing campaigns might (and should) be planned on the basis of agent availability.

Few call centers use advanced forecasting methods; most forecasters have no scientific background and implement their own method in Excel. Next to limited functionality, this has all the disadvantages of spreadsheets: it is error-prone and hard to maintain and transfer (Powell et al. 2009). Currently, we slowly see that data scientists also get involved in forecasting, moving away from spreadsheets, to better tooling such as R or Python.

A typical method, easily implemented in a spreadsheet, is a simple decomposition approach that adds the increase over a year to last year's volume. Written in a formula with h as the historical volumes, $\hat{h}$ the forecast, and w and y time periods of a week and a year,

$$\hat{h}_t = h_{t-y}\frac{h_{t-w}}{h_{t-y-w}}.$$

This forecast is adapted using estimations of the impact of events on t, $t-w$, $t-y$, and $t-y-w$. It can be made more sophisticated by separately predicting weekly volumes and intraweek profiles and by estimating the yearly increase by averaging over multiple weeks. Some scheduling tools offer forecasting functionality but rarely more advanced than this. Very few call centers employ advanced forecasting methods.

In the scientific literature, many methods are proposed and applied to call center data. Taylor (2008), Jalal et al. (2016), Antipov and Meade (2002), Weinberg et al. (2007), Ibrahim and L'Ecuyer (2013), and Huang et al. (2019) are some of them. An elaborate overview is given in Ibrahim et al. (2016b). All these models, using different (combinations of) algorithms from statistics and AI, are successful in the context in which they are described, but there is no consensus on which method is preferable in the most common situations. Although various lead times (i.e., the difference in time between the moment of forecasting and the moment for which the forecast is made) are considered, many papers focus on short-term forecasts such as daily or intraday forecasts, without considering intrayear seasonality, which cannot be ignored in practice. Also, most methods lack certain features that drive call center volume. In particular, events are rarely included in the models. Some exceptions are Aldor-Noiman et al. (2009), Antipov and Meade (2002), and Soyer and Tarimcilar (2008), which incorporate the effect of events (e.g., marketing strategy and special calendar effects) as exogenous variables in their mixed Poisson arrival count models, but they do not address all factors mentioned before.

A good, robust method that deals with all aspects that drive call volume is described in Hyndman (2013), using smoothing methods as in Hyndman and Athanasopoulos (2018) and a separate regression for events using dummy

variables. Also a linear regression model with a polynomial for the trend and dummy variables for the seasonal components and the different types of events works quite well (Koole 2013). Because call center actuals are multiplicative in their components (Ding and Koole 2022) it is advisable to use Poisson regression, that is, to execute the regression on the logs of the actuals.

Decomposition methods, by which we mean methods that determine the factors that influence volume one by one, only work in a multipass setting because of the dependencies of the underlying variables. For example, the occurrence of outliers can only be determined if you know the seasonalities. But the seasonalities can be better estimated if the events and outliers are filtered out. Whereas some form of decomposition is commonly used by forecasters in practice, multipass methods are rarely used nor studied in the literature.

In the urge to explain the forecast, forecasters, and especially their managers, like to include time series, such as sales, in the forecast. Forecasts made this way are called "ratio forecasts" because of a (known) fraction of new customer's calls. However, again, a forecast of the external variable is needed, whereas the trend most of the time shows considerable collinearity with the external variable, such as the sales. Furthermore, the ratio might change. Therefore, it is questionable whether including, for example, a sales forecast improves the forecast. Testing it is the only way to find out, and often, it is indeed not the case.

Forecasting errors have to be measured using some criterion. It is an important task of the forecaster to report and explain forecasting errors to management. Therefore, the error measure should be easy to interpret. The weighted absolute percentage error (WAPE) is a good candidate because it is less prone to outliers in small volumes than the mean absolute percentage error and the WAPE is linear in the intraday management costs (Ding and Koole 2022).

Call center arrivals can well be modeled as coming from an inhomogeneous Poisson process (Kim and Whitt 2014); thus, we forecast the arrival rate. This gives a minimal error, which is, in terms of the absolute percentage error (APE), equal to $\mathbb{E}|N_\lambda - \lambda|/\lambda$ with λ the forecast and $N_\lambda \approx \text{Poisson}(\lambda)$. A formula is given in Crow (1958); the (very good) normal approximation is $\sqrt{2/(\lambda\pi)}$. Taking a weighted average of the minimal APE for each interval gives the minimal WAPE for any time period consisting of multiple intervals that is to be forecasted. It is interesting to note that managers sometimes require a WAPE that is lower than this minimal WAPE, which is impossible.

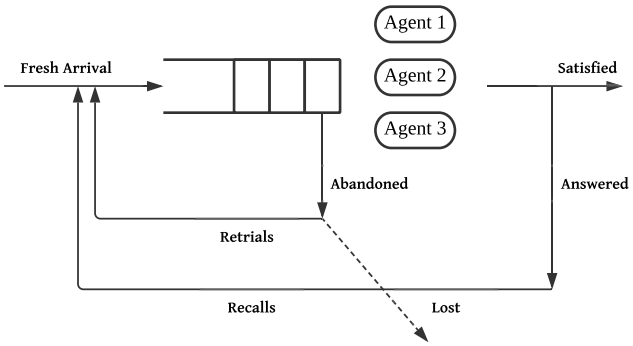
It is observed that call center data are overdispersed with respect to the Poisson distribution (see Jongbloed and Koole 2001, Avramidis et al. 2004). However, this simply means that there is a considerable error, which might well be largely explained by adding more features, such as events. Evidently, the risk of overfitting is present, but a good forecasting model taking all relevant features into account can reduce the overdispersion enormously. These features include, next to events such as marketing actions, special days, and IT issues, also the weather and other time series, such as sales forecasts.

The weather, especially the derivative of the temperature, has an impact on call volume: the first day of nice weather sees a decrease in calls in countries such as the Netherlands. To forecast call volume using the weather, you need a good weather forecast. Therefore, including the weather only works for short-term forecasting. A simple implementation is to include the first day with nice weather as a recurring event of which the impact can be determined from previous days with nice weather. Depending on the forecasting granularity, horizon, and other characteristics, a considerable WAPE on top of the minimal WAPE might remain. Although 5% is considered the gold standard, it varies wildly in practice and is regularly much higher than 5%. It would be interesting to be able to attribute the error to the forecast components. For example, a WAPE of 20% can be split up in a Poisson noise of 5% and error in the intraweek seasonality of 10%, etc.

Forecasting is often done for the total number of offered calls. However, this includes retrials: callers who abandoned earlier and called again later; see Figure 2. Data analysis shows that retrials often occur shortly after the first attempt, usually within the same day (Ding et al. 2015). It is common to know the numbers of connected and abandoned calls, but the fraction of retrials usually is not unless the callers can be identified. An empirical study of retrial behavior can be found in Hathaway et al. (2017). Ding et al. (2015) propose a statistical method to determine the "fresh" volume by using the retrial percentage as a variable in the forecasting model. In practice, taking the average between the offered and handled numbers of calls often works well, corresponding to 50% retrials. Note that there are also recalls, callers who call a second time to get further advice. Recalls add to the call volume and, therefore, to the workload, but they are also a very important driver of customer dissatisfaction (Dixon et al. 2013). Reducing it, however, is outside the scope of WFM, but the consequences of a reduction are that it decreases the volume and with that the staffing needs.

After determining daily volumes, they have to be drilled down to the intraday volume. Typically, forecasters base themselves on what they consider to be similar days to the one they are about to forecast: same day of the week, not too long ago, similar events. Then, they take the average of the profiles, the normalized volumes, and multiply that with the daily forecasts. However, this method leads to considerable overfitting; the average profile

Figure 2. Retrials and Recalls



often shows quite some variability. Much better results are obtained by using the fact that you expect neighboring intervals not to vary that much and by fitting a polynomial or a smoothing spline. This proves to work quite well (Soyer and Tarimcilar 2008, Channouf and L'Ecuyer 2012, Bakker et al. 2019).

In this section, we fully focus on forecasting inbound calls, but the same methods apply to other forms of customer contact, such as email and chat. Next to that, other parameters needed for capacity planning and agent scheduling require forecasting as well. Examples are handling times and forms of shrinkage, such as sick leave. There are some differences, and usually, not the same granularity is required (sick leave is a parameter at the week level), but overall, the same methods can be used. Note that handling times also show fluctuations during a day. Aktekin (2014) and Tan and Netessine (2014) observe that the time of the day affects agent handling times. For example, they may speed up their service when the call center is busy.

Concerning research opportunities, we already mentioned that it is interesting to be able to decompose the forecasting error on top of the error caused by the Poisson distribution. For example, what is the relation between the amount of data and the accuracy of the estimates of the seasonal parameters? Another interesting topic is applying so-called "hierarchical forecasting" methods to call volumes. Instead of forecasting independently each given service type, combining similar types to lines can sometimes improve the forecast (Ibrahim and L'Ecuyer 2013). But how to detect similar service types? What are optimal combinations giving the smallest possible error?

4.2. Staffing

Agent scheduling concerns the construction of schedules such that, among other objectives, SLAs are supposed to be met to the extent possible. Commonly used SLAs are one minus the tail of the waiting time distribution (also referred to as the SL, often taken as 80% answered within 20 seconds) and the expected waiting time (the average speed of answer). With the SLA as a constraint, the minimum required staffing in every interval is determined, sometimes explicitly or implicitly in the scheduling algorithm. If it is done explicitly and then given as input to the scheduling algorithm, then it is often done by the forecasters and even called workforce forecasting. We call it (safety) staffing as it entails planning overcapacity to deal with fluctuations in workload. Staffing is probably the best studied part of WFM and the starting point of many scientists interested in WFM, explaining why many queueing scientists (used to) work on call centers.

We specify a difference between single and multichannel and single and multiskill operations. Staffing is done at the interval level, usually 15 minutes. Even though agents can often handle multiple skills and/or channels, they are often scheduled during one or more intervals to work on a single skill and/or channel. We first look at staffing for these single-skill, single-channel operations, starting with inbound. Then, we look at staffing in a blended multichannel environment and in the presence of skill(s)-based routing, in which, in real time, a contact from the optimal channel or skill is being pushed to the agent.

It is commonly assumed that arrival rates and numbers of agents are stepwise constant functions, constant during each quarter. This is motivated by the fact that arrival rates are expected to change little during each quarter, and schedule changes are only possible at the quarter. In this situation, the so-called stationary independent period by period (SIPP) approach (Green and Kolesar 1991) is an obvious choice: you assume stationarity in each interval and use a stationary queueing model. The $M|M|s$ or Erlang C model is most commonly used in practice. Allowing for customers to abandon is a relevant feature that improves the approximation. The model including abandonments is commonly written as $M|M|s+G$ with $+G$ denoting the generally distributed patience. In the case of exponential patience, we call it Erlang A. Seminal work on these models was done by Palm (1953) and Baccelli and

Hébuterne (1981). Zeltyn and Mandelbaum (2005) give simple formulas for the $M|M|s+G$ using integrals over the patience distribution. Sze (1984) adds retrials.

Compared with the Erlang C, the Erlang A requires one more parameter: the patience distribution or its expectation, depending on the exact model used. The waiting time of a customer is the minimum of the customer's patience and the time to service; thus, patience is a censored variable. The famous Kaplan–Meier method (Kaplan and Meier 1958) can be applied, leading to results such as in Brown et al. (2005) who do a thorough analysis of call center data. Note that, in practice, patience is usually underestimated because practitioners often only look at the abandoned calls. However, taking an expected patience of 5 or 10 minutes is already much better than applying Erlang C. You can also use the patience as a tuning parameter, but then, you need data on the achieved service levels.

A handful of papers focuses on the patience distribution and how it affects call center performance and staffing decisions. For example, Mandelbaum and Zeltyn (2004) study the impact of various patience distributions on $M|M|s+G$ queues. They observe approximate linearity between the probability to abandon and the average waiting time when there is a low-to-moderate abandonment rate. Roubos and Jouini (2013) empirically show that the hyperexponential distribution is an accurate representation of the patience distribution. Aktekin and Soyer (2014) conduct Bayesian analysis built upon different families of distributions. Ye et al. (2020) estimate the hazard function of customer patience time with a nonparametric approach. It is worth mentioning that Whitt (2005) shows that the patience distribution has a bigger impact than the service time distribution for the same expectations.

Erlang A gives also the possibility to include the abandonment rate in the SLA separately as the fraction of abandonments needs to stay below, for example, 5% or implicitly in the SL (Jouini et al. 2013). It is common in science to use the virtual waiting time: the time an arbitrary customer would have to wait if the customer's patience were ∞. However, this measure is not measured in a call center; thus, its performance cannot be verified. In practice, other definitions are used, for example, the fraction of all calls being answered within the time to answer. For definitions and ways to compute the SL for different definitions, see Jouini et al. (2013).

Hardly any of the specialized scheduling tools use Erlang A. When forecasters do staffing, they usually employ some spreadsheet add-in. The widely used Erlang97 Excel add-in (Bromley 2001) also has the option to compute abandonments. It is based, however, on a waiting-time quantile of the Erlang C, thereby making two errors: it does not model the fact that Erlang A generally has a better SL than Erlang C because some customers leave the queue, and it assumes the patience is the same for all customers (van Eeden et al. 2013).

Gans et al. (2010) and Ibrahim et al. (2016a) report considerable agent heterogeneity: the average service time (called average handling time (AHT)) differs significantly. Taking the overall average as input for the Erlang model is a simple solution, but the error can be big depending on the agents to be scheduled. Taking the (weighted) average of scheduled agents is more accurate, but this requires the staffing and scheduling step to be done together. Note that a group of new agents usually has a big impact on the AHT because of their longer handling times.

It is interesting to note that delay announcements, which provide waiting time estimates, influence customer patience. Psychologically, "uncertain waits feel longer than known finite waits" (Maister 1985). On the other hand, the delay announcement can also induce some customers to balk or abandon earlier, leading to peaks in abandonments. This, in turn, influences the waiting times (Feigin 2006; Armony et al. 2009; Ibrahim and Whitt 2009, 2011; Jouini et al. 2011). Recently, Akşin et al. (2017) applied a series of Cox regressions to call center data of a bank with delay announcements every 60 seconds. The results reveal that the details of the announcements (e.g., how the waiting time information is offered and if it represents a positive or a negative change for the customers), the congestion levels of the call center and the characteristics of customers have statistically significant impacts on abandonment behavior. However, because of the complexity, no staffing decisions considering delay announcements have been studied yet.

Although SIPP combined with an Erlang model is the most commonly used method in practice, there are a number of problems with such an approach. We discuss them one by one.

In the first place, the queue is not in a stationary situation at the beginning of each interval. Depending on the parameters of the previous intervals, you might, for example, expect a backlog. There are a number of methods available to handle this, of which, according to Babat (2015), the stationary backlog carryover approach from Stolletz (2008) performs best.

In the second place, the SIPP predicts expected performance. If you schedule using SIPP at the level of your SLA, then, in roughly 50% of the cases, you won't reach your daily SLA. The error can be quite big (Roubos et al. 2012), which is, in practice, unacceptable. A solution might be to look at quantiles of the service level distribution (Roubos et al. 2012), but the problem is usually solved by intraday management.

In the third place, on top of the transient effect we just discussed, there is uncertainty about the arrival rate, the overdispersion we find in the previous section. Scheduling according to the expected rate is suboptimal; Ding and Koole (2022) propose a method that integrates the intraday adaptations into the staffing step, leading to a newsvendor-type

staffing method. Whitt (2006), Steckley et al. (2004), and Liao et al. (2012) also incorporate uncertain arrival rates into staffing methods.

In the fourth place, there are many factors that influence performance that are not modeled by Erlang C or Erlang A, such as the impact of short, unscheduled breaks and the behavior of agents under longer periods of high workload. Only recently the first attempt to validate the Erlang models based on realized service levels was undertaken (Ding et al. 2020). By studying agent data together with call data, it is indeed found that breaks have a huge impact on performance. One way to solve this is to move to a statistical/machine learning approach that takes all features into account as in Li et al. (2020b). This is even better than simulation because, implicitly, the behavior of the agents is taken into account; to use simulation, it has to be modeled explicitly, which is hard because of differences between agents and the lack of knowledge on how and when, for example, breaks are taken.

It is hard to obtain qualitative insights from the Erlang formulas, for example, how they behave when you increase scale. Square-root staffing does. For λ, the arrival rate, and β, the average handling time, it says that staffing should be at $\lambda\beta + \alpha\sqrt{\lambda\beta}$ with α a parameter depending on the SL only. The square root can intuitively be interpreted: if you add two independent and identically distributed random variables, then the standard deviation is multiplied by $\sqrt{2}$. The same holds for safety staffing because it is there to handle fluctuations in load. This clearly shows the economies of scale, which is one of the reasons why we want agents to be multiskilled. It also shows decreasing returns as $\lambda\beta + \alpha\sqrt{\lambda\beta}$ is concave in $\lambda\beta$, which tells us that not all agents need to be multiskilled. Halfin and Whitt (1981) introduce this Halfin–Whitt regime in which the load increases but the delay probability is held constant. Since then, many papers study this regime in many different variants; see Braverman (2020), Gamarnik and Stolyar (2012), Harrison and Zeevi (2004), and Reed (2009) for recent ones. Unfortunately, these ideas are very little used in practice.

Next to inbound, a variety of other channels are used. They can be divided into synchronous and asynchronous communication. Email, webforms, and old-fashioned mail and fax are asynchronous. Usually the time to answer is multiple hours or days—at least multiple intervals. This means that fluctuations have to be dealt with by flexibility in scheduling, not by safety staffing. Inbound is synchronous. Another noteworthy synchronous channel is chat. Its difference with inbound from the point of view of WFM is that a chat agent can do multiple chats in parallel, usually two or three. When all agents are saturated, customers wait in the queue, just as with inbound. This parallelism increases efficiency. When an agent answers one customer, the other(s) can formulate their responses. However, it makes the total handling time per chat longer: sometimes a customer has to wait for a chat to become available. Quantifying the durations are somewhat challenging but can then be used to extend the Erlang models for chat systems. See Koole (2013) for more details on implementation.

Moving decisions to a later moment when better information is available is a general principle to improve decision making. One way to do this is to move the decision on which type of task to do from the schedule to the routing. In a multichannel environment, this leads to blending in a multiskill environment to SBR. We now discuss how staffing can be done in these environments.

Blending is usually executed by blending synchronous and asynchronous channels, such as inbound and email or outbound. When the asynchronous channel can be interrupted to deal with priority with inbound, then staffing is easy: inbound is staffed as discussed, and the overcapacity with respect to the expected load is filled with email. Things get more complicated when the asynchronous channel cannot be interrupted as in the case of outbound. This case is studied in Bhulai and Koole (2003) and Gans and Zhou (2003). Both the routing policy and, implicitly, staffing are determined. Capacity allocation between inbound and back office via robust optimization is considered in Mattia et al. (2017). No papers discuss the blending of synchronous channels, such as inbound and chat. A simple policy could be to assign an agent to the channel with the longest waiting customer. To use as few agents as possible for chat, chats should be assigned to agents already handling chats but who are not yet saturated (as multiple chats can be handled in parallel by the same agent). Note that the way in which blending is done as part of the routing should be reflected in the staffing method that is used.

Now, we move to SBR. Because of the lack of closed-form formulas, simulation is the only viable option for SBR apart from some approximations based on models without waiting (Chevalier et al. 2004b, Pot et al. 2008) or based on fluid models considering abandonment targets as the QoS constraints (Bassamboo et al. 2006, Gurvich et al. 2010, Bodur and Luedtke 2017). But why run a long-term simulation to find stationary behavior when it is easier to do a short-term transient analysis, for example, of a day? For this reason, most studies tackle right away the scheduling problem, resulting in little literature on multiskill staffing alone. Some notable exceptions are Cezik and L'Ecuyer (2008), Gurvich et al. (2010), Harrison and Zeevi (2005), and Chan et al. (2014b, 2016).

Note that, in practice, for the schedule to be realistic, the routing rules of the ACD should be used in the scheduling software. Unfortunately, this is rarely done, often because the scheduling tool doesn't allow it. This puts an additional burden on real-time management because the scheduling algorithm consistently makes a bias schedule.

Most research opportunities in the area of staffing are, in our opinion, related to data-driven methods for blended and multiskill systems. As many parameters are unknown and hard to quantify, black-box methods might perform well. This is especially the case for other channels than inbound. Similar ideas are also discussed in healthcare; see Armony et al. (2015).

4.3. Agent Scheduling

Agent scheduling is the operational process in which agents get assigned to shifts and activities during these shifts. Activities include the channel and/or skills they have to work on and also paid breaks, meetings, training sessions, etc. Next to the routing, which is part of the telephony/omnichannel switch, it is the part of WFM that is most often supported by specialized software. These tools are crucial because they administer the schedules and communicate them to the agents, which is nearly impossible to do with a spreadsheet. This has led to a wide choice of software vendors; see, for example, TrustRadius (2020). However, little is known about their exact workings. Erlang C and simulation are used for staffing, and at least some use heuristics for scheduling. This can lead to long running times, especially for the tools using simulation. Fukunaga et al. (2002) give some details about Verint (called Blue Pumpkin at the time). Smaller call centers and also the ones with fewer scheduling issues (for example, because they are only open during business hours), often schedule using a spreadsheet. Agent scheduling in general is hardly studied in the literature: usually unpersonalized shifts are determined without activities within the shifts, which is actually shift scheduling.

Agent scheduling consists of multiple steps and is more complicated in multiskill and/or multichannel settings. In this section, we first discuss the advantages of integrating (some of) the steps. After that, we discuss methods for single-skill inbound call centers, then multiskill, and finally multichannel methods.

In its simplest form, agent scheduling consists of the following four steps:

- For each interval, the required staffing levels are determined (e.g., using an Erlang formula).
- The most efficient way to cover the staffing needs by the available shifts is determined (potentially using integer linear programming (ILP)).
- These shifts are assigned to agents in some way (for example, by letting them choose in the order of seniority).
- Activities are assigned during the scheduled working time of the agents.

The first step, staffing, has just been discussed. The first to formulate a solution for the covering problem of the second step was Dantzig (1954), in which he considered toll booths at a U.S. bridge.

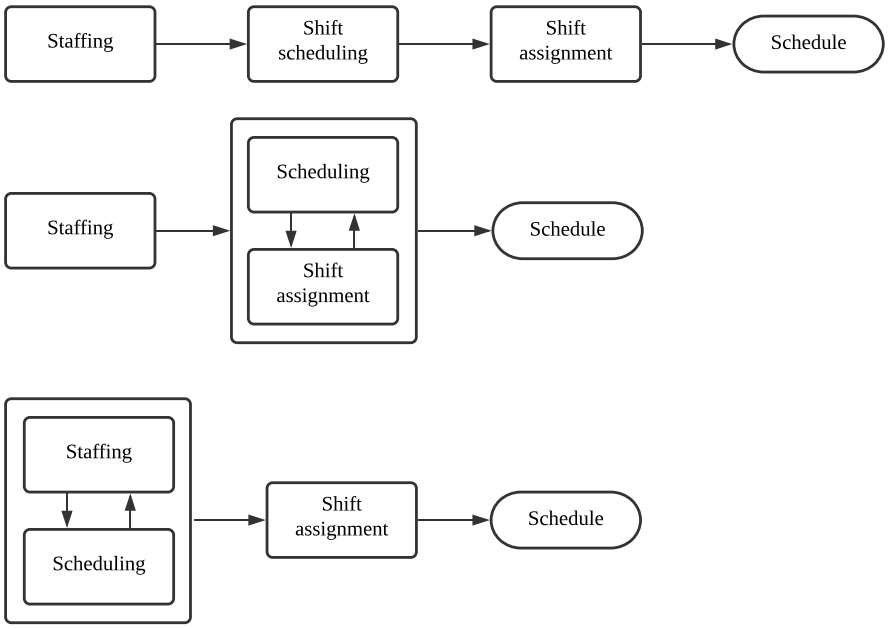
There are various reasons why successively executing the four steps is highly suboptimal and even infeasible. Often, employees have different types of contracts; therefore, in step 2, different groups of shifts should be identified; otherwise, no match between agents and shifts can be made. Furthermore, many agents have personal preferences. Satisfying them as much as possible is crucial for employee satisfaction, making that the schedule should be made at the individual level, integrating steps 2 and 3. Dealing with these personal preferences is evidently part of WFM software but hardly studied scientifically. Step 4 is even less studied, and the requirement that, for example, meetings are attended by a group of agents requires an integration of steps; otherwise, their shifts might not even be overlapping.

A much better studied subject is the integration of steps 1 and 2 as in the lower part of Figure 3. The reason for combining them is that the staffing levels of step 1 are hard to cover with shifts, leading to considerable overstaffing. Often SLAs are formulated at the daily level; thus, SLs are allowed to fluctuate a bit—certainly if that leads to more efficient schedules and if the daily constraints are met. Integrating steps 1 and 2 makes the optimization problem nonlinear. It can be rewritten as ILP but at the cost of having many binary variables. Add to this the fact that we should schedule at the weekly level (necessary because of constraints on the schedules related to numbers of working days per week and start times); then, we have a problem so big that it can only be solved with heuristics, such as local search. In multiskill settings, we need simulation to get a reliable evaluation of possible solutions, leading us to simulation optimization with stochasticity on the SL constraints—problems that are known to be notoriously difficult.

Now, we discuss single-skill scheduling methods, then multiskill scheduling, and finally multichannel methods. We focus on steps 1 and 2, which can be separated or integrated. Some methods allow for different groups of shifts and constraints on their numbers, which comes close to step 3.

In a single-skill setting, Koole and van der Sluis (2003) use SIPP to determine staffing levels and show that, under a very simple shift structure, a suitable local algorithm can find optimal schedules. In a transient single-skill setting, Atlason et al. (2008) integrate staffing and shift scheduling and use simulation to generate cutting planes used in the shift-optimization module. Liao et al. (2013) also use simulation. They combine stochastic programming and robust optimization to work out scheduling with uncertain arrival rates. Robbins and Harrison (2010) solve a stochastic scheduling problem to minimize the combined cost of agents and missing QoS targets. Gans et al. (2015)

Figure 3. Options for Integrating the Steps of the Scheduling Process



consider a two-stage scheduling problem that allows adding and removing agents based on updated forecasts at midday.

In a multiskill situation, Pot et al. (2008) and Bhulai et al. (2008) use an overflow approximation for SBR, similar to Chevalier and Tabordon (2004), to build a multiskill scheduling algorithm. Bodur and Luedtke (2017) solve a two-stage stochastic programming for scheduling with Benders decomposition. A main drawback of these approximations are unrealistic assumptions or unrealistic fluid approximations. Moreover, service levels cannot be approximated because the models are based on rejection models without queues. Again using simulation, Cezik and L'Ecuyer (2008) extend the approach of Atlason et al. (2008) to multiskill staffing. Avramidis et al. (2010) extend the cutting plane method to solve the scheduling problem over a day (i.e., multiple periods). Running times, however, are very long.

The current state of the art is Li et al. (2020a), which combines simulation optimization with ML: an ML model is fit to a number of simulations, and then, a local search over the ML approximation is done. This makes it possible to solve industrial-size, weekly, multiskill, multichannel problems in several minutes. Solving the same problem without using ML takes much longer; see Li et al. (2019). Note that it is inevitable that the SL fluctuates because of our transient daily SL objective. In call centers, planners spend long hours adapting schedules manually to get smooth service levels—from a mathematical perspective, a useless and expensive practice.

Note that all these problems consider shift scheduling: they determine shifts but do not assign agents and do not determine the activities within the shifts. This adds multiple layers of complexity far beyond the current state of the art, but it is required in operations and done by WFM software. On the other hand, it can be argued that the activity assignment should be done at the routing level although some activities (such as meetings) need to be planned in advance. The fact that these methods are not at the level of agents, but, at best, at the agent-group level makes them better suitable for capacity planning, which is the subject of the next section.

The big remaining challenge not yet addressed in the literature is the construction of even faster simulation-optimization methods or (meta)heuristics using very accurate approximations that solve the combined problem of shift and task scheduling in several minutes. Most WFM systems either use very crude approximations based on Erlang C or have excessive runtimes of, for example, a whole night on a fast computer. These methods should work for a multiskill, multichannel environment and take all shareholder interests (agent satisfaction, costs, and SL) into account.

5. Capacity Planning

Capacity planning is the holy grail of WFM. Hiring and training new agents is a lengthy process that can easily take three months or more; thus, the capacity has to be planned well in advance. To be able to do long-term planning,

you have to take into account how you deal with all the shorter term processes. Thus, all decisions at all levels impact capacity planning, which makes it potentially highly complicated. There are two ways to do capacity planning. If the objective is to determine the total required workforce, then an approach focused on obtaining weekly totals suffices. This is what most call centers do, using a homemade spreadsheet. If, also, the types of contracts of new hires and/or the training of new skills need to be determined, then a more detailed approach is necessary.

Both approaches start again with forecasting. Tactical forecasting is similar to operational forecasting with the exception that events play less of a role: events such as marketing actions are often not yet scheduled and, therefore, should not be modeled separately.

Of big impact can be the increase in proficiency of the agents, and with that, a reduction of the handling times. Therefore, the AHT needs to be forecasted as well. Gans et al. (2010) and Ibrahim et al. (2016a) analyze call center data showing significant heterogeneity and learning effects although Ibrahim et al. (2016a) also report on agents whose AHT increases. Ding et al. (2020) fit curves that are exponential in time to the AHT of individual agents.

Although the (net) workload can often be calculated at the weekly level, it is still required to use safety staffing to compute the (net) workforce. In theory, staffing models have to be used at the interval level because the required staffing level is a nonlinear function of the volume. However, practice shows that staffing according to the average volume is very close to the average staffing level. From the concavity of the square root staffing approximation it follows that it slightly overstaffs. To translate the gross workload (the net workload plus safety staffing) to gross workforce, we have to add all forms of shrinkage and the shift inefficiency. Shrinkage is the term used for all activities that prevent agents from being available for phone work (or other types of contacts) from holidays and illness to meetings and paid breaks. Shift inefficiency is the fact that we cannot cover the required staffing levels exactly with shifts. Note that the levels of shrinkage, such as sick leave, can vary over time and need to be forecasted.

More complicated are decisions related to the shifts and initial skill sets of new agents and decisions about the training of new skills for existing agents. To determine which types of agents to hire, shift scheduling has to be done over a longer period, starting from the current pool of agents, taking agent resignations and shrinkage into account. The operational schedule should take activities such as meetings and short breaks into account, capacity planning all of forms of shrinkage. Note that they are sometimes unpredictable, such as illness, and sometimes planable, such as when agents go on holidays or when meetings take place. Both types complicate capacity planning. Many call centers do capacity planning in a grossly simplified way by replacing all randomness and advanced calculations by calculations based on historical fractions between the types of workload and workforce. Probably even more call centers use no calculations at all but make rough estimates, potentially making big errors in the amount of required agents and especially in the optimal contract and skill mix. This regularly leads to long periods with understaffing and/or the excessive use of flexibility offered by third parties. Very few call centers utilize more advanced technology. Finding the optimal agent pool and determining which agents are best to be added to the current pool is hardly done.

There are no papers solving the pool optimization problem completely. Some papers, such as Avramidis et al. (2010), Bhulai et al. (2008), and Li et al. (2019), as discussed in the previous section, solve the shift-scheduling problem for a week or a day, but methods have to be found to extend this to longer periods or to somehow aggregate weekly results to, say, a year. Furthermore, all forms of shrinkage have to be added. In our opinion, this is the biggest remaining challenge in WFM, and the only possible solution method we see is a time-consuming simulation-optimization procedure, possibly sped up using ML as in Li et al. (2020a).

A simpler solution to the pool-composition problem might be to use some rule of thumb. Chevalier et al. (2004a) claim, using approximations based on networks of overflow queues, that 80% specialized and 20% fully flexible agents works surprisingly well in many situations. This holds for the staffing problem; random forms of shrinkage likely make the need for flexible agents higher in the pool-composition problem. Also Wallace and Whitt (2005) show, using simulations, that a little flexibility goes a long way in a situation in which agents have one or two skills and a topology that "connects" all skills.

6. Budget Planning

Budget planning is often done one to two years in advance and involves higher levels of management than capacity planning. This has two important consequences:

- The forecast on which the budget plan is based is likely to be highly unreliable.
- Forecasts purely based on historical volumes are not accepted by the users.

Indeed, long-term forecasts are mainly determined by the trend. Extrapolating the trend misses all trend changes that are likely to occur, leading to very bad strategic forecasts (Makridakis 1990). Next to that, the decision makers, external to the call center, who need to approve the budget want to understand the forecast; they prefer a story,

such as, "Next year, we will have a growth of 20% in all product categories, but an improved self-service reduces the call-to-customer ratio by 10%; therefore, our increase is only 8%." Note that this moves the responsibility of the correctness of the forecast from the call center forecast to the sales forecast. Of course, throughout the year, the validity of the assumptions should be checked, and the budget should be adapted accordingly. Instead of a single forecast, multiple scenarios could be formulated and adapted, leading to different budget requirements.

Based on the forecast, schedules can be made just as for agent scheduling. Then, the costs of these schedules could be determined leading to the budget. However, apart from some practicalities, such as the lack of information on agents still to be hired, runs are often too long to compute the multiyear horizon needed for the budget, especially because Excel, which is used for this in 99% of all organizations, is not appropriate for this kind of calculation. A simple, fast calculation is to estimate the budget proportional to the volumes: if the volume increases by x%, then the costs will also increase by x%. Of course, we make an error: costs are not linear in the forecast, but for small changes, the error is expected to be small—probably much smaller than the forecasting error. More advanced methods, such as an ML model to estimate costs based on the forecast and other parameters, are also successfully used in practice. Note that, although successful in practice, these methods lack a scientific basis, offering a research opportunity.

7. Conclusion

A call center needs to adapt to changes in the environment. We not only refer to changes in call arrivals, but also to changes in the system. We went from single-skill systems to multiskill a few decades ago and, nowadays, from single to multichannel systems. As a consequence, call centers also face new challenges related to routing, blending, and so on. One can find that the open research questions discussed here are not the same as the ones addressed in similar work years ago. In our opinion, operations research should be aligned with industry, understanding the real need, working out solutions for current problems. We hope this overview will help in better achieving this goal.

We also stress the importance of analyzing real data, not only in call center systems, but also in other areas, such as healthcare and manufacturing systems. Although call center data are not as difficult to obtain as hospital data (Armony et al. 2015), it can also be an obstacle to some researchers. Making algorithms and data publicly available on, for example, github can make sharing much easier.

Sometimes, system changes can be very sudden, for example, the ones caused by the recent coronavirus pandemic. Because of the surge of the number of agents and customers working from home, longer handling times and longer customer patience are observed. If we keep using old models to predict the parameters, it can easily result in big errors and huge costs. What we need are adaptive approaches that can quickly detect and respond to the changes. Moreover, working from home also increases the flexibility of shifts. Because agents do not have traveling costs, they are more willing to take short shifts or work overtime if required. This should be considered in capacity planning and agent scheduling or even in the design. It is also interesting to note that some activities, such as chats/emails, which do not need to have real-time communication, become more preferred by agents when they work from home. Consequently, the amount of time spent on these activities is required to be more fair proportionally among agents. We do not see these changes disappearing in a short time. Many of these new phenomena raise interesting research opportunities.

Acknowledgments

The authors are grateful to the stimulating environment that CCmath offers and that made this paper and its research possible. The authors are especially grateful to Giuseppe Catanese, Alex Roubos, and Wout Bakker for their feedback. CCmath has its own algorithms for forecasting, staffing, and scheduling of which we were not allowed to disclose the details for commercial reasons. Note that this overview might be biased toward the situation found at CCmath's clients. The second author also thanks the Vrije Universiteit for the hospitality that was offered to her over multiple years.

References

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

Akşin OZ, de Vericourt F, Karaesmen F (2008) Call center outsourcing contract analysis and choice. *Management Sci.* 54(2):354–368.

Akşin Z, Ata B, Emadi SM, Su C-L (2017) Impact of delay announcements in call centers: An empirical approach. *Oper. Res.* 65(1):242–265.

Aktekin T (2014) Call center service process analysis: Bayesian parametric and semi-parametric mixture modeling. *Eur. J. Oper. Res.* 234(3):709–719.

Aktekin T, Soyer R (2014) Bayesian analysis of abandonment in call center operations. *Appl. Stochastic Models Bus. Indust.* 30(2):141–156.

Aldor-Noiman S, Feigin PD, Mandelbaum A (2009) Workload forecasting for a call center: Methodology and a case study. *Ann. Appl. Statist.* 3(4):1403–1447.

Antipov A, Meade N (2002) Forecasting call frequency at a financial services call centre. *J. Oper. Res. Soc.* 53(9):953–960.

Armony M, Maglaras C (2004a) Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52(4):527–545.

Armony M, Maglaras C (2004b) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.
Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* 58(3):624–637.
Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.
Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.
Atar R (2005) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15(4):2606–2650.
Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.* 58(5):1427–1439.
Atlason J, Epelman MA, Henderson SG (2008) Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Sci.* 54(2):295–309.
Avaya (2011) How to balance business goals with Avaya Business Advocate. Accessed June 14, 2023, http://www.stlcom.com/wp-content/uploads/2016/07/49_AvayaBusinessAdvocateBrochure.pdf.
Avramidis AN, L'Ecuyer P (2005) Modeling and simulation of call centers. *Proc. 2005 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 144–151.
Avramidis AN, Deslauriers A, L'Ecuyer P (2004) Modeling daily arrivals to a telephone call center. *Management Sci.* 50(7):896–908.
Avramidis AN, Chan W, Gendreau M, L'Ecuyer P, Pisacane O (2010) Optimizing daily agent scheduling in a multiskill call center. *Eur. J. Oper. Res.* 200(3):822–832.
Babat F (2015) Evaluation of service level approximations in call centers. Accessed September 6, 2020, beta.vu.nl/en/Images/werkstuk-babat_tcm235-458387.pdf.
Baccelli F, Hébuterne G (1981) On queues with impatient customers. Kylstra FJ, ed. *Performance '81* (North-Holland, Amsterdam), 159–179.
Bakker W, Catanese G, Koole G, Li S, Roubos A (2019) More precision with less data: A new approach to intra-day patterns. Accessed September 2, 2020, www.ccmath.com/spline-method.
Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* 54(3):419–435.
Bhulai S, Koole GM (2003) A queueing model for call blending in call centers. *IEEE Trans. Automatic Control* 48(8):1434–1438.
Bhulai S, Pot SA, Koole GM (2008) Simple methods for shift scheduling in multi-skill call centers. *Manufacturing Service Oper. Management* 10(3):411–420.
Bodur M, Luedtke JR (2017) Mixed-integer rounding enhanced Benders decomposition for multiclass service-system staffing and scheduling with arrival rate uncertainty. *Management Sci.* 63(7):2073–2091.
Braverman A (2020) Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Math. Oper. Res.* 45(3):1069–1103.
Bromley L (2001) Erlang for Excel. Accessed September 7, 2020, www.erlang.co.uk.
Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.
Cezik MT, L'Ecuyer P (2008) Staffing multiskill call centers via linear programming and simulation. *Management Sci.* 54(2):310–323.
Chan W, Koole G, L'Ecuyer P (2014a) Dynamic call center routing policies using call waiting and agent idle times. *Manufacturing Service Oper. Management* 16(4):544–560.
Chan W, Ta TA, L'Ecuyer P, Bastin F (2016) Two-stage chance-constrained staffing with agent recourse for multi-skill call centers. *2016 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3189–3200.
Chan W, Ta TA, L'Ecuyer P, Bastin F, Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S (2014b) Chance-constrained staffing with recourse for multi-skill call centers with arrival-rate uncertainty. *Proc. 2014 Winter Simulation Conf.* (IEEE Press, Piscataway, NJ), 4103–4104.
Channouf N, L'Ecuyer P (2012) A normal copula model for the arrival process in a call center. *Internat. Trans. Oper. Res.* 19(6):771–787.
Chevalier P, Tabordon N (2004) Overflow analysis and cross-trained servers. *Internat. J. Production Econom.* 85:47–60.
Chevalier P, Shumsky RA, Tabordon N (2004a) Routing and staffing in large call centers with specialized and fully flexible servers. Google Scholar Accessed September 8, 2020.
Chevalier P, Shumsky RA, Tabordon N (2004b) Routing and staffing in large call centers with specialized and fully flexible servers. *Manufacturing Service Oper. Management*. Accessed 14 June 2023, http://mba.tuck.dartmouth.edu/pages/faculty/robert.shumsky/xtrain_large_cc.pdf.
Cleveland B (2019) *Call Center Management on Fast Forward*. ICMI, Colorado Springs (CO) (Call Center Press).
Crow EL (1958) The mean deviation of the Poisson distribution. *Biometrika* 45(3–4):556–562.
Cui L, Tezcan T (2016) Approximations for chat service systems using many-server diffusion limits. *Math. Oper. Res.* 41(3):775–807.
Dantzig GB (1954) A comment on Edie's "Traffic delays at toll booths." *J. Oper. Res. Soc. Amer.* 2(3):339–341.
Deslauriers A, L'Ecuyer P, Pichitlamken J, Ingolfsson A, Avramidis AN (2007) Markov chain models of a telephone call center with call blending. *Comput. Oper. Res.* 34(6):1616–1645.
Ding S, Koole GM (2022) Optimal call center forecasting and staffing. *Probab. Engrg. Informational Sci.* 36(2):254–263.
Ding S, Koole G, van der Mei RD (2015) On the estimation of the true demand in call centers with redials and reconnects. *Eur. J. Oper. Res.* 246(1):250–262.
Ding S, Li S, Koole G, Yuce EI, van der Mei R, Stolletz R (2020) Data analysis and validation of call center staffing and workforce models. Working paper, Vrije Universiteit, Amsterdam.
Dixon M, Toman N, Delisi R (2013) *The Effortless Experience* (Penguin, New York).
Feigin P (2006) Analysis of customer patience in a bank call center. Working paper, The Technion, Haifa, Israel.
Fukunaga A, Hamilton E, Fama J, Andre D, Matan O, Nourbakhsh I (2002) Staff scheduling for inbound call centers and customer contact centers. *AI Magazine* 23(4):30–40.
Gamarnik D, Stolyar AL (2012) Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: Asymptotics of the stationary distribution. *Queueing Systems* 71(1–2):25–51.
Gans N, Zhou Y-P (2003) A call-routing problem with service-level constraints. *Oper. Res.* 51(2):255–271.
Gans N, Zhou Y-P (2007) Call-routing schemes for call-center outsourcing. *Manufacturing Service Oper. Management* 9(1):33–50.
Gans N, Koole GM, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Gans N, Liu N, Mandelbaum A, Shen H, Ye H (2010) Service times in call centers: Agent heterogeneity and learning with some operational consequences. *IMS Collections* 6:99–123.

Gans N, Shen H, Zhou Y-P, Korolev N, McCord A, Ristock H (2015) Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing Service Oper. Management* 17(4):571–588.

Ghareeb ET, Abd Elatif MM, El Bakry HM (2016) Optimal routing selection using analytical hierarchy process. *Internat. J. Adv. Comput. Techn.* 8(3):44–57.

Global Industry Analysts, Inc. (2023) Call Centers: Global Strategic Business Report. Accessed June 14, 2023, https://www.researchandmarkets.com/reports/338444/call_centers_global_strategic_business_report.

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

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

Gurvich I, Luedtke J, Tezcan T (2010) Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Sci.* 56(7):1093–1115.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–587.

Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Oper. Res.* 52(2):243–257.

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

Hasija S, Pinker EJ, Shumsky RA (2008) Call center outsourcing contracts under information asymmetry. *Management Sci.* 54(4):793–807.

Hathaway BA, Emadi SM, Deshpande V (2017) Queue now or queue later: An empirical study of callers' redial behaviors. Working paper. Accessed June 14, 2023, https://kenaninstitute.unc.edu/publication/queue-now-or-queue-later-an-empirical-study-of-callers-redial-behaviors/.

Hathaway BA, Emadi SM, Deshpande V (2020) Don't call us, we'll call you: An empirical study of caller behavior under a callback option. *Management Sci.* 67(3):1508–1526.

Huang H, Jiang M, Ding Z, Zhou M (2019) Forecasting emergency calls with a Poisson neural network-based assemble model. *IEEE Access* 7:18061–18069.

Hyndman RJ (2013) Forecasting with daily data. Accessed June 26, 2020, robjhyndman.com/hyndsight/dailydata.

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (O Texts).

Ibrahim R, L'Ecuyer P (2013) Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing Service Oper. Management* 15(1):72–85.

Ibrahim R, Whitt W (2009) Real-time delay estimation based on delay history. *Manufacturing Service Oper. Management* 11(3):397–415.

Ibrahim R, Whitt W (2011) Wait-time predictors for customer service systems with time-varying demand and capacity. *Oper. Res.* 59(5):1106–1118.

Ibrahim R, L'Ecuyer P, Shen H, Thiongane M (2016a) Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *Eur. J. Oper. Res.* 250(2):480–492.

Ibrahim R, Ye H, L'Ecuyer P, Shen H (2016b) Modeling and forecasting call center arrivals: A literature survey and a case study. *Internat. J. Forecasting* 32(3):865–874.

Jack EP, Bedics TA, McCary C (2006) Operational challenges in the call center industry: A case study and resource based framework. *Management Service Quality* 16(5):477–500.

Jalal ME, Hosseini M, Karlsson S (2016) Forecasting incoming call volumes in call centers with recurrent neural networks. *J. Bus. Res.* 69(11):4811–4814.

Jongbloed G, Koole GM (2001) Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.* 17(4):307–318.

Jouini O, Akşin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* 13(4):534–548.

Jouini O, Dallery Y, Rabie N-A (2008) Analysis of the impact of team-based organizations in call center management. *Management Sci.* 54(2):400–414.

Jouini O, Koole GM, Roubos A (2013) Performance indicators for call centers with impatience. *IIE Trans.* 45(3):341–354.

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53(282):457–481.

Kim S, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing Service Oper. Management* 16(3):464–480.

Koole GM (2013) *Call Center Optimization* (MG books, Amsterdam).

Koole GM, van der Sluis HJ (2003) Optimal shift scheduling with a global service level constraint. *IIE Trans.* 35(11):1049–1055.

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

Legros B, Jouini O (2019) On the scheduling of operations in a chat contact center. *Eur. J. Oper. Res.* 274(1):303–316.

Legros B, Jouini O, Dallery Y (2015a) A flexible architecture for call centers with skill-based routing. *Internat. J. Production Econom.* 159:192–207.

Legros B, Jouini O, Koole G (2016) Optimal scheduling in call centers with a callback option. *Performance Evaluation* 95:1–40.

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

Legros B, Ding S, van der Mei R, Jouini O (2017) Call centers with a postponed callback offer. *OR Spectrum* 39(4):1097–1125.

Li S, Koole G (2020) An adaptive call center routing policy. Working paper.

Li S, Koole G, Jouini O (2019) A simple solution for optimizing weekly agent scheduling in a multi-skill multi-channel contact center. *Proc. 2019 Winter Simulation Conf.*(IEEE, Piscataway, NJ), 3657–3668.

Li S, Wang Q, Koole G (2020a) Optimal contact center staffing and scheduling with machine learning. Working paper.

Li S, Koole G, Yuce EI, Catanese G (2020b) A machine learning approach to call center staffing. Working paper.

Liao S, van Delft C, Vial J-P (2013) Distributionally robust workforce scheduling in call centres with uncertain arrival rates. *Optim. Methods Software* 28(3):501–522.

Liao S, Koole G, van Delft C, Jouini O (2012) Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum* 34(3):691–721.

Luo J, Zhang J (2013) Staffing and control of instant messaging contact centers. *Oper. Res.* 61(2):328–343.

Maister DH (1985) The psychology of waiting lines. Czepiel J, Solomon MR, Surprenant CF, eds. *The Service Encounter* (Lexington Books).

Makridakis SG (1990) *Forecasting, Planning, and Strategies for the 21st Century* (The Free Press).

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Oper. Res.* 52(6):836–855.

Mandelbaum A, Zeltyn S (2004) The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the $m/m/n+g$ queue. *OR Spectrum* 26(3):377–411.
Mattia S, Rossi F, Servilio M, Smriglio S (2017) Staffing and scheduling flexible call centers by two-stage robust optimization. *Omega* 72:25–37.
Mazareanu E (2019) Call center market size by region 2012-2017. Accessed June 22, 2020, www.statista.com/statistics/881033/call-center-market-size-region.
Mehrotra V, Ross K, Ryder G, Zhou Y-P (2012) Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing Service Oper. Management* 14(1):66–81.
Milner JM, Olsen TL (2008) Service-level agreements in call centers: Perils and prescriptions. *Management Sci.* 54(2):238–252.
Palm C (1953) Methods of judging the annoyance caused by congestion. *Tele* 4:189–208.
Pang G, Perry O (2015) A logarithmic safety staffing rule for contact centers with call blending. *Management Sci.* 61(1):73–91.
Pot SA, Bhulai S, Koole GM (2008) A simple staffing method for multi-skill call centers. *Manufacturing Service Oper. Management* 10(3):421–428.
Powell SG, Baker KR, Lawson B (2009) Impact of errors in operational spreadsheets. *Decision Support Systems* 47(2):126–132.
Price B, Jaffe D (2008) *The Best Service Is No Service* (Wiley).
Reed J (2009) The $g/gi/n$ queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* 19(6):2211–2269.
Robbins TR, Harrison TP (2010) A stochastic programming model for scheduling call centers with global service level agreements. *Eur. J. Oper. Res.* 207(3):1608–1619.
Roubos A, Jouini O (2013) Call centers with hyperexponential patience modeling. *Internat. J. Production Econom.* 141(1):307–315.
Roubos A, Bhulai S, Koole GM (2017) Flexible staffing for call centers with non-stationary arrival rates. Boucherie RJ, van Dijk NM, eds. *Markov Decision Processes in Practice* (Springer), 487–503.
Roubos A, Koole G, Stolletz R (2012) Service-level variability of inbound call centers. *Manufacturing Service Oper. Management* 14(3):402–413.
Saltzman RM, Mehrotra V (2001) A call center uses simulation to drive strategic change. *Interfaces* 31(3):87–101.
Shen H, Huang J (2008) Interday forecasting and intraday updating of call center arrivals. *Manufacturing Service Oper. Management* 10(3):391–410.
Soyer R, Tarimcilar MM (2008) Modeling and analysis of call center arrival data: A Bayesian approach. *Management Sci.* 54(2):266–278.
Steckley SG, Henderson SG, Mehrotra V (2004) Service system planning in the presence of a random arrival rate. Technical report, Cornell University Operations Research and Industrial Engineering.
Stolletz R (2008) Approximation of the non-stationary $m(t)/m(t)/c(t)$-queue: The stationary backlog-carryover approach. *Eur. J. Oper. Res.* 190(2):478–493.
Sze DY (1984) A queueing model for telephone operator staffing. *Oper. Res.* 32(2):229–249.
Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* 60(6):1574–1593.
Taylor JW (2008) A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Sci.* 54(2):253–265.
Tezcan T, Zhang J (2014) Routing and staffing in customer service chat systems with impatient customers. *Oper. Res.* 62(4):943–956.
TrustRadius (2020) Call center workforce optimization software. Accessed September 8, 2020, www.trustradius.com/call-center-workforce-optimization.
van Eeden K, van der Hilst E, Koole G (2013) Errors in the Erlang97 Excel add-in module. Accessed September 7, 2020, www.gerkoole.com/publications.
Wallace RB, Whitt W (2005) A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* 7(4):276–294.
Ward AR, Armony M (2013) Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Oper. Res.* 61(1):228–243.
Weinberg J, Brown LD, Stroud JR (2007) Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *J. Amer. Statist. Assoc.* 102(480):1185–1198.
Whitt W (2005) Engineering solution of a basic call-center model. *Management Sci.* 51(2):221–235.
Whitt W (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* 15(1):88–102.
Ye H, Brown LD, Shen H (2020) Hazard rate estimation for call center customer patience time. *IISE Trans.* 52(8):890–903.
Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51:361–402.
Zhan D, Ward AR (2014) Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing Service Oper. Management* 16(2):220–237.