Determining optimal staffing levels in multiple skill inbound call centers : a literature survey

Determining Optimal Staffing Levels in Multiple Skill Inbound Call Centers - a Literature Survey

M.J. Stegeman and M.H. Jansen-Vullers
Technische Universiteit Eindhoven

Preface

This literature review aims to find techniques for determining the optimal staffing levels in a multiple skill inbound call center. The domain of the research is therefore set to mainly inbound call centers (although also models of call centers handling in- and outbound calls are described) that can handle multiple types of customers and service requests. Three phases of the (re)design process of staffing call center staffing are discussed to answer the research assignment:

- Performance measurements in call centers
- Forecasting input parameters for call center modeling
- Determining (optimal) staffing levels in call centers

This resulted in a table overview with the different authors and their approaches towards determining optimal staffing levels in call centers. Of course conclusions have been drawn from the research and a list of future research areas and learning experiences is provided.

In the literature research a number of existing approaches is being described. This list is not complete, of course. On the one hand because of the available time to perform the literature research. On the other hand because the authors and approaches listed in this literature research are by far the most referred to in scientific literature. Other authors and approaches dealing with the research subject exist, but are not as renowned and widely referred to as the major part of the authors discussed in this literature review.

The most important characteristics, objectives and differences with other approaches will be provided for every author and approach described. Details on specific characteristics or underlying mathematical models of the approaches can either be found in the appendices or in the articles of the referred authors.

Mark Stegeman
Monique Jansen-Vullers

Contents

1. Introduction

Browsing the amply available literature on call centers and all of its aspects one can find one specific definition of a call center that is referred to a lot. Mehrota (1997) gives a rather broad definition of a call center:

"Any group whose principal business is talking on the telephone to customers or prospects."

The group can be centralized, geographically spread or composed with agents in individual offices. Nowadays call centers play a considerably important role in the service processes of companies. They are closest to the (potential) customers and therefore have a major influence on the customers' perception of the company. In different media, results from surveys at customer services (e.g. help desks and call centers) are published, with performance statistics and company comparisons. The influence of call centers in our service-driven economies has undoubtedly grown last decade. This has been recognized in a lot of articles and by companies involved in the call center processes. For example Gans et al. (2003), Grossman et al. (2001), Mehrota (1997), Dawson (2004) and Stolletz (2003) mention the rise of call center presence in the service business and state this with impressing numbers of growth in call centers, employees and turnover. In the early years call centers were seen as cost centers (Bapat and Pruitte, 1998). Nowadays service has become more important and the call center is often the most important way to communicate with customers. Processes in call centers have (often) become more complex (Lin, Lai and Hung, 1998) and available new technologies offer new possibilities. To take advantage and benefit from these new developments a business and managerial approach is a necessity. That is why companies start looking at call centers as profit centers. Because of this newly adopted approach, techniques are developed to manage the call center processes at higher levels (strategic and tactical). Integration with back-end processes and (information) systems within the company becomes necessary.

Research has resulted in many techniques to support decisions to be made in the (design of) call center processes where a balance has to be found among different objectives. Generally four main dimensions are distinguished in the effects of (re)design measures:

- time,
- cost,
- quality and
- flexibility.

The trade-off that has to be made between the different dimensions is often difficult. Brand and Van der Kolk (1995) clarify this with the devil's quadrangle (Appendix A).

1.1 Defining research area

To categorize the call center, the characteristics mentioned in Stolletz (2003) are used.

- The functionality of the call center is offering help desk services;
- the initiation of contact has an inbound character;
- the size of the help desk is small (maximum of 15 - 40);
- the geography feature is centralized;
- the communication channels that are mainly used are telephone and mail.

Help desk services and inbound calls are the most important features to keep in mind. Furthermore the calls have multiple types and therefore need multiple types of resources; generalists and specialists. The multiple communication channels (telephone and mail) mean that a call center process can also be seen as a contact center. According to Dawson (2004):

"Contact centers handle more than the traditional voice call. These would be centers that handle voice plus fax, or email, live Web chat centers, video interactions – all the many real and hypothetical customer interactions that are now possible".

Scientific literature identifies different activities in the (re)design process for staffing call centers. Grossman et al. (2001) mention:

1. forecasting,
2. performance estimation,
3. staff requirements,
4. shift scheduling and
5. rostering.

Stolletz (2003) and Stolletz and Helber (2004) mention the same activities for (re)design but leave out performance estimation. All authors state that each stage is a research area itself and that the planning process has a sequential and hierarchical structure. In particular the term forecasting needs some explanation. By forecasting, the different authors mean the estimation of arrival rates and service times of the different types of cases/calls that arrive at an inbound call center. In this survey, we will focus on the phases of forecasting, staff requirements planning and performance estimation. Shift scheduling and rostering are considered to be other research areas and therefore out of the scope of this literature review. They deal with satisfying employee and organizational requirements and preventing under-utilization of resources. The other three phases are more aimed at long term planning of (overall) staffing levels.

1.2 Research question

The objective of this report will be to "*find techniques for determining the optimal staffing levels in a multiple skill inbound call center* "

The result of the literature review/this report will be an overview of available techniques to determine optimal staffing levels from a strategic and tactical point of view, including the techniques' advantages and disadvantages. Optimality is reached when (a mix of) objectives and performance measures (is) are satisfied. The possible/existing criteria for measuring objectives and performance will also be subject of the literature review. It will not be a summary of methods of performance evaluation, but a search for techniques that support finding an optimal design.

Terminology throughout scientific literature appears not to be consistent. Especially the terms used for different types of resources are widespread and will be defined in the upcoming chapter. Koole, Gans, Mandelbaum (2004) and Koole and Mandelbaum (2002) mention two techniques for call center analysis: to model the processes and to simulate them. According to these authors the two types of analyses can not be seen as separately independent ways of support for call center process (re)design, but should be used in combination to obtain optimal, fine-tuned results. In this business reengineering process they will both be used (for slightly different goals) to obtain the optimal mix.

Scientific literature about call centers has to be up-to-date, because of the rather fast developments and changes made in the area of customer service and call center management (Koole and Mandelbaum, 2002). Books usually provide a general and managerial approach and not the latest techniques for solving complex problems. An effective overview of existing literature and research areas can be found on a webpage, containing an up-to-date research bibliography (Mandelbaum, 2004).

1.3 Methods and Approach

The remainder of this literature review is organized as follows. In chapter 2 an explanation will be given on what the general structure of a call center is; the way call centers are usually represented in scientific literature and the common terms and techniques used in call centers. Chapter 3 deals with the different existing approaches towards (re)designing call center processes. A number of authors and their approaches will be subject of discussion. The scope of (re)design will be to determine optimal staffing levels. Three relevant activities in the (re)design process of call center staffing (performance estimation, forecasting and determining staffing requirements) will be discussed in the subsequent chapters. In chapter 4 the different means of measuring performance of call center processes will be discussed by mentioning different authors and their approaches towards performance measurement. Chapter 5 will describe forecasting techniques to determine/estimate input parameters (arrival rates, service times and abandonment behavior) for modeling. Chapter 6 will discuss approaches towards determining optimal staffing levels. Queueing theory and simulation models will be an important subject of chapter 6, next to mathematical tools to solve minimization problems.

The most important findings in chapters 2 to 6 will be collected in a table overview (appendix K). This table will represent some of the most relevant characteristics of the identified different approaches towards determining optimal staffing levels for a multiple skill inbound call center.

In the final section, chapter 7, conclusions and reflections on the literature review will be discussed, together with limitations, future research and learning experiences.

2. Call center representation in scientific literature

Because of differences in research backgrounds, authors, countries, cultures, languages, time periods and research objectives, different terminologies in and approaches towards describing, analyzing and (re)designing call center processes (or service processes in general) exist. For this reason the different terms in and representation of call center processes that appear in scientific literature will be summed up and described in this chapter. First (section 2.1) the terminology for all kinds of process parts will be dealt with. Then, in section 2.2, a summary of characteristics will be provided that different authors mention in their articles. In the area of call center process representation (in models and figures) a wide range of different forms of representation exists. Some useful and practical representations of different structures and models will be given (section 2.3).

2.1 Terminologies in Call Center processes

2.1.1 ACD and Human Resources

When a customer calls for a service he or she usually enters an Automated Call Distributor (ACD) that is at the very beginning of the call center (process). The ACD is part of the telephony-switch infrastructure (typically hardware-, but recently more software-based) and routes calls to agents, while tracing and capturing the history of each call. The call center or contact center thus basically performs two activities (Zapf, 2004) to handle incoming customer or service requests:

(1) classification of the incoming request (by the ACD or an agent) and if necessary forward it to a qualified employee/agent. The request may have a standard or a special nature. The request volume is partitioned because of specialization and communication reasons.
(2) handling the request by providing required information or by performing necessary actions

The term agent has already been used for personnel occupying the call center in various positions. In scientific literature a wide range of terms for ‘agents’ comes forward as was already mentioned in the introduction. Zapf (2004) and Reijers and Limam (2005) for example use the terms generalists and specialists (figure 1) as a distinction between different classes of agents in a call center.

Figure 1: specialist vs. generalist

Generalists and specialists are preferred in this literature review. Coherence with other typologies will be explained when necessary. For example flexible resources and single skill servers have the same meaning as generalists and specialists.

- Stolletz (2003) defines a generalist as an agent that can handle all types of customers, and a specialist as an agent that can handle only one type of customer.
- Pinker and Shumsky (2000) prefer to use the terms flexible and specialized workers for generalists and specialists,
- whereas Hasija et al. (2005) and Shumsky and Pinker (2003) use gatekeepers and experts for generalists and specialists.
- A bit different from these terms are seniors and juniors (Lin, Lai and Hung, 1998) and have to more to do with differences in experience and responsibility than with different skill levels.
- Blend agents are mentioned by Pichitlamken et al. (2003) and next to outbound calls they also handle overflow from inbound (when waiting), in the case of occupied inbound only agents.
- Mazzuchi and Wallace (2004) do not make the distinction between generalists and specialists, but set up a skill matrix for every individual agent. The matrix contains the skills agents have or have not. The assumption made by the authors that service times would not depend on server experience is a rather unrealistic one. This type of agent seems to be a generalist with some specialist skills.

As already stated in the introduction, the terms generalists and specialists are preferred in the remainder of this literature report. Definitions or qualifications of generalists and specialists are not unambiguous in scientific literature. Every model or author uses a specific qualification for generalists and specialists and usually the two dimensions to which the agents are defined are:

- Type/number of skills an agent has
- The level at which the skills can be performed

In general terms (provided by Zapf, 2004) one would define generalists and specialists as:

Generalists usually handle the standard requests dealing with the processing of simple transactions, the modification of customer data or general enterprise or product information; actions where basic knowledge is required.

Specialists with more specific, in-depth knowledge or special skills, deal with the more difficult requests that refer to technical problems, extensive consultations or complaints.

2.1.2 Important parameters

Some other terminologies that show up in almost every article (for example Koole and Mandelbaum, 2002 and Gans et al., 2003) are λ, lambda and μ, mu which stand for respectively the arrival rate of service requests and the handling time of a service request by an agent (specialist or generalist). These are important process parameters that influence the occupancy of the agents and the speed of processing. The number of servers (= agents, generalists and specialists) also is part of the performance of the call center process because of cost aspects.

In the ideal situation all customers that enter a call center with a service request are being served. Naturally this is not the case. The telephone system might be overloaded because the trunk capacity is too low. Mazzuchi and Wallace (2004) state that the trunk's capacity is C + K, where C is the number of present agents in the call center and K is the number of waiting spaces or buffers to hold waiting callers. If the trunk is full, new callers are blocked. If callers/customers are set on hold or have too wait too long they become impatient and hang up. This is called abandonment in almost all scientific literature. On occasion the terms balking (Stolletz, 2003) and reneging are used. Some customers try to get in again, this is called retrial and/or jockeying (calling many times). The behavior of abandoning customers (because of impatience) is difficult to model. This subject will be described in chapter 5.

2.1.3 Time

Planning staff on a short term basis is out of the scope of this research since the goal is to find the optimal staffing levels over some period of time (long term planning). In call center business rostering and scheduling is typically performed on a daily and weekly basis, with maybe a maximum of four weeks ahead (Henderson and Mason, 2004).

Fact is that call centers have to deal with (overlapping) shifts and that rosters maybe overstaffed in certain time intervals and understaffed in other time intervals (Ernst et al., 2004) in order to obtain good, low-cost rosters that cover the workforce requirements adequately. Timeblocks (15, 30 or 60 minutes) are usually used to decide how many employees are needed for a (particular part of the) day. The ACD also records the call center data per timeblock, for standard reports on the hourly, daily, weekly and monthly performance.

Another important restriction for determining staffing requirements is the fact that an employee is not available for 100%, during its attendance. Mehrota and Fama (2003) use a shrinkage factor (for example 10%) to take into account a certain amount of agent time that will be lost, either in large blocks (unanticipated shift cancellations, partial day absences for personal reasons) or in small blocks (late arrivals to the call center, extra-long breaks, trips to the bathroom).

2.2 Characteristics in call center processes

2.2.1 Categorization of call centers

A call center can be characterized by all kinds of characteristics. Depending on the objectives, authors use a specific set of characteristics to describe a call center. Stolletz (2003) uses four characteristics to describe the queueing model (chapter 6) he uses for modeling call center processes:

- customer profile: arrival process per calling customer type and the patience of a particular customer class are described.
- agent characteristic: two dimensions, the general qualification to handle a specific type of call (skill) and the distribution of the service time (skill level).
- routing policy: which agent serves which customer next. These decisions may depend on the state of the system (number of waiting customers of different classes or the number of busy agents).
- the size of the waiting rooms: defines the maximum number of customers in the system.

Another approach is mentioned by Koole and Mandelbaum (2001) who categorize along the following dimensions:

- functionality (help desk, emergency, telemarketing, information providers, etc.),
- size (from a few to several thousands of agent seats),
- geography (single- vs. multi-location),
- agents charateristics (low-skilled vs. highly-trained, single- vs. multiskilled), and
- whether a call center handles inbound or outbound calls.

Other important characteristics are the type of calls (single or multiple) and whether a call center is multi-layered or not (with or without a backoffice of specialists).

Zapf (2004) distinguishes two important characteristics to describe call center processes:

- the level of difficulty of requests (standard vs. special requests, already described before)
- the communication channel (synchronous vs. a-synchronous).

Synchronous communication takes place if customer and agent are communicating with each other at the same time (e.g., phone or chat). E-mail and fax are examples for a-synchronous communication channels, where customer and agent do not need to get in contact at the same time and longer time intervals pass by between single communication steps. Assigning service requests from a particular channel to a suitable agent will be dealt with later on in this literature report. In Appendix B, table 1 an overview is given of possible tasks of generalists and specialists in various call center configurations (two- or one-level and back-office).

An interesting issue that Zapf brings up is the fact that often an agent has to call back to a customer to complete the service, a characteristic which is often ignored by other authors when modeling a call center.

Henderson and Mason (1998) use characteristics to describe call centers that are mentioned before and they also provide factors that can complicate the rather simple descriptions. For example, varying arrival rates of calls and abandoning customers, customers can have higher or lower priorities, multiple types of calls (and therefore need to be assigned (if possible) to a particular subset of agents) and the number of available agents may vary through the day. Varying arrival rates may be caused by unpredictable factors but also by predictable factors and triggers (Gans et al., 2003). At certain times external factors (e.g. seasonal or promotional activities) can influence the pattern of arrivals.

Another important aspect that was found in scientific literature (Gans et al., 2003, Mehrota 1997) about call centers is the fact that human resources often account for 50%-75% of the operating expenses of a call center, which means human resources act as a bottleneck during call center process (re)design. Some discussions are found concerning shared resources (e.g. information system or database), acting as a bottleneck in a call center (Akşin and Harker, 2003). Since only few authors backup this thesis and use it for a certain type of call center, the shared resource is generally not seen as bottleneck.

2.2.2 Skilled-Based-Routing (SBR)

The routing policy that is mentioned by Koole and Mandelbaum (2002) is the so called Skill-Based-Routing (SBR), which differs for example from the First-Come-First-Served (FCFS) strategy. SBR can be part of an advanced ACD (also according to Mazzuchi and Wallace, 2004) and can be seen as the strategy to match callers and agents. Especially when call centers have multiple types of customers and multiple types of tasks to perform. A common way of implementing skill-based routing is by specifying two selection rules:

1. agent selection - how does an arriving call select an idle agent, if there is one;
2. call selection - how does an idle agent select a waiting call, if there is one.

Gans et al. (2003) use the same description for the SBR routing policy. Every agent has an individual subset of skills, so each skill has a group of suitable agents. The authors also provide a nice example of SBR in appendix C, figure 19. Becker et al (1999) also talk about SBR as the routing policy for routing service requests. In Gans et al. (2003) and Garnett and Mandelbaum (2001) a series of canonical designs (appendix C, figure 18) have been provided to give an overview of possible routings of calls through a call center.

Mazzuchi and Wallace (2004) provide a practical policy for SBR by using the LIAR policy for arriving service requests. The LIAR policy stands for Longest-Idle-Agent-Routing and sends calls to the agents that have been waiting the longest for a call since the completion of their last job (i.e., idle the longest). To adjust for priorities, the LIAR policy sends calls to the agents that have been waiting the longest (or idle the longest) and have the highest skill-level to handle the call. When an agent becomes free and if there are no customers in the agent's primary skill queue, the first customer in the agents secondary skill queue is taken.

2.3 Representations of call center processes

Some examples of how to schematically represent the (general) process in a call center will be provided in this section. A very basic representation (figure 2) can be found in Koole and Mandelbaum (2002) and in Gans et al. (2003) and it describes the relationships among the main terms in call center processes. Calls come in at a certain arrival rate (arrivals) and are lost, queued, lost because of longtime queueing or re-queued (retrial). Calls can thus either be lost or solved by an agent. After a call has been solved or dealt with, a customer can return (redial) with a call related to the same problem.

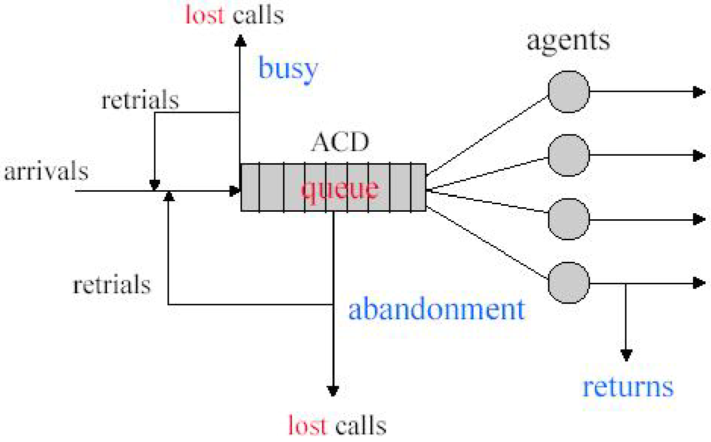


Figure 2: simple representation of a call center process

Gans et al. (2003) also give a more technical approach towards the representation of a call center (figure 3). PSTN stands for ‘public switched telephone network’ and it leads callers through the trunk lines to the PABX (private automatic branche exchange, a private switch) of the company’s call center. The PABX leads calls to either the IVR/VRU (interactive voice response / voice response unit) unit or the ACD. The IVR unit or VRU is a specialized computer that allows customers to communicate their needs and to “self-serve.” Customers interacting with an IVR use their telephone key pads or voices to provide information. CTI (computer telephone integration) server and the customer data server are used to more closely integrate the telephone and information systems (e.g. Customer Relationship Management, CRM). With help of IVR/VRU, ACD and servers, calls are led to the right agent.

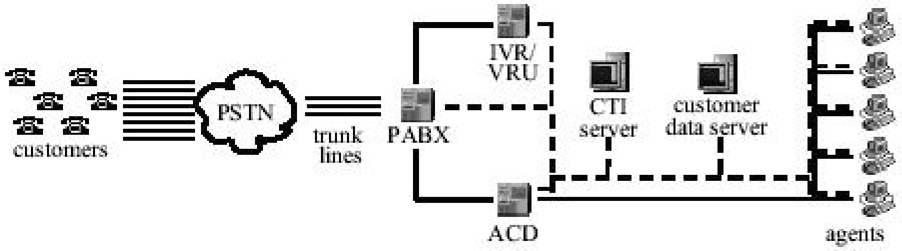


Figure 3: schematic technology diagram of a call center

In the terminology (gatekeepers and referrals) of Hasija et al. (2005) and Shumsky and Pinker (2003) the representation of a call center process looks like figure 4. Again a very basic view, on the principles of gatekeepers and referrals. According to the authors the performance of a call center and the behavior of an agent (gatekeeper) mainly depend on the prescribed referral rate (the rate at which work is routed from generalists to specialists). If a call is too difficult to be handled and solved by the gatekeeper, he or she refers the call to a specialist. The authors call it a triaging system, because customers first interact with a generalist who determines if the customer requires the attention of a specialist or not.

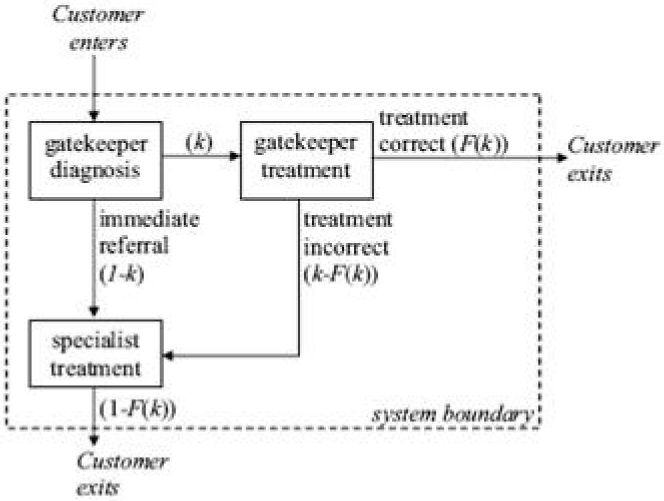


Figure 4: gatekeeper and referral configuration of a call center

Zapf (2004) uses a kind of Petri-net way of representing the process in a call center (figure 5). He differs from Hasija et al. (2005) and Shumsky and Pinker (2003) because, as mentioned in section 2.2.1, he distinguishes between two types of request (standard and special) which are handled both in a different way. The difficulty of a service request is known upfront, whereas Hasija et al. (2005) and Shumsky and Pinker (2003) assume an agent (gatekeeper) can first try to successfully treat a service request.

Standard requests are classified and handled by generalists and special requests are first classified by generalist and then handled by specialists. The Petri-net principles are used, which means that a call can only be classified (and possibly handled) when a requests arrives and a generalist is free.

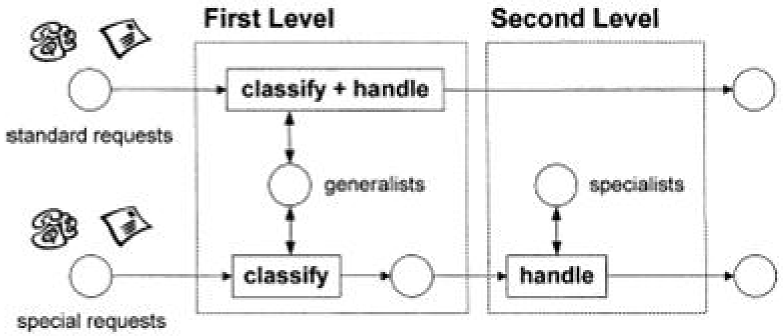


Figure 5: Two-level design of a call center process

In Appendix B, figures 12 to 15, more types of configurations of a call center process can be found. A short description is provided with the figures. Also very interesting process models of a call center with out-calls and with waiting tolerance of customers with synchronous requests are provided by Zapf (2004). The model can also be found in Appendix B (figure 12).

3 (Re)design of Call Center staffing

Section 1.1 about defining the research already mentioned redesign approaches from a number of different authors (Grossman et al. 2001, Stolletz and Helber 2004 and Stolletz 2003). The five general, distinct activities in staffing process that are distinguished by Grossman et al. (2001) are:

1. forecasting,
2. performance estimation,
3. staff requirements,
4. shift scheduling, and
5. rostering.

Each of these activities is a research area itself. Of course details may vary at different call centers. As explained in section 1.1, the latter two activities our outside of the scope.

Usually the redesign of an existing call center process is subject of discussion in scientific literature. The use of historical and real data is namely very important in case of re-engineering processes that are driven by stochastic factors. Koole and Mandelbaum (2002) support this by stating that the use of forecast and staffing models must be based on analytical models and real data if it wants to have any practical value.

Management decides whether or not a redesign is necessary. A redesign can be necessary because of lacking performance (for example because of a bottleneck) of the existing call center, changing objectives of a call center or minimizing costs. The relationship of redesign with performance measures is close. Conclusions from performance measurement are used to change the structure, staffing or procedures of a call center. A redesign can be directly implemented or first tested in a model; of course with corresponding performance measures.

In this chapter a number of authors and their view on (re)designing call center staffing will be listed (section 3.1 to 3.8). At the end of the chapter a short explanation for the content of the subsequent chapters will be provided (section 3.9).

3.1 Mehrota (1997) and Mehrota and Fama (2003) and (re)design

Mehrota (1997) provides 4 fundamental questions for the (re)design of a call center:

- How many calls will we get?
- How many people do we need on staff?
- When/how should these agents be hired, trained and scheduled?
- What will this cost?

To make sure all these questions are answered correctly for mid- and long-term planning, a tool, generally called Workforce Management (WFM) software is available. WFM is particularly suitable for day-to-day operations (Gans et al., 2003). WFM uses other tools to forecast calls (for a particular time period), to determine the required number of agents in seats and to assign agents to schedules. This last tool is outside the scope of the research question in this report. Mehrota (1997) slightly ignores the importance of a more strategic approach which would deal with subjects like types of agents, routing policies and quality of service. A better job was done by Mehrota and Fama (2003) who again use the dimensions from figure 6 (chapter 4) to base their decision support model on. The trade-off is again between costs, service quality and employee satisfaction. The dilemma call center managers should deal with is reflected by a list of important (and more specific) questions:

- How many agents should we have on staff with which particular skills? How should we schedule these agents' shifts, breaks, lunches, training, meetings and other activities?
- How many calls of which type do we expect at which times?
- How quickly do we want to respond to each type of inbound call?
- How should we cross-train our agents? How should we route our calls to make the best use of these resources?
- Given a forecast, a routing design, and an agent schedule, how well will our system perform?
- What is our overall capacity? How will a spike in call volumes impact our overall performance?
- How is our center doing right now? What has changed since we did our last forecast and published our schedules? If the changes are significant what can I do to respond to minimize the impact on the rest of the day or week?

This provides a much wider range of areas on which managers of call centers have to make decisions. To solve these problems, mathematical models, workload forecast models and simulation of different possible design solutions are used to support decisions.

3.2 Gans et al. (2003) and redesign

Gans et al. (2003) speak of capacity management on different hierarchical levels. At every level an analytical model supports decision-making:

1. Queueing performance models (section 6.1) for low-level staffing decisions,
2. mathematical programming models (section 6.3) for intermediate-level personnel scheduling,
3. and long-term planning models for hiring and training.

The primary interest for this literature investigating report is the lowest level: queueing performance. At the intermediate-level it might be interesting to see (for a week or a month) how many agents should generally be available. To have input for the model at the lowest level, correct forecasting models and estimation procedures are very important. An interesting remark the authors made at the highest level is the fact that employee turnover can be significant at certain call centers, which possibly causes problems with hiring and training.

Another important design consideration management faces, is the choice for an effective routing policy. The authors state that dynamic programming (DP) is a way to solve this problem, although it is a rather impractical one. In Hillier and Lieberman (1995) DP is described as a very useful technique for making a sequence of interrelated decisions. It requires formulating an appropriate recursive relationship for each individual problem. Especially time-consuming and difficult when the subject of research or redesign is a complex and large call center with many types of calls. This would cause an explosion of the dimensionality of the state space. Though with modern computer technology such enumerations should not take too much time. For small call centers DP could be a good solution, but for more complex call centers a model for reducing complexity (Gans et al., 2003) is necessary. This leads to:

1. typology simplification by using classical canonical designs for SBR which was shown before in section 2.2 and can also be found in appendix C, figures 18 and 19.
2. control simplification, which uses for example fixed, static priority policies.
3. asymptotic analysis, which is meant for heavy traffic call centers. Two asymptotic regimes have been considered by the authors:
    a. Efficiency Driven regime, which turns out to be inappropriate for inbound call centers with heavy traffic. Too much delay in the processing of different types of calls is allowed because of efficiency objectives.
    b. Quality-Efficiency Driven regime has the more qualitative goal to find a routing of calls where a significant fraction of the customers find idle servers upon arrival. The QED complexity stems from the absence of complete resource pooling (section 6.4) and the fact that the agent-selection problem plays an important role.

In the area of QED routing policies, a lot of research is being done at the moment on square-root laws and V-design (see also appendix C, figure18). Given a V-design, the QED regime is straightforward to characterize as simply maintaining square-root safety staffing. Hasija et al. (2005) also use the square-root staffing rule and heuristics to determine optimal staffing for both tiers, given any particular referral rate in their gatekeeper-specialist system. The rules and heuristics provide quick solving possibilities and characterization of effects (sensitivity) of certain parameters on the optimal solution. Furthermore it allows for direct comparison between one- and two-tier systems and thus comparison between extreme systems (only generalists or only specialists) and mixes of different types of agents. The square-root staffing rule and heuristics will be dealt with more specifically in section 6.3.2.

3.3 Principal agent model

Gans et al. (2003) shortly mention future directions of multidisciplinary research in (re)designing call center processes. One of their propositions is the use of a model from microeconomics to provide insight in the possible outcomes of proposed system designs: the principal agent model. Hasija et al. (2005) and Shumsky and Pinker (2003) base their model on the assumptions and tools of the principal-agent model. They define it as the difference in preferences between gatekeeper and principal and there may be information asymmetry between the principal and the agent (in our case the gatekeeper - but not the firm - may see the details of each customer's problem and the suitability of the gatekeeper's skills for that customer). The agent faces a decision that is, in some ways, more general than the standard "effort-level" decision. Here, the gatekeeper may or may not prefer to put in effort (treat customers), and the firm's profit is not monotone in the treatment rate. A gatekeeper who provides too little or too much treatment may significantly reduce the firm's profits.

3.4 Zapf (2004) and redesign

Another approach towards (re)designing call center processes can be derived from Zapf (2004) who defines 3 design dimensions:

- Task allocation to generalists and/or specialists
- Front-office and back-office roles
- Degree of integration of synchronous and a-synchronous requests

In figure 5 and appendix B (figures 12 to 15) several possible configurations of call center processes can be found, designed along the three dimensions mentioned above. The modeling complexity of these configurations can be quite high because of the use of queueing theory and linear programming, LP (section 6.3). Only simple designs have been analyzed under strong restrictions. Zapf therefore uses discrete event simulation, which overcomes the restrictions and therefore process designs are more close to reality.

3.5 Ernst et al. (2004) and redesign

Ernst et al. (2004) use mathematical models and algorithms to build a rostering tool. The process to develop the rostering tool is interesting, since it shows quite an overlap with identified activities in other design processes. The authors identify three activities to develop the rostering tool:

- A demand modeling study that collects and uses historical data to forecast demand for services and converts these to the staffing levels needed to satisfy service standards
- Consideration of the solution techniques required for a personnel scheduling tool that satisfies the constraints arising from workplace regulations while best meeting a range of objectives including coverage of staff demand, minimum cost and maximum employee satisfaction
- Specification of a reporting tool that displays solutions and provides performance reports

Actually only the first and part of the last activity are interesting for this literature research.

Furthermore the authors state that queuing models are elegant and may give analytical results but generally many real world simplifications need to be made. Simulation can take many practical factors into account but these may be very computationally expensive solutions. Sometimes, queueing models and simulation are combined to obtain ideal staff requirements. Other authors mention this option as well. For example Mehrota (1997) and Grossman et al. (1999). Pichitlamken et al. (2003) describe it as follows: the simulation model is used to perform what-if scenarios, because of high flexibility. The queueing model (CTMC, Markov) is used to approximate the system performance measures. CTMC models are insightful, relatively easier to construct.

3.6 Henderson and Mason (1998) and redesign

Two phases or activities in the rostering problem are identified by Henderson and Mason (1998) and also by Mehrota (1997). First staffing requirements for each period of the day are determined using queueing models and/or simulation. Three approaches to complete this first phase exist:

1. Steady-state queueing models provide excellent approximations to the number of service agents required. Unfortunately, fast convergence to steady-state is not typical for these systems.
2. Attempt to numerically calculate, or approximate, the time varying distribution of GOS (overall Grade of Service).
3. Use simulation to obtain a required number of agents in each period of the day.

In the second phase, one attempts to build staff rosters that cover these staffing requirements using integer programming formulations of set covering problems (out of scope). This two phase procedure is an improvement over heuristic rostering, although it is not entirely satisfactory. The linking between adjacent timeblocks is missing. A solution can be iteration of the design process until convergence (RIIPS: Rostering by Iterating Integer Programming and Simulation).

3.7 Chan (2003) and redesign

Based on the identified Key Output Performance Variables (KOPV) and the Key Input Performance Variables (KIPV) by Chan (2003) a number of activities can be performed to design an effective workflow for call centers using simulation tools. Simulation enables dynamic analysis, which is necessary to analyze processes. A list of steps in the design process is provided in appendix G.

3.8 Lin, Lai and Hung (1998) and overlapping schedules and shifts

Overlapping staffing schedules and shifts are subject in a lot of scientific literature. Many call centres have to deal with that subject too, although not all in such an extreme manner as mentioned in Lin, Lai and Hung (1998). Their case deals with a 24H hotline service, while a particular call center may only have a few hours of overlapping (employee) schedules. The authors use an integrated approach, on a monthly basis, in which one of the final stages towards scheduling is the activity that implies the use of historical data, the choice for call center configuration and other information (appendix F, figure 21). A regression model leads to a simulation model, with outcomes that can be used to make decisions for rostering and scheduling (heuristics).

3.9 In the upcoming chapters

As mentioned before, only three of five (re)design phases are within the scope of this literature review. In the following chapters the three relevant phases will be subject of discussion:

- Chapter 4: Performance measurements in Call Centers
- Chapter 5: Forecasting input parameters for modeling Call Centers
- Chapter 6: Approaches to determine (optimal) staffing levels

The findings from the different chapters will be summarized, by author (and thus by approach), in a table overview which can be found in appendix K (table 5). In appendix K a summary/listing is provided from the results of chapters 2 to 6. A number of the authors presenting a (re)design approach towards determining staffing levels for call centers will be listed. For these authors the table represents the authors' name, the characteristics of the (re)design approach, the typology of the human resources occupying the call center, the type of measurements and the special characteristics (compared to the other approaches and authors).

4. Performance measurements in Call Centers

One of the activities in the process of staffing call centers is to determine how performance of a call center is measured. Depending on the main objectives of a call center, one might use a specific criterion or a mix of criteria to measure a call center's performance. Koole and Mandelbaum (2002) state that usually the service level must hold for every time interval (timeblock), while Koole and van der Sluis (2003) emphasize the advantages of looking at an overall service level. For deeper needs (e.g. statistical analysis) performance should be observed at high resolution, so the appropriate frequency would be every 30 or 60 minutes. For overall performance of a call center it is allowed to use averaging over a series of time intervals.

Whether or not a call center performs well, mainly depends on the way it is measured and on which criteria the judgment is based. The choice for a particular performance measure is a fundamental trade-off (Koole and Mandelbaum, 2002) within an organization.

- Measuring the operational service level is typically done with performance measures focused on abandonment, waiting and/or retrials; useful in an efficiency-driven environment.
- At the other extreme one finds the quality-driven environment, in which for example the utilization of agents is measured.
- Usually quality and efficiency are balanced in a "rationalized" game.

A possible mix of performance measurements (Koole and Mandelbaum, 2002) could be abandonment, average speed of answer (ASA), average handling time (AHT, service duration) and agent utilization.

Reijers and Limam (2005) state that usually 4 main dimensions are distinguished in the effects of (re)design measures: time, cost, quality and flexibility. The trade-off that has to be made is often difficult. Brand and Van der Kolk (1995) state this with their view on the problem: the devil's quadrangle (Appendix A). In call center terminology time would be average waiting and service time, cost would be the cost for agents on the phone, quality and flexibility would be related to the number of generalists and specialists and their (qualified) skills. Sometimes not all dimensions are relevant. This depends on the situation and the circumstances. An example of competing effects is provided in figure 7. In a general situation with adding more specialists, the costs of skilled agents and customer's time in the system will roughly remain the same, the flexibility will decrease (less generalists) and the quality of service will increase.

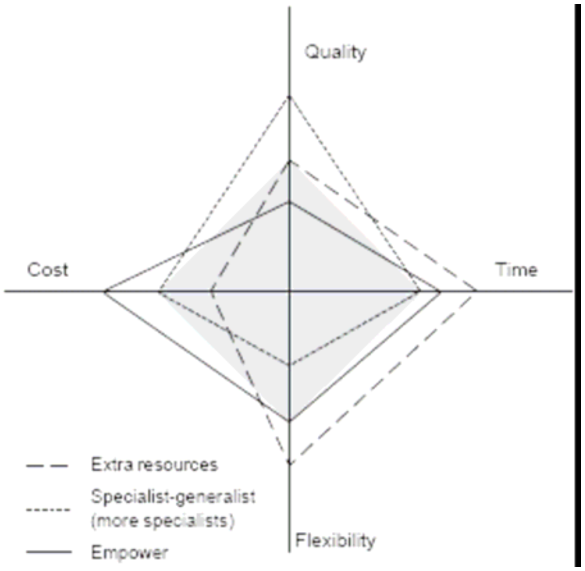


Figure 6: effects of more specialists

In this chapter a number of known and widely used performance measurements will be listed by author, from section 4.1 up to section 4.8.

4.1 According to Gans et al. (2003)

According to Gans et al. (2003) there are three commonly used views on quality in call center business:

- Accessibility (measured with waiting times, abandonment, etc.)
- Effectiveness (for example measured by: solved the problem or not?)
- Content of interactions (measured by: listening to specific calls, customer perception surveys)

Daily practice usually deals with the operational side of service quality. With WFM the central dilemma is the utilization rate of the agents. Higher rates mean longer waiting times in queues and thus lower accessibility. Call center goals are formulated as the provision of a given level of accessibility, subject to a specified budget constraint. Common practice is that upper management decides on the desired service level and then call center managers are called on to defend their budget.

System performance can be measured by queueing models which are based on given assumptions on primitives (arrivals, service times and number of agents) and the relationships among them. Furthermore the desired service level has to be based on customer patience, which can be different for a number of types of customers. Customers with Internet questions may be more patient than customers with telephone questions. Some authors even think of two dimensions of customer patience and distinguish between time willing and expecting to wait. A Patience Index can be derived from expectations of the both dimensions (Gans et al., 2003) and is used to evaluate the real behavior of customers.

Most performance measures are correlated. For example, the average waiting time is linearly related to the fraction of abandoning customers. This implies that only one statistic has to be measured and the other one can be derived through interference. Research by Brown et al. (2002) shows peaking of arrival rates, service times and delay in queues all tends to happen at the same time. Analysis reveals the fact that peak hours are the convenient hours for customers with longer service times to call.

4.2 According to Pinker and Shumsky (2000)

Pinker and Shumsky (2000) mention Quality of Service as the main measure for system performance. Quality of Service (QoS) depends mainly on the fraction of customers served and therefore the QoS is directly related to the revenues of the system. Queues and waiting times are ignored, which is rather a doubtful proceeding, because they are not being ignored by other researchers and authors. Zapf (2004) for example states that the basic knowledge to understand the efficiency of process designs comes from queueing theory. Pinker and Shumsky (2000) though state that this structure captures the key relationships between service standards, labor costs, and server utilization while being more amenable to analysis than systems with queueing. The service standards are determined exogenously and appear as constraints in the model (average waiting time, fraction of customers served).

4.3 According to Mehrota (1997)

According to Mehrota (1997) the key performance metric for call center managers is service level, which is defined as the percentage of customers who wait less than some target before reaching an agent. Looking from a queueing theory perspective the service level is more a function of several variables, including call arrivals, call handling time and the number of agents on staff. Faced with tight budgets the usual trade-off is between service level and costs. Mehrota and Fama (2003) add an extra dimension to this trade-off with employee satisfaction (figure 7). Due to the increased importance of call centers and its service delivery, the authors recognize the importance of the wellbeing of call center agents.

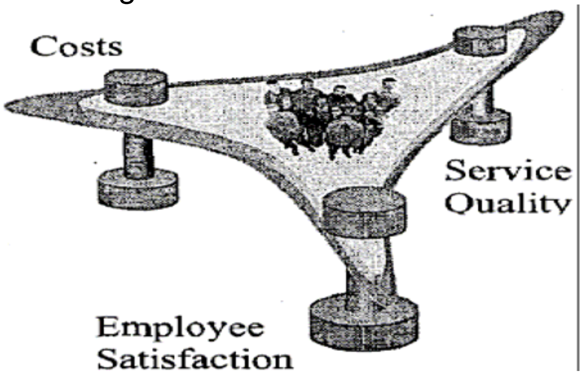


Figure 7: the call center balancing act

4.4 According to Chan (2003)

A very operational and quantitative approach towards call center workflow and performance is provided by Chan (2003). Chan defines Key Performance Input and Output variables: KPIV and KPOV. Main performance measure for the output of a call center according to Chan is the effectiveness and its KPOVs are:

- The number of processed calls
- Throughput time
- Waiting time
- The number of reneging customers
- The cost of processing a single call

Chan also identifies several possible KPIVs that affect the KPOVs:

- The arrival rate. An increase will increase the load of the call center, which will result in longer waiting times for each call.
- The number of resources available. This affects costs, capacity, average waiting time and throughput time of the call center.
- The number of research call. Time to collect more information on a problem.
- The process time of each call. The time required to complete each call.
- Time to renege: allowable time a caller will tolerate before hanging up.

4.5 According to Henderson and Mason (1998)

Henderson and Mason (1998) also have a rather one-sided approach towards quality of customer's service, which they define in terms of a measure called customer grade of service (CGOS). The CGOS typically depends on the customer's waiting time in queue. By averaging one may obtain an overall grade of service (GOS). The authors developed a model using utility curves, reflecting the effect of different waiting times in the queue on the customer's satisfaction. These utility curves can vary from customer to customer. A single utility curve was chosen the authors believed to be representative, and they attempted to minimize some statistic associated with customer utilities. The utility curve is called a CGOS. Customers receive a CGOS corresponding to their waiting times. This way various customer service requirements may be specified. For example:

- CGOS should exceed 50 for all customers.
- 95% of customers must receive a CGOS > 80.
- In any 2 hour window, the average customer CGOS should exceed 80.
- During peak times average CGOS should exceed 50, otherwise it should exceed 80.
- The expected CGOS for a customer arriving at any time throughout the day should exceed 80.

4.6 According to Akşin and Harker (2003)

Optimality is defined in the same economic sense as in Andrews and Parsons (1993). This allows one to capture the characteristic that revenues are a direct function of staffing decisions, and is different from the prevailing approach of minimizing costs. This type of objective function is the appropriate one for service systems that are profit centers of some sort. Since measuring quality is extremely difficult, the authors want to determine economically optimal staffing levels, defined as those levels that maximize total revenues net of staffing costs for the service system. These economically optimal staffing levels can be determined for a loss system. The authors state that revenue is generated by serving a customer. Each time a customer is lost, the system incurs a revenue loss. Thus, in order to relate staffing decisions to revenues, one needs to characterize the customer loss as a function of the number of servers.

4.7 According to Mazzuchi and Wallace (2004)

Mazzuchi and Wallace (2004) provide a list with some commonly used performance metrics.

- the probability that an arriving caller is blocked
- speed-to-answer performance measures
- tracking agent's utilization

The first two performance metrics are usually included in the Service Level Agreement. The last performance metric normally contains several sub-measurements. In appendix E some more detailed information is provided about the specific formulas of the performance metrics.

4.8 According to Lin, Lai and Hung (1998)

Lin, Lai and Hung (1998) use a very simple performance measure. By using the ACD report (Appendix F) the system performance is measured by means of the call abandonment rate, defined by (aband calls)/(ACD calls + aband calls) , where aband calls and ACD calls denote the numbers of abandoned calls (after entering the queue) and completed calls, respectively.

5 Forecasting input parameters for modeling call centers

In section 1.1 and chapter 3 forecasting has already been mentioned as part of the (re)design activities of call center processes. Since the objective of many approaches towards (re)design is to first build a model of the existing call center process, determining input parameters for the model is a very important part of the (re)design process. Determination is mainly done with forecasting techniques and with making assumptions. Sometimes assumptions can be made about certain events that will happen in the future, like upcoming shipments, advertisements or in other words: predictable events.
In this chapter we will describe forecasting and the different approaches towards forecasting and the available techniques to do so. By forecasting, the different authors of scientific literature mean the estimation of arrival rates and service times of the different types of cases/calls that arrive at an inbound call center.

Forecasting calls is typically driven by a combination of historical data, time series models and expert judgment (Mehrota and Fama, 2003) to determine call volumes and average handling time. Forecasts should be determined for each queue for each time interval in the simulation period. Mehrota (1997) notes the importance of abandonment and calling back by stating that along with traditional forecasting issues such as data availability, data integrity, seasonality and non-stationary randomness, abandonment makes call forecasting more challenging than simply fitting a regression model to historical call volumes, particularly since there is usually no way to tell if an abandoned call led to a call back later on.

In the remainder of this chapter a number of authors and the forecasting techniques they deal with will be described (section 5.1). Section 5.2 will describe different approaches towards determining abandonment and retrial behavior of callers. In forecasting there are many ways to make mistakes with historical data. Section 5.3 will deal with the possible pitfalls.

5.1 Different approaches towards forecasting

5.1.1 Model types

Gans et al. (2003) and Koole and Mandelbaum (2002) both mention two main activities in determining input data for further call center modeling. Input data can be for example arrival patterns, service durations and caller's patience behavior. In the first stage data can be analyzed with three possible types of models:

- Empirical or descriptive model. Suitable for organizing and summarizing the data being analyzed. The simplest of these are tables or histograms of parameters and performance.
- Theoretical model, which seeks to test whether or not the phenomenon being observed, conforms to various mathematical or statistical theories. Examples include the identification of an arrival process as a Poisson

process (queueing model) or of service durations as being exponentially distributed.

- Explanatory model. In-between the descriptive and the theoretical model falls the explanatory model. It is often created in the context of regression and time-series analysis. The explanatory model goes beyond, say, histograms by identifying and capturing relationships in terms of explanatory variables. At the same time, this model falls short of the theoretical model in that there is no attempt to develop or test a formal mathematical theory to explain the relationships.

The empirical and theoretical models are mainly used to draw conclusions on the primitives (basic parameters) of call center processes. Queueing models are theoretical models which mathematically define relationships among building blocks, for example, arrivals and services. Queueing analysis of a given model starts with assumptions concerning its primitives and leads to properties of performance measures, such as the distribution of delay in queue or the abandonment rate. Validation of the model then amounts to a comparison of its primitives and performance measures - typically theoretical - against their corresponding parameters in a given call center - mostly empirical. In the second stage, after data analysis, the forecasting activities can be executed on the basis of the models mentioned above.

5.1.2 Data gathering

Usually data is gathered from the ACD (and IVR) and put in ACD reports. Ideally would be the recording of transactional data per individual call, for the purpose of analysis (Koole and Mandelbaum, 2002). With the number of agents and (historical) data per timeblock, arrival process and service times can be analyzed. Also data of abandoning customers is recorded by the ACD. Pichitlamken et al. (2003) warn that a lack of specific call data may complicate the forecasting on arrival rates, service times and other call center parameters. Gans et al. (2003) describe different types of data which can be used for analysis:

- Operational customer data; for specific call data, arrival patterns, delay in queue and service times. Operational customer data provide listings of every call handled by a site or network of call centers. Each record includes time stamps for when the call arrived, when it entered service or abandoned, when it ended service, as well as other identifiers, such as who was the CSR.
- Operational agent data; for agent utilization, availability, duration of being idle and service times.
- Marketing data; from the corporate IS (Information System), for example for qualitative measurements.
- Human Resources data; to find out what skills and skill-levels agents have, for example for use with SBR.
- Psychological data; for qualitative measurements (deal with subjective perceptions of customers)

The authors also collect data from the ACD reports and with support from the WFM the reported numbers are aggregated to monthly totals. These totals form the (historical) basis for forecasting input parameters for modeling. The term 'grand averages' is used for aggregating different totals (arrival rates and/or service times). An example of hierarchical views of arrival rates can be found in appendix D. Ernst et al. (2004) also mention the use of historical data to forecast demand for services. Point of departure for this author is flexible demand where the likelihood of future incidents is less well known and must be modeled using forecasting techniques. Requests for service may have random arrival rates and possibly random service times. Zapf (2004) recognizes the importance of recording historical data with the ACD and adds an interesting source of data gathering: expert information. Experts are for example experienced agents who make estimates on arrival rates and handling times (per type of call, per type of agent). The author uses the historical data and expert estimates to dimension incoming requests and waiting tolerance.

5.1.3 Data processing and analysis

Timeblocks – Arrival rates

Once the top-level (monthly totals) forecasts are set, they are split into day-of-the-week/day-of-the-month, as well as by time-of-the-day specific numbers of arrivals of calls (Gans et al., 2003). Each period or interval is allocated with a number or a percentage of the total call volume for the aggregate period. For these periods and intervals constant arrival rates are assumed, which allows the use of standard, steady-state models, where steady-state is reached quickly and forecasted λ and μ may be used.

Another approach mentioned by the authors (Gans et al., 2003) is breaking cycles into smaller intervals to get a sample mean that is used as the arrival rate for the subinterval. The analysis shows that, when the underlying arrival process is time-inhomogeneous Poisson and cyclical (explained in upcoming alineas), the limiting sample-rate function is a consistent estimator of the original arrival-rate function. Massey et al. (1996) prove this with a linear function of arrival rates, $\lambda(t)=a_i+b_it$.

Timeblock specific arrival rates are also proposed by Ernst et al. (2004), because after using queueing theory to determine staff levels, these levels could be specified for each hourly interval over a four week planning horizon.

Other authors who suggest aggregating data over some period of time (in a timeblock, typically 30 minutes) are Pichitlamken et al. (2003). The goodness of fit is much better with split-up arrival patterns for different intervals. Furthermore they state that from empirical study, call center arrivals are known to have a variance that is considerably higher than implied by Poisson arrival (their sources: Jongbloed and Koole (2001) and Deslauriers et al. (2004)) and strong positive association between the arrivals in different time periods (their sources: Tanir and Booth, 1999 and Brown et al., 2002).

Poisson nature

Arrivals to call centers are typically random (Gans et al., 2001). There are many potential, statistically identical callers to the call center; there is a very small yet non-negligible probability for each of them calling at any given minute, independently of each other. Under such circumstances, theory dictates that the arrival process fits well a Poisson process. If more customers are likely to call say at 10:30 am than at 1:00 pm, one gets a time-inhomogeneous Poisson process.

Call arrivals can be determined with use of descriptive models to reflect patterns, explanatory models for forecasting future arrival rates and classical theory models to prove the time-inhomogeneous Poisson nature of the arrival process. The use of classical theory models for proving the time-inhomogeneous Poisson nature of the call center process has also been put forward by Pichitlamken et al. (2003), backed up by empirical evidence in Brown et al. (2002). They found out that a Poisson process with a deterministic time-varying arrival rate cannot realistically model the call center arrivals. Assuming deterministic arrival rates reduces traffic variability and congestion and improves the quality of service and performance measurements. A stochastic rate would be more realistic. They tested this assumption and concluded that a better goodness of fit is obtained when the arrival process is time-of-the-day and day-of-the-week dependent. With that finding they partly reject their own remark on the deterministic time-varying arrival rate. The arrival rate may be time dependent but should be more stochastic in the specific time blocks. The authors end this discussion with the remark that possible errors or successes of the assumptions made, are strongly related to the amount and reliability of the data.

Zapf (2004) states that the arrival pattern of incoming requests has a Poisson nature, because of findings in scientific literature. The exact Poisson nature can be derived from the ACD data. The average volume of synchronous requests, which is required as input parameter for this distribution, can be derived empirically from the output of the ACD. Based on one typical week without extreme work loads an average request volume per day has been determined. This volume contains standard calls and special calls. The request volume of a-synchronous requests has been derived from the overall request volume per month which leads to standard and special requests per day. Here too, the call volumes have been split up to get time and interval specific arrival parameters.

Predictability

There are scenarios where the Poisson assumptions are violated. For example by an external event, such as a telephone number shown in a TV commercial, which can be modeled by adding a Poissonian number of arrivals at a predictable point in time. This is still referred to a Poisson point process, which experiences discontinuity in its cumulative arrival rate.

For short term periods one is encountered with stochastic variability in the arrival process. Over longer periods even a predictable variability, such as a seasonal pattern, can sometimes be discovered. Lin, Lai and Hung (1998) recognize this and even add the option of influencing their forecasts and input parameters by factor for the adjustment on call arrivals on particular hours or days specified (see also appendix E).

Erlang distributions
Erlang is a unit of measurement of traffic density in a telecommunications system. The erlang describes the total traffic volume for a certain period of time. In the early days of call center literature the distribution used for call arrivals was a standard Erlang C distribution, which nowadays is too restrictive for the much more complicated call center processes that exist. Ernst et al. (2004) state that the assumption of the Erlang C nature of the arrival process in call centers is a rather simple and shortsighted one. During their literature research Koole and Mandelbaum (2002) found many analytical models for performance analysis with Erlang B and mostly Erlang C distributions for the arrival process rates. According to the authors the first is typically inappropriate for not acknowledging waiting and the second lacks central features, notably customer abandonment and heterogeneity.
Koole (2005) notices that the workhorse queueing models have been the Erlang B (loss) and C (delay) models, known as M|M|s in the standard Kendall queueing notation. The most common extensions considered attempt to account for customer abandonment, customer retrials, non-exponential call-holding-time distributions and timevarying arrival rates, but even these familiar phenomena pose serious analysis challenges.
While heterogeneity could require a leap in modeling capabilities, the Erlang A model is ripe for applications. For details on this queueing model and its possible extensions we refer to the article by Gans et al. (2003).

Service or handling times
In the field of determining service times Gans et al. (2003) discovered that scientific literature concentrated exclusively on description (like histograms) and validation of theoretical models (tests for goodness of fit). Hardly any explanatory work can be found in literature. Erlang and exponential distributions are mentioned as possible patterns for the distribution of service times, but the theoretical justification for using an exponential distribution is usually quite doubtful. Analytical traceability along with a lack of empirical evidence to the contrary is not really a strong basis to draw solid conclusions.

Pichitlamken et al. (2003) confirm the lack of reliable and constructive literature on the determination of distributions of service times. Their estimates are based on the sum of service times available and they used a case to determine whether the service times are time-of-the-day independent or dependent. The authors got a better fit with the latter. The assumption was tested for a gamma and a lognormal distribution, with a simulation model. The gamma distribution turned out to be much easier to test and the goodness of fit was reasonably. Besides this the gamma is very usefull for modeling random experiments (Montgomery and Runger, 1999).

Zapf (2004) argues that a stochastic distribution is the right way for modeling service times in a call center, because the times are not constant in reality. In order to determine the appropriate stochastic distribution and average values one typical week of data has been statistically analyzed gathered from the ACD. A sample of real call times can be compared with data from exponential distributions. The X^2-test will deliver a P-value. (In Montgomery and Runger (1999) it can be found that the X^2-test is a special case of the gamma distribution and can be used for interval estimation and tests of hypothesis.) The test of Kolmogorov/Smimov will result in a lower bound for the P-value. A P-value of more than 0.10 stands for a good correspondence between the distribution and the sample data. Based on the test results the exponential distribution can be used for modeling the relevant handling times: classification time, call time and after-call time. The average values for the call time and after-call time can be derived from the ACD system, the average classification time has to be estimated by experts. For a-synchronous requests no distinction between call handling and after-call work is necessary, therefore only one handling activity can be modeled for each process design. The average values for the corresponding handling time have to be estimated by experts.

Example

A nice example of deriving forecasts from an ACD report is provided by Lin, Lai and Hung (1998). Based on hourly ACD data an appropriate model is built with the purpose of showing the relationship between abandonment rate and workload (based on demand volume and service capacity). See appendix E for detailed information on the method followed to determine the relationship. Timeblocks are used to get specific forecasts per interval and since no particular trend was observed, a moving average (for each hour in the week) is adopted as forecasting model. The methodology's objective is to directly evaluate net staff levels for certain periods of time in order to stay below pre-fixed abandonment numbers and/or percentages. In case of insufficient past data for constructing the model, a queueing simulation model is adopted to evaluate net staff level. For the call arrival pattern Poisson call arrivals are used to simulate incoming requests by customers. Talk time and abandonment time will then be normally distributed.

5.2 Determining abandonment and retrial behavior of callers

Modeling abandonment parameters is extremely complex (Mehrota and Fama, 2003) because of mathematical implications from queueing dynamics and also because of a lack of proper and detailed observable data about customer abandonment and retrial. Two main questions are identified:

1. What is the customer's tolerance for waiting, and at what point will this customer hang up and thereby leave the queue?
2. How likely is the customer to call back, and after how long?

Mehrota and Fama leave the second question unanswered and state that, for answering the first question the distribution of patience has to be extracted (estimated) from historical data about callers' time in queue. For the simulation model used by the authors, a life span for the waiting time of customers is drawn from an exponential distribution.

In figure 4, Gans et al. (2003) already acknowledged the existence of abandonment, lost calls and retrials. They also recognize that a lot of published papers pay attention to abandonment and patience, and few about retrying. The impatience function distinguishes between regular and high priority customers, where regular customers are less patient than high-priority customers. This could be a reflection of a more urgent need on the part of priority customers to speak with an agent, or it could reflect their higher level of trust that they will be served soon after arrival.

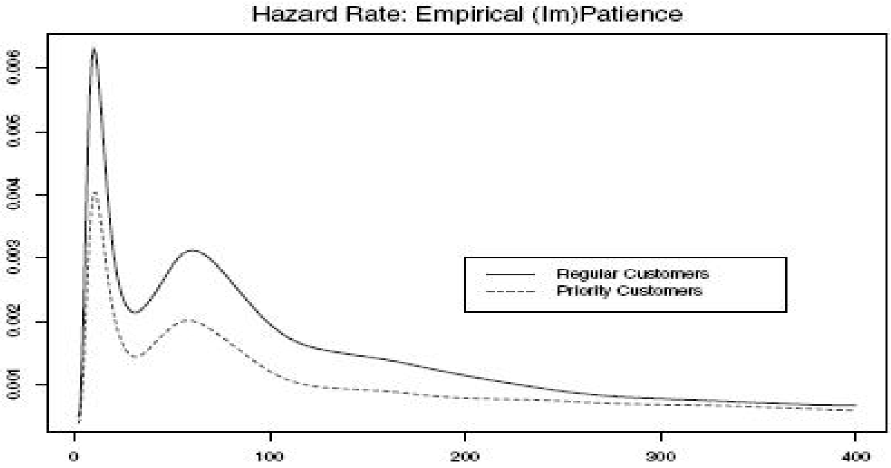


Figure 8: Impatience Functions of Regular and Priority Customers

Second, the impatience functions (figure 8) of both types of customers are not monotone and have two peaks: the first near the origin, due to those who simply decide not to wait, and the second at about 60 seconds. The second, as it might happen in a particular situation, reflects an announcement to customers who have waited 60 seconds, informing them of their relative place in the tele-queue (but not their anticipated waiting time). As can be seen, the information here encourages abandonment. This could be in contrast to its original goal, namely preventing abandonment by reducing the uncertainty about waiting times.

Many models of impatience have been developed in the call center field. The authors provide a small list of specific (case-based) examples with different distributions for (im)patience behavior. Distributions like Weibull, Erlang A and lognormal are mentioned. In the different cases (im)patience is based on irritation (based on inconvenience) or short, medium or call-back delays.

Impatience behavior

Data generated from the IVR and/or ACD also shows the number of abandoning customers and the times after which a customer decides to leave, abandonment times (Gans et al., 2003). Most research on the subject of abandonment and retrials by customers in 'tele-queues' originated in psychology and marketing

(Koole and Mandelbaum, 2002). In scientific literature most attention is given to the (im)patience of visiting customers, with little devotion to the tendency to retry. In chapter 3 some findings of Gans et al. (2003) are already presented in the field of the impatience factor. Every type of customer has its own (im)patience behavior and there is a difference between the customer's time willing and expecting to wait.

Interdependency and equilibrium

To indicate the influence of abandonment on call center process performance, the authors state that in heavy traffic, even a small fraction of busy-signals or abandonment could have a dramatic effect on performance. Therefore decisions on agent staffing must take into account customer patience. In turn, customer patience is influenced by the waiting experience which, circularly, depends on staffing levels. An appropriate framework, therefore, is that of an equilibrium (Game Theory), arrived at through customer self-optimizing and learning. Abandonment arises as an equilibrium behavior of rational customers who optimally compare their expected remaining waiting time with their subjective value of service. A simplified model can be derived from the equilibrium behavior, which enables some support for adaptive behavior (learning) of customers.

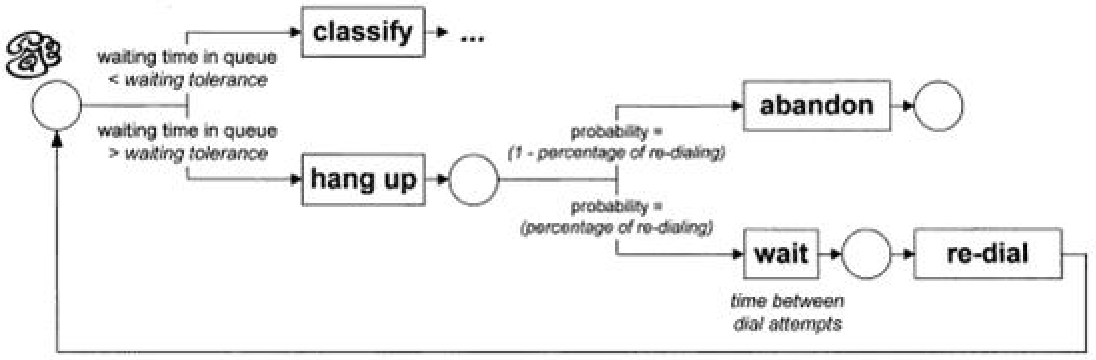


Figure 9: Call abandonment process for synchronous requests

Abandonment and redialing modeled

Zapf (2004) provides a nice representation (figure 9) of the abandonment process and redial possibility for synchronous requests. Customers do not wait for an agent as long as the service provider would like but only for a particular time period. If calls are in the waiting queue for a longer time the customer hangs up.

This period is called the waiting tolerance which is determined by the customers' preferences and his current situation. The data used are from a particular case and company. Since the ACD system does not log the waiting tolerance, an estimate from experts is used which results in an average waiting tolerance of 1:00 min per customer request. The variation of the waiting tolerance between single requests is reflected by using the exponential distribution for this parameter. After hanging up, some customers re-dial in order to get a free agent. The part of re-dialers is represented by the percentage of re-dialing. The parameter time between dial attempts defines the time interval between hanging up and re-dialing and is also supposed as exponential distributed. Communication center experts estimated 75% percentage of re-dialing and an average time between dial attempts of 0:06 min. The waiting tolerance and re-dial process have also been modeled for calls which have been classified by a generalist and have been transferred in a waiting queue of a specialist agent.

5.3 Errors and pitfalls in forecasting

Estimation vs. prediction
In scientific literature on call center analysis some warnings are given by several authors about estimations and determinations of call center parameters based on historical data. First of all Gans et al. (2003) make the important distinction between estimation and prediction. These are two closely related, but different, statistical tasks. Estimation concerns the use of existing (historical) data to make interferences about the parameter values of a statistical model. Prediction concerns the use of the estimated parameters to forecast the behavior of a sample outside of the original data set (used to make the estimation). Predictions are "noisier" than estimates because, in addition to uncertainty concerning the estimated parameters, they contain additional sources of potential errors.

Sources of uncertainty – Gans et al. (2003)
Forecasting or prediction may not be trustworthy because of insufficient historical data, which can lead to unpredictable factors not discovered during data analysis. Furthermore the authors provide three sources of uncertainty in available data:

1. Process uncertainty. Inherent random because exact call arrival times are usually not known in advance.
2. Model uncertainty. Any model is an approximation of reality, and therefore necessarily misspecifies the underlying phenomenon to some extent. The estimation of the arrival process rate might not be totally correct.
3. Model uncertainty on performance measures. System performance may not be insensitive to the form of the service-time distribution. Given a model that describes reality satisfactorily, there still may exist parameter uncertainty, as in the case of a arrival process with an uncertain rate.

Data from the ACD can be pretty censored when it is used to analyze and determine abandonment parameters. Since the data of served customers who did not abandon the system is not available, patience is not fully observed. Only the maximum patience times of those customers who abandon are observed. Pichitlamken et al. (2003) do mention this concern as well, next to the fact that the time an agent is available to take a call is very likely to be less than the time for which they are scheduled, because of coffee breaks, trips to restrooms, absenteeism, etc.

Sources of uncertainty – Chen and Henderson (2003)
These authors have a typical approach towards forecasting input parameters for a queueing model. Determination of forecasts is based on historical data and the

authors mention sources of uncertainty and provide tools to detect and model random arrival patterns. At least three potential sources of uncertainty in estimating the arrival rate for a future period:

1. Estimation error. The arrival rate estimator will not deliver an exact value of the arrival rate parameter, because it is an average of a finite number of random variables.
2. Non-stationarities in the data available. Arrival rates for future periods may not be well-predicted by the number of arrivals in the corresponding previous periods. The use of several seasons of data may reveal and predict such non-stationarity.
3. Random arrival rate. The historical data of a particular period may show a call arrival process with a Poisson nature. Estimations for the future will be based on these findings. In future particular periods one might measure arrival rates that seem to be random. This could be caused by the weather for example, or other external (and thus random) factors.

Detecting and modeling a random arrival rate is explained by the authors with a practical method using statistical analysis and hypotheses. The random arrival rate should be explicitly modeled if the underlying performance measure is highly nonlinear over the range of the random arrival rate. Otherwise the presence of a random arrival rate will lead to over-predictions of service performance and an underestimation of staff required is made. Of course, the main concern from a practical point of view is the degree of this effect. A practical method is supplied for detecting and modeling a random arrival rate, and it is described how to compute performance in this setting. The approach is very general, and in particular does not rely on the use of very specific models, nor does it rely on any convexity assumptions. One can use this method in a simple "pilot study" to compare performance assuming a deterministic arrival rate and assuming a random arrival rate. For more details on the method followed view the article by Chen and Henderson (2001).

6. Approaches to determine (optimal) staffing levels

Once activities as forecasting inputs and setting the right performance measurements have been completed, call center managers can start thinking of determining the optimal staff mix. Without a model it is difficult to specify and justify the performance of solutions. Depending on the number of types of customers one can think of different combinations of generalists and specialists in the optimal mix. For sure, the manager will think an intermediate solution is the best solution (Pinker and Shumsky, 2000). The choice for a certain workforce configuration involves a trade-off between the efficiency of cross-trained workers (generalists) and the higher experience and quality of specialized workers (specialists). There are several techniques that can help the manager in confirming or contradicting his intuition. Among others, Ernst et al. (2004) state that queuing theory and simulation modeling are the two approaches most commonly used for translating customer arrivals during different time intervals into the staffing levels (demand) needed to maintain the required service standards. In section 3.5 Ernst et al. (2004) even recommend to use the two in combination to obtain best results.

In this chapter a number of approaches towards determining staffing levels will be described to get a better view on the specific characteristics and (dis)advantages of these approaches. If applicable, measuring sensitivity with the tool will also be dealt with. First the queueing models will be described (section 6.1). Then simulation is subject of discussion (section 6.2). Furthermore a number of tools for minimization will be described, mainly mathematical techniques like Integer Programming and square-root staffing (section 6.3). At the end of this chapter some findings of the different tools will be presented in the conclusions section (6.4).

6.1 Queueing theory models

According to Koole and Mandelbaum (2002) and referring to figure 2, call centers can be viewed as queueing systems. A lot of the principles in queueing theory match with the characteristics in call center processes. In section 5.1.1 the usefulness of queueing theory is already explained:

Queueing models are theoretical models which mathematically define relationships among building blocks, for example, arrivals and services. Queueing analysis of a given model starts with assumptions concerning its primitives and leads to properties of performance measures, such as the distribution of delay in queue or the abandonment rate.

6.1.1 Erlang and Poisson

In a queueing model of a call center, the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and tele-queues consist of callers that await service by a system resource. The simplest and most-widely used model is the M|M|s queue, also known in call center circles as Erlang C. As already stated before, for most real-world applications Erlang C is an oversimplification. It assumes out busy signals, customer's impatience and services spanned over multiple visits (Gans et al., 2003). The basic operational model of a call center is the M|M|s queue with parameters λ, μ and s, the primitives:

- the arrival process, assumed Poisson at a constant rate λ
- the service times, assumed exponentially distributed with mean μ
- the number of agents, s

Furthermore there are implicit assumptions, of which independence among the primitives and FCFS service disciplines are the most important.

The popularity of using M|M|s queues is related to the fact that closed form expressions exist for most of its performance measures. When modeling call centers, the useful approximations are typically those in heavy-traffic, namely high agents' utilization levels at peak hours. To explore other possible queue types, the authors shortly mention some researched queue types.

For example the M|G|s queue, as a result of non-exponential service times. Unfortunately analytically intractable. One must then resort to approximations and it turns out that performance improves as stochastic variability in service times increases (decreases). For small to moderate numbers of agents s, research asserts that waiting time is approximately exponential. Large s, on the other hand, gives rise to a different asymptotic behavior. Some mathematical solutions have been developed to deal with this approximation problem. Two of them will be described further on in this chapter (sections 6.3 and 6.4); square-root staffing and site pooling.

6.1.2 Gatekeepers and referrals

Hasija et al. (2005) use queueing theory to show that there are essentially two queues in series:

- n_g gatekeepers
- n_e experts

with both different staffing costs. Customers arrive to the gatekeepers according to a Poisson process with rate λ. Each call has a complexity x, which means the probability of a customer is being treated successfully by the gatekeeper is f(x): the treatment function. In section 2.3, figure 4 a representation of a gatekeeper configuration was presented. The referral rate (k) is used as a variable input parameter and can be seen as the policy for the maximum of calls not treated (well) by the gatekeepers. In other words, 1-k is the call resolution rate of the gatekeepers. Fraction k of the calls ends up at the desk of experts. The arrival

rate at experts therefore is the sum of the rate of calls untreated and mistreated by the gatekeepers.
Minimizing costs is the objective. Some parameters can be fixed, others variable. Solving this problem can be done numerically (minimizing the objective function under certain restrictions) and will be described in section 6.3, focusing on mathematical solutions for minimization problems.

The authors state that for planning staffing needs, first referral rates and gatekeeper treatment workloads have to be determined, in order to have information for the higher level staffing model. A few assumptions:

1. The referral (rate) is not influenced by queue lengths at gatekeepers and specialists. This leads to a simpler structure, allowing to focus on the long-term impact of incentives on gatekeeper behavior (principal-agent model).
2. Strict separation between gatekeeper's diagnosis and treatment steps.
3. All gatekeepers have same diagnostic capabilities and all specialists are homogeneous in terms of cost. Expected cost is independent from identity of gatekeeper or specialist.
4. Incorrect treatment by gatekeeper leads to directly sending the service request to a specialist.

Hasija et al. (2005) also researched sensitivity of the models and processes used. The cost-minimizing referral rate varies with changes in parameters related to queuing, i.e., arrival rates and service rates. When the arrival rate increases, the optimal referral rate converges to the optimal referral rate for the deterministic case. For very large λ, waiting costs are relatively small, compared to the sum of staffing and mistreatment costs. Therefore it is optimal to use the treatment threshold from a deterministic model, which only considers staffing and mistreatment costs. The authors also provide a rule of thumb for choosing the optimal system. Treatment threshold k_d is the main parameter the rule is based on. Appendix I includes a more broadly description of the rule and of the methodology followed by Hasija et al. (2005). The authors found that with a certain skill level of the gatekeeper a direct access system (only experts) is optimal. So, to justify gatekeepers (assumed high waiting costs) they should have high level skills. This effect can be explained by the fact that a one-tier system offers benefits from pooling (section 6.4) and that these benefits are more powerful when waiting costs are high.

6.1.3 Shared resource as bottleneck server

Akşin and Harker (2003) developed a model that is interesting in queueing perspectives, despite the unusual assumption of the IS as a shared bottleneck resource (see also section 2.2.1). The problem being considered is a staff dimensioning problem for a service system, which determines the optimal number of servers that is allocated to multiple customer classes. A lot of methodologies have the objective to minimize costs. As mentioned in the introduction already, call centers can nowadays be more seen as profit centers.

That is the reason why the authors take revenues as a direct function of staffing decisions. As with most of the earlier call center staffing papers, the underlying performance model is a queueing system. What is unique about the performance model, however, is the explicit consideration of buffer and server resources, as well as a resource that is shared among different customer types. Each customer class has its own server and buffer resources, but shares an information processing resource with other customer classes.

Since measuring quality is extremely difficult, the authors want to determine economically optimal staffing levels, defined as those levels that maximize total revenues net of staffing costs for the service system. These economically optimal staffing levels can be determined for a loss system. A more detailed description of the authors' model is provided in appendix J. In the model, the service time is assumed dependent on the number of customers in the system, because of the shared resource. The formulated sizing problem is in particular difficult to solve for the blocking and renege probabilities, since they are non-linear in the number of servers. For the general multi-class case, heuristics are developed that make use of the structural properties of a single class system.

The authors did not model the possibilities of unsuccessful treatment of customer requests and multiple skilled servers are also not considered in the article.

6.2 Simulation models

In section 3.5 a number of authors are quoted for emphasizing the usefulness of combining queueing theory and simulation for determining optimal staffing levels. The simulation model is used to perform what-if scenarios, because of high flexibility. The queueing model (CTMC, Markov) is used to approximate the system performance measures. CTMC models are insightful and relatively easier to construct than a simulation model.
Mehrota (1997) already recognized the importance of simulation for future planning. Call center managers are relying increasingly on simulation models as the source of answers to key "what-if" questions, and as the right way to design and/or modify different aspects of their call centers. Furthermore he states that randomness, complexity and interactions can be very well modeled with simulation, which is also backed up by Bapat and Pruitte (1998).

Koole (2005) goes further into this subject and states that after building a mathematical model of the call center and estimating all relevant parameters, conclusions can be drawn from a thorough analysis of the model. This analysis often involves simulation of the whole system. This approach is time consuming, usually performed by external consultants, and whether it really gives good results is sometimes doubtful. It works best for operational problems of a repetitive nature. Furthermore he concludes that for evaluating more complex call centers (two or more skills per agent) the analytical models based on Erlang formulas are no longer suitable. Instead one has to rely on mathematically involved and long computations, on approximations, or on simulation. More or less the use of simulation in evaluating complex call centers becomes a necessity.

Chan (2003) provides a nice overview of steps to be followed to design an effective workflow using simulation. A description of the steps can be found in appendix G.

6.2.1 Zapf (2004): Discrete event simulation

Zapf (2004) uses discrete event simulation to overcome certain complexity restrictions of modeling and therefore the evaluation of process designs is close to reality. For every design, stochastic discrete event simulation models are made. With the simulation tool ARENA, based on the SIMAN simulation language, a model of a call center process was designed. Some parts of the model were developed with the call center specific extension Call$im, other parts have been implemented through individual routines. Zapf (2004) also models outcalls (related to accepted customer requests) in the simulation models. A representation of the process with outcalls can be seen in appendix B, figure 16. After initialization of the simulation model (setting a number of agents per group, per design), different scenarios are run. Conclusions on these results will be discussed in section 6.4.

6.2.2 Testing validity and sensitivity with simulation

Mazzuchi and Wallace (2004) validated their simulation model with the help of some call center experts and did not use any real-world data or system. The authors admit the limitedness of this procedure and in their opinion good validation should be done with real-world output. They verified their simulation models with a number of industry-accepted verification techniques: modular testing, sensitivity testing, stress testing, trace analysis, and output comparison against known models. See Wallace (2004) for details.

Pichitlamken et al. (2003) compare simulation results to the collected empirical data from the call center. In their research this leads to slight differences which can be related to the percentage of time that employees are not available to take calls.
Furthermore the authors examined sensitivity of the assumptions of the simulation model. Distributions of arrival process and service times are changed and tested with the CTMC models that are modeled parallel to the simulation models.

Zapf (2004) presented the simulation study and discussed it with communication center professionals in order to get feedback from practice concerning the applied method and the obtained results. Details on the validation and verification steps identified by Zapf (2004) can be found in Appendix H.

Mehrota and Fama (2003) discuss a very interesting part of simulation procedures. They ran multiple replications of the simulation model and computed estimates for performance measures based on the average of the run length. This was done for each of the individual scenarios that were developed. For purposes of determining the number of runs for each scenario, they focused on average weekly Service Level for the inbound queue as the statistic of interest. After each run, overall standard deviation of this statistic was examined across all runs to date. They continued to run additional iterations until this overall standard deviation was under 2.5%, which was set arbitrarily as the confidence threshold.

Koole (2005) warns for the possible time-consuming verification and validation activities, certainly with complex systems. He states verification and validation are crucial to a proper use of simulation.

6.3 Mathematical tools for minimization problems

Mathematical models are usually used for approximation of staffing solutions, especially where minimizing costs is the objective.

In section 3.5 the hierarchical approach of Gans et al. (2003) towards capacity management for call centers was already mentioned. The intermediate-level of personnel scheduling is mainly done with mathematical programming models. Point forecasts for system parameters are derived from the low-level queueing performance models and used as input for mathematical programming models. Ernst et al. (2004) also note that the literature is heavily skewed towards mathematical programming and metaheuristic approaches for rostering as opposed to CP (constraint programming) and other techniques arising out of artificial intelligence research.

As mentioned in section 6.1.2, Hasija et al. (2005) use linear programming to numerically approximate the optimal staffing levels. With the costs of staffing per unit time a minimization with certain restrictions is used to solve the problem. Akşin and Harker (2003) use a greedy allocation algorithm to solve the problem for the loss system. Servers are assigned sequentially to the customer class that improves revenues net of staffing costs the most, and continues until no such activity can be found within the feasible set.

6.3.1 Integer Programming (IP)

In their literature research Koole and Mandelbaum (2002) concluded that with modeling staff, integer programming is more focussed on rostering than on finding the optimal mix.
As said before, Gans et al. (2003) use outcomes of lower level queueing performance models as input for staff scheduling with integer programming models. All blocks together, and their specific forecasted arrival rates and service times, give rise to a target staffing level over longer periods. The authors distinguish between two elements of the scheduling process: shifts and

schedules. A shift denotes a set of half -hour intervals during which a CSR works over the course of the day. A schedule is a set of daily shifts to which an employee is assigned over the course of a week or month. Both shifts and schedules are often restricted by union rules or other legal requirements and can be quite complex. Then determination of an optimal set of schedules can be described as the solution to an integer program (IP).

In section 3.6 an iteration and convergence tool RIIPS (Henderson and Mason, 1998) was introduced. This tool tries to link between adjacent timeblocks in the staffing process. Flexibility in rostering solutions increases with this tool. RIIPS is quite a complex, time consuming problem to solve. Two algorithms are provided to save time in IP modeling and finding optimal solutions. GOS is used to measure and test possible solutions by simulation and the question to answer is whether or not performance criteria are met. The algorithms are performed until convergence.

6.3.2 Square-root staffing

Another approximation tool for staffing is the square-root method. Koole and Mandelbaum (2002) discovered a lot of attention for this tool in scientific literature. For all kinds of M-queues and especially in case of heavily loaded call centers. Gans et al. (2003) also mention the heavy traffic characteristic and state that with high N (a lot of service requests) an asymptotic regime can be followed with the square-root staffing method. The square-root method provides three advantages (Hasija et al., 2005) over numerically solving a problem:

1. Solve staffing problems more quickly
2. Enables characterization of the effects of certain parameters on the optimal solution
3. Direct comparison possible between the one-tier (only experts/specialists) and two-tier (gatekeeper) systems.

As mentioned in section 6.1.2 Hasija et al. (2005) use the square-root staffing rule to find optimal staffing for both tiers (gatekeepers and specialists). The number of servers for each level can be determined for any particular referral rate. In other words: for any routing strategy.
With the objective of minimizing total staffing and waiting costs and λ allowed to near ∞, the ratio of staffing and waiting costs is bounded. Such a system can be described as being in the rationalized game. Then, a simple square-root staffing heuristic provides asymptotically optimal results. A description of the heuristics for one-tier and two-tier (including treatment threshold k) systems can be found in Hasija et al. (2005). With the obtained number of servers and the static routing strategy (depending on the referral rate and treatment threshold), the determination of total costs of operating is done with minimization.

6.3.3 The “pμ rule”, de Vèricourt and Zhou (2003)

In the model of Hasija et al. (2005), gatekeepers have some probability of success with a particular call. The same is done by de Vèricourt and Zhou (2003). They assume that each server has a different call resolution probability (p), and they also assume that each server may have a different service rate (μ). They identify the routing policy of calls to servers (a “pμ rule”) that minimizes the total time a call spends in the system, including re-calls. While de Vèricourt and Zhou (2003) assume that the staffing level is given - one server of each type – the model of Hasija et al. (2005) considers both the staffing and routing problem for large systems. The structure of their service system is also quite different. They assume that there are two pools of servers: the expert pool has a resolution probability equal to 1 and the gatekeeper pool attempts to treat calls or passes them along the expert pool.

6.4 Conclusions from modeling techniques

In scientific research one can find many interesting experiences from different authors when modeling the process of a call center. Especially when playing with the numbers of servers, generalists and specialists and with task division (skills and referral rates).

6.4.1 Resource pooling

Mazzuchi and Wallace (2004) provide a nice description of the phenomenon called “resource pooling”. In a skill-based routing call center environment, agents are flexible and can support multiple skills. If agents have only one skill in an SBR environment in which there are n different work groups, then it is well-known that the system will behave as a collection of much smaller independent call centers (assuming blocking is negligible). At the other extreme, if each agent can support all service requests or skills, then the system behaves as one big call center or single multi-server system. Under this big call center scenario, there is no situation in which there are waiting customers and idle agents. When this situation occurs, the system exhibits full resource pooling or simply resource pooling.

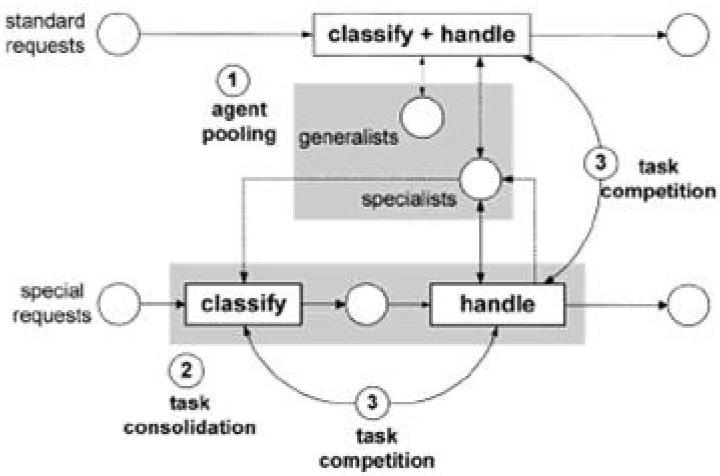


Figure 10: Competing effects

Zapf (2004) also recognizes the impact of resource pooling and observes an interesting fact when looking at competing effects in his model. A representation of the observed effects can be seen in figure 10. Agent or resource pooling is achieved by using specialists for classifying and handling standard requests (dotted lines) in case generalists cannot cope with the offered load of service requests. The number of resources for one task is increased and therefore the performance for standard requests is better. On the other hand fewer resources are now available for handling special requests since the total number of resources remains the same. This task competition (3 in the figure), between “classify + handle” and “handle” leads to a worse performance for special requests.

Another way of task competition can be observed in case of task consolidation, which means that two tasks which have been performed by different agent groups before are consolidated and handled by one agent group afterwards. The tasks ‘‘classify’’ and ‘‘handle’’ are consolidated for special requests and handled by specialist agents (dotted lines). This consolidation leads to less handling time and therefore to a better performance for special requests. Contrary to this acceleration the specialists have an additional work load through the classification and have less capacity for handling special requests. The tasks ‘‘classify’’ and ‘‘handle’’ compete for the same resources which reduces the number of handled tasks and therefore the overall performance for special requests.

Resource pooling was also identified by other authors like Gans et al. (2003) and Wallace and Whit (2004). For example the latter show that extreme resource pooling (agents have all skills) is not necessary to perform better in terms of service and waiting times. Using a one-factor-at-a-time SBR analysis they show that the system where agents have two skills performs nearly as well as the system where agents have all skills.

Mazzuchi and Wallace (2004) identified several ways to characterize the existence of resource pooling. With some fixed parameters like routing policy and size of the trunk, the authors prove no interactions between different call rate factors exist while the system is experiencing resource pooling. Interaction between factors occurs when the difference in response between the levels of one factor is not the same at all levels of the other factors. The existence of interaction can be determined by drawing interaction graphs. For more details on determining interactions please read the article by Mazzuchi and Wallace (2004)

6.4.2 Sensitivity

Measuring sensitivity was already subject of discussion in section 6.2.2. There it was tested by means of simulation. Here some more common interactions between several parameters will be dealt with. Pinker and Shumsky investigate if the performance of the system is sensitive to the staffing configuration choice. For small systems with high learning rates, the optimal staff mix provides significant benefits over either extreme case (a completely specialized or completely flexible workforce). If the system is small and the rate of learning is slow, flexible servers are preferred. For large systems with high learning rates, the model leans toward specialized servers. The choice of workforce configuration involves a trade-off between the efficiency of cross-trained workers and the higher experience and quality of specialized workers. The dynamics of queueing systems show that the size of the system influences the efficiency gains created by cross-trained workers. On the other hand, the rate at which workers improve their QoS through experience influences the impact of the staffing decisions on the QoS experienced by the customer. Similarly the tenure process also affects the QoS. The authors state that there are a lot of feasible and optimal configurations of workforce and that for several reasons all flexible system turns out to be the worst staffing solution. Costly flexible servers make the system costly as well. Also the QoS is at stake in this extreme situation. The learning rate is high for difficult service requests and customers who need specialist help are likely to meet an inexperienced server.

Zapf (2004) identified some strengths and weaknesses for the different qualification- and communication-mixtures he came up with (section 2.1.1 and appendix B, figure 12 to figure 15). Table 2 in appendix B gives a rough summary of the identified strengths and weaknesses per qualification-mixture. The one-level design (figure 14, appendix B) is very good for handling standard requests because of pooling generalists and specialists. The design has weaknesses for synchronous special requests in overload situations and for special a-synchronous requests, since specialists are additionally occupied with standard requests. The back-office design (figure 13, appendix B) has strengths in handling asynchronous requests since specialists are reserved for these requests. Classifying synchronous requests is the weakness of the back-office design because of a small generalist group. The two-level design (figure 12, appendix B) does badly for synchronous and a-synchronous requests.

The integration of communication channels is in most cases more efficient than the separation (table 3, appendix B). Only in the back-office design (figure 13, appendix B) the integration of communication channels leads to worse performance for special synchronous requests which can be explained by task competition between handling synchronous and a-synchronous requests.

Some totally other effects that are found in a research to data mining approaches for call center performance, are mentioned by Paprzycki et al. (2004). The research analyzed the sensitivity of inputs and found that products, agents and dates could affect the quality of performance more than time management. The CSRs serving in some product areas have more opportunity to exceed the expected service time than the ones in some other product areas. The top performers constantly “exceed” or “far-exceed” the expectation. The performance of CSRs whose evaluation results fall into “met” or “below”, is not stable. The research suggest that the call center management team should focus on training and coaching the individuals and product areas which constantly have low quality instead of emphasizing balancing the length of times spent on calls.

7 Conclusion, future research and reflections

In this final chapter, conclusions and reflections on the literature review will be discussed. First, in section 7.1, the research assignment will be shortly repeated. The major findings on subjects discussed in chapters 2 to 6 will be listed (section 7.2). Furthermore some findings on future research (section 7.3), managerial implications (section 7.4) and limitations and learning experiences (section 7.5) will be provided.

7.1 Research assignment

The objective of this report is to "*find techniques for determining the optimal staffing levels in a multiple skill inbound call center* ". The literature studied was from scientific literature and included top-ranked journals and authors.

7.2 Major findings

The surevy revealed a table overview of existing approaches to determine optimal staffing levels. Apart from that, more insight is created on quantitative and qualitative performance measurements, redesign approaches, basic terminologies and generalists vs. specialists, forecasting and approaches to determine optimal staffing levels. These are summarised below.

Table overview of existing approaches towards determining optimal staffing levels

From the results of the literature research on different subjects, described in chapters 2 to 6, a table overview can be found in Appendix K. A summary is provided of the researched authors and their developed techniques to determine optimal staffing levels. In the table one can find different characteristics of the techniques or models such as:

- Author
- Typology of resources
- (Re)design approach
- Quantitative and/or qualitative performance measurements
- Special characteristics

No single ideal tool exists to solve the problem of determining optimal staffing levels. Every type of call center (case) is specific and demands a specific approach. Authors only provide different approaches to shape the tool to solve the problem.

Quantitative and qualitative performance measurements
From the researched scientific literature one can conclude that quantitative performance measurements are most widely used, because they are easy to determine and to measure. In chapters 3 and 4 it became clear that with the

(re)design approach and determining the performance criteria it is very hard to find a good balance or mix of performance measurements which reflects best the characteristics of a specific process. This dilemma is represented by the devil's quadrangle (appendix A).

Redesign approaches
For (re)design purposes, mainly form a managerial perspective, a number of authors have been discussed and it turns out that there is always a number of general steps/activities in the (re)design process. The same questions have to be answered over and over again. The general approach of (re)design is:

1. The use of real/historical data to forecast process parameters for future model use.
2. Establish the performance criteria to which the model/process has to perform.
3. Determine the staff requirements, based on a model
    - Queueing Theory leading to CTMC-models,
    - Simulation models or
    - A mix of queueing and simulation models.

A standard part of the (re)design approach should be the verification and validation of the process model and of estimates of (input) parameters.

Basic terminologies and generalists vs. specialists
Furthermore some basic terminologies in call center processes were introduced in chapter 2 like call center agents, ACD, timeblocks, and process parameters; the usual dimensions to which a call center is being categorized or analyzed. A definition of generalists and specialists was provided. Advantages and disadvantages of the both came forward and the conclusion is that a mix of generalists and specialists would be optimal in most situations in terms of flexibility and quality of service. Also different kinds of call center representations and the principles of SBR were provided in chapter 2.

Forecasting
On forecasting (chapter 5) it can be concluded that historical data (of all kinds of types) is a necessity to make reliable estimations and predictions on input parameters for call center modeling. The three main input parameters to determine are:

- the arrival rates,
- service times and
- abandonment behavior by customers.

An important part of data analysis is to study uncertainties. A lot of authors warn for different types of uncertainties, especially for model uncertainty which may have a negative effect on performance measures. A number of authors emphasizes (again) the importance of verification and validation of estimates and forecasts found with statistical analysis.

Furthermore there is a number of authors that emphasizes the importance of using timeblocks with aggregated data for forecasting input parameters. Under the assumption of reaching steady-state quickly, aggregating data (total number of calls, average waiting time, total abandoned calls, etc.) is a useful technique to estimate time-of-the-day, day-of-the-week dependent input parameters. The importance of abandonment and calling back is also noted, by stating that along with traditional forecasting issues such as data availability, data integrity, seasonality and non-stationary randomness, abandonment makes call forecasting more challenging than simply fitting a regression model to historical call volumes, particularly since there is usually no way to tell if an abandoned call led to a call back later on.

Specifically on the arrival rates some clear conclusions can be drawn from the literature research. The process of arrival rates can usually be very well represented by a Poisson process. Furthermore it is stated that instead of the traditional (and simple) Erlang C distribution for the distribution of arrival rates, it would me more appropriate nowadays (with more complex processes) to use the Erlang A distribution with some extensions to overcome customer abandonment, customer retrials, non-exponential call-holding-time distributions and timevarying arrival rates.

Approaches to determine optimal staffing levels

Main conclusion that can be drawn from the researched approaches to determine optimal staffing levels (chapter 6) is the ideal combination of using simulation models next to queueing theory/models to obtain best results. The simulation model can be used to perform what-if scenarios, because of high flexibility. Apart from the fact that simulation provides the possibility to model more complex processes. The queueing model (CTMC, Markov) is used to approximate the system performance measures. Furthermore CTMC models are insightful and relatively easier to construct.

Using queueing theory and simulation models, optimization is not directly possible. One can run different scenarios, but it remains unclear whether or not that scenario or solution is optimal or not. This problem can be solved by using mathematical models and techniques (like IP and square-root staffing rules) to perform minimizations (and thus optimizations), in combination with the tools mentioned before. By performing such algorithms one can find an optimal solution through convergence or reaching asymptotically optimal results.

A specific application of queueing models is represented by the model with gatekeepers and referral rates where the determination of the ideal referral rate is partly based on the principal agent model.

Again validation of developed models is very important and mentioned by a number of authors. Next to this it is always interesting to look at the sensitivity of a process model. Especially in terms of how sensitive the relationship is between certain parameters of a call center process model and the system (overall) performance. The presence of such an analysis really backs up the understanding of the system model in terms of interactions and system dynamics and puts it in the right perspective to judge on the real value of the developed model.

Another clear conclusion is the effect of resource pooling on system performance. When the number of available and suitable agents to perform as many tasks as possible is increased, waiting times will get smaller and system performance will be better (if that is a performance criteria of course). On the other hand, occupying the call center with more flexible agents (generalists) could be bad for the quality of service. Of course the pooling effects of different agents depend on (waiting, personnel) costs and levels of skills of generalists and specialists.

7.3 Future research

In most of the studied scientific literature a rather specific case or situation is being modeled and no real general models or statements are being developed. Therefore it is no wonder that almost every author advises to research call center processes from other domains and with other structures to gain more insight in the general way of modeling call center processes. With simulation it is sometimes proposed to study other (developed) scenarios to measure system performance and sensitivity.

Simplifications and too easily made assumptions are usually the cause of non-realistic process models. This is allowed of course for exploring the area of modeling call center processes, but in making progress in call center analysis, more complex processes should be modeled. Simulation modeling provides the opportunity to do so. In combination with queueing theory and algorithms to solve minimization problems. Future research should concentrate on this subject, because the changing role of call centers (from cost-center to profit-center) asks for more complex call center processes and thus analysis and modeling techniques for these more complex processes.

All authors state that the existing techniques to determine abandonment and waiting behavior (for tele-queues) and service times are still in its infancy. Further and future research will be necessary to determine realistic parameters as input for modeling call centers. Research on service/handling times is insufficient and is usually based on expert opinions. The lack of real and hard data and the lack of reliable and constructive literature make this area a real challenge for future research.

Due to the importance of the availability of good historical data, a number of authors emphasizes the increasing influence of database management and analysis. Rich databases do not always mean rich information. More valuable data could be extracted from the databases.

In the part of the major findings in this chapter it was already stated that most of the performance measurements used are quantitative. Qualitative measurements though should become more important, because (again) call centers are more and more seen as profit centers. Authors mention research in marketing and psychology areas to provide insight in the perception of customers in the call center processes. In order to determine how to measure the quality of call center processes. One author (Zapf, 2004) mentions conversation quality or customer satisfaction, which is very import for the overall performance, but if no agent is accessible, no conversation takes place and the customer could not be satisfied at all. So quantitative measures are the basis but not sufficient for an overall evaluation of organizational designs. Qualitative measurements should accompany the traditional quantitative performance measurements.

Furthermore, a number of authors mentions the lack of guidelines for the use of SBR options in the ACD. The full scale of possibilities of SBR remains unclear and in this area there is much to be discovered and researched about this subject. Gans et al. (2003) state this with "... the technology has raced ahead of managers' and academics' understanding of how it may best be used, and the characterization of effective strategies for skill-based routing is an open question at all levels of the capacity-planning hierarchy ..."

7.4 Managerial implications

As mentioned before in the conclusions managers have to consider many aspects when (re)designing a call center process. Especially in case of new staffing levels since costs for human resources often account for 50% to 75% of the operating expenses of a call center. After 'solving' the devil's quadrangle, the right performance measurements can be set up in order to measure when the call center's overall performance is optimal. Managers also have to bear in mind that competition effect and task consolidation play a major role in pooling resources and combining or separating tasks.

From the dilemma of the devil's quadrangle managers can also learn that (re)designing and staffing call centers optimally is not just a matter of looking at costs or flexibility, but calls for an inherently multidisciplinary research for better understanding of customer and CSR behavior. All aspects have to be taken into account. When modeling the call center process it is important to make sure the model represents the real world as much as possible. Assumptions and simplifications can be made, but not too rigorously. Estimates and distributions may be used for input parameters, but should be statistically derived from good historical data. This is where database experts could be enormously valuable. With any modern call center software application all data of a service request handled by a call center is being recorded into the database. According to some authors from the researched literature the truth may lay within the databases.

To complete and secure good modeling it always remains important to follow the verification and validation steps necessary to obtain a reliable representation of reality.

7.5 Limitations and Learning Experiences

Limitations

The sections of future research and managerial implications already clarify the fact that there are no ready-to-use and –implement tools for the dilemma of determining optimal staffing levels. Partly due to the predetermined scope of the literature research and the available time to carry out the research, I am very much aware of the incompleteness of the total amount of scientific literature I studied, researched and discussed in this literature report. Though the authors I dealt with in the different sections of this report are referred to a lot in many scientific articles dealing with this subject, so a certain scientific value can be attributed to the summarized approaches. The table overview with existing tools to analyze and solve staffing problems for call centers thus is far from complete, but provides a good representation of known approaches.

As stated before a lot of the models and approaches from different authors dealt with in this report are far from real, because of too many simplifications and assumptions. Especially the lack of possibilities to measure quality of a process is a limitation of existing techniques. Next to this, usually no sophisticated characteristics of a call center are modeled. For example, SBR, overflow strategies, call routing between different locations, integration of outbound activities and integration of outsourcing to providers. Again, this is partly caused by the inexperience of working with database data and (more important) database information.

Summarizing: some important things to bear in mind when choosing tools to solve staffing problems in call centers.

Furthermore a number of authors warns for the expensive and time-consuming simulation techniques and models. Nowadays though some simulation tools are on the market that are relatively easy to use and with the progress in computer technology, time (to run scenarios and simulations) should not be a problem.

Another weak point in all of the researched literature is, in our opinion, the lack of use of other distributions than Erlang C and A. Some authors mention another distribution but do not do anything to prove the goodness-of-fit of such a distribution.

Learning experiences

Reading and analyzing scientific literature on call center processes and issues is very useful when experiencing the practice of a real call center. Comparison between practice and literature is very helpful when studying the characteristics,

agent types and managerial considerations in a real call center. The processes and specific characteristics of a call center become more visible.
Moreover some important aspects arose in this survey:

- Managerial approach towards designing a model of an existing call center process by following three identified relevant activities towards determining optimal staffing levels.
- Use of historic (e.g. ACD) data with forecasting input parameters.
- Recognizing the importance of uncertainties, lack of good data, corrections, (un)predictability of events, presence of trends (over time).
- Verifying and validating models of call center processes (simulation and queueing).

References

Aksin OZ. and Harker PT. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. European Journal of Operational Research 2003; 147(3):464–483.

Andrews B and Parsons H. Establishing telephone-agent staffing levels through economic optimization. Interfaces 1993; 23(2):14–20.

Brand N, van der Kolk H. Workflow analysis and design. Deventer: Kluwer Bedrijfswetenschappen 1995 [in Dutch].

Bapat V and Pruitte Jr EB. Using simulation in call centers. 1998 Winter Simulation Conference. Proceedings, IEEE, Piscataway, NJ, USA 1998; 1395–1399.

Becker KJ, Gaver DP, Glazebrook KD, Jacobs PA, Lawphongpanich S. Theory and Methodology. Allocation of tasks to specialized processors: A planning approach. European Journal of Operational Research 2000; 126:80-88.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L. Statistical analysis of a telephone call center: A queueing science perspective. Working paper, The Wharton School, University of Pennsylvania, Philadelphia. Electronically available as iew3.technion.ac .il/serveng/References/callcenter.pdf, 2002.

Chan DYK. Design an effective workflow in simulation. The 8th international conference on computer supported cooperative work in design proceedings. IEEE 2003.

Chen BPK and Henderson SG. Two issues in call centre staffing levels. Annuals of Operations Research 2001; 108:175-192.

Dawson K. The call center handbook: The complete guide to starting, running and improving your call center. San Francisco: CMP books, 2004.

Deslauriers A, L'Ecuyer P, Pichitlamken J, Ingolfssony A, Avramidis N. Markov chain models of a telephone call center with call blending. Technical report, GERAD and DIRO, University of Montreal, Canada 2004. Electronically available as
www.iro.umontreal.ca/~lecuyer/myftp/papers/ctmc1.pdf

de Véricourt F, Zhou Y-P. Managing Response Time in a Call Routing Problem with Service Failure. Working paper, Fuqua School of Business, Duke University, Durham, North Carolina, USA. Electronically available as faculty.washington.edu/yongpin/abs_callback.pdf, 2004.

Ernst AT, Jiang H, Krishnamoorthy M and Sier D. Staff scheduling and rostering: A review of applications, methods and models. European Journal of Operational Research 2004; 153:3–27.

Gans N, Koole G and Mandelbaum A. Commissioned Paper Telephone Call Centers: Tutorial, Review, and Research Prospects. Manufacturing & Service Operations Management 2003; 5(2): 79–141.

Garnett, O., A. Mandelbaum. An introduction to skillsbased routing and its operational complexities. Teaching note, Technion, Haifa, Israel. Full version available upon request, from AM. Electronically available as iew3.technion.ac.il/serveng/Lectures/lectures.html, 2001.

Grossman TA, Oh SL and Rohleder TR and Samuelson DA. Call centers. University of Calgary, Canada. Kluwer online reference works. Electronically available as reference.springerlink.com.janus.libr.tue.nl/kapxsl.asp?xmlid=079237827x/c_sec2, 2001

Hasija S, Pinker EJ and Shumsky RA. Staffing and Routing in a Two-Tier Call Center. Simon School, University of Rochester, Rochester, New York 2005.

Henderson SG and Mason AJ. Rostering by Iterating Integer Programming and Simulation. Proceedings of the 1998 Winter Simulation Conference 1998.

Hillier FS and Lieberman GJ. Introduction to Operations Research. Sixth Edition, McGraw-Hill International Editions, Industrial Engineering Series, 1995.

Jongbloed, G and Koole, GM. Managing uncertainty in call centers using Poisson mixtures. Applied Stochastic Models in Business and Industry 2001; 17:307-318.

Koole GM. Call Center Mathematics: A scientific method for understanding and improving contact centers. Vrije Universiteit, Department of Mathematics, Amsterdam, The Netherlands. Electronically available as www.math.vu.nl/~koole/ccmath, 2005.

Koole GM and Mandelbaum A. Queueing Models of Call Centers: An Introduction. Annals of Operations Research 2002; 113:41–59.

Koole GM and Van der Sluis E. Optimal shift scheduling with a global service level constraint, IEE Transactions on Scheduling and Logistics 2000; 35:1049-1055.

Lin CKY, Lai KF and Hung SL. Development of a workforce management system for a customer hotline service. Computer & Operations Management 2000; 27:987-1004.

Mandelbaum A. Call Centers (Centres): Research Bibliography with Abstracts. Electronically available as ie.technion.ac.il/serving, 2004.

Mazzuchi, TA and Wallace, RB. Analyzing skill-based routing call centers using discrete-event simulation and design experiment. Proceedings of the 2004 Winter Simulation Conference, 2004.

Massey WA, Parker GA, Whitt W. Estimating the parameters of a non-homogeneous Poisson process with a linear rate. Telecommunication Systems 1996; 5:361–388.

Mehrotra, V. Ringing Up Big Business. OR/MS Today 1997; August:18-24.

Mehrota V and Fama J. Call center simulation modeling: methods, challenges, and opportunities. Proceedings of the 2003 Winter Simulation Conference, 2003.

Montgomery DC and Runger GC. Applied statistics and probability for engineers. Second edition. John Wiley & Sons, Inc, 1999.

Paprzycki M, Abraham A, Guo R, and Mukkamala S. Data Mining Approach for Analyzing Call Center Performance. The 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Canada, Springer Verlag, Germany, 2004. LNAI 3029: 1092-1101.

Pinker EJ and Shumsky RA. The Efficiency-Quality Trade-Off of Cross-Trained Workers. Manufacturing & Service Operations Management 2000; 2(1):32-48.

Pichitlamken J, Deslauriers A, L'Ecuyer P, and Avramidis AN. Modelling and simulation of a telephone call center. Proceedings of the 2003 Winter Simulation Conference, 2003.

Reijers HA and Limam Mansar S. Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics. Omega 2005: 33:283 – 306.

Shumsky, RA, and Pinker, EJ. Gatekeepers and Referrals in Service, Management Science, 2003; 49(7), 839—856.

Stolletz R. Performance Analysis and Optimization of Inbound Call Centers. Lecture notes in economics and mathematical systems 528, Springer-Verlag Berlin Heidelberg, 2003.

Stolletz R and Helber S. Performance analysis of an inbound call center with skills-based routing: A priority queueing system with two classes of impatient customers and heterogeneous agents. OR Spectrum 2004; 26:331–352.

Tanir O, and Booth RJ. 1999. Call center simulation in Bell Canada. Proceedings of the 1999 Winter Simulation Conference, IEEE Press. Electronically available as informs-cs.org/wsc99papers/237, 1999.

Wallace, RB. Performance Modeling and Design of Call Centers with Skill-Based Routing, D.Sc. Dissertation (summary), The George Washington University 2004.

Wallace. RB. and Whitt W. Resource Pooling and Staffing in Call Centers with Skill-Based Routing. Submitted to Operations Research. Electronically available as www-2.cs.cmu.edu/~harchol/WORMS04/people/whitt/whitt.pdf, 2004.

Zapf, M. From the customer to the firm: evaluating generic service process designs for incoming customer requests. Computers in Industry 2004; 55:53–71.

Appendices

Appendix A: Devil's Quadrangle

Brand and Van der Kolk (1995) distinguish four main dimensions in the effects of redesign measures: time, cost, quality and flexibility. Ideally, a redesign of a business process decreases the time required to handle an order, it decreases the required cost of executing the business process, it improves the quality of the service delivered and it improves the ability of the business process to react to variation. The attractive property of their model is that, in general, improving upon one dimension may have a weakening effect on another. For example, reconciliation tasks may be added in a business process to improve on the quality of the delivered service, but this may have a drawback on the timeliness of the service delivery. To signify the difficult trade-offs that sometimes have to be made they refer to their model as the devil's quadrangle.

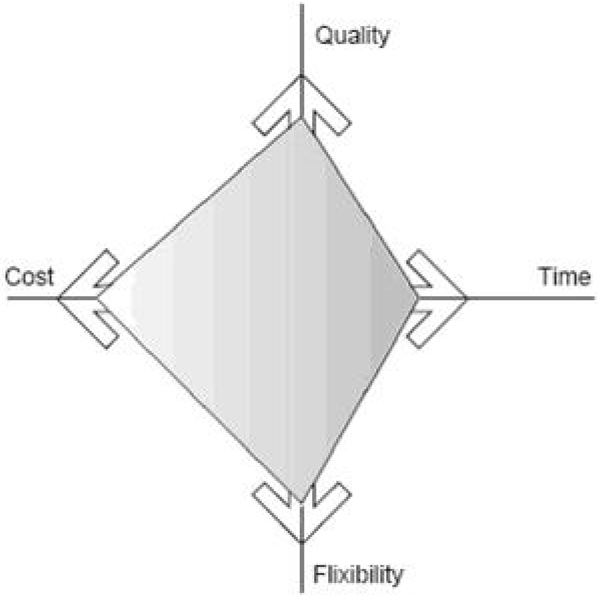


Figure 11: the devil's quadrangle according to Brand and Van der Kolk (1995)

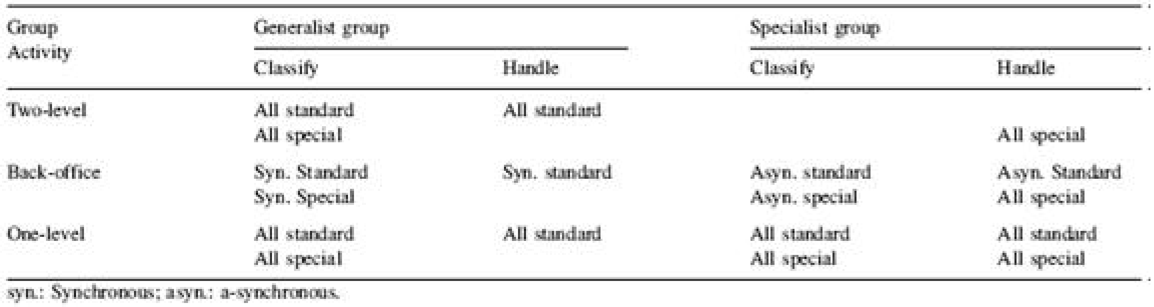
Appendix B: Call center configurations based on Zapf (2004) dimensions

Zapf (2004) distinguishes two very important characteristics he uses to describe call center processes:

- the level of difficulty (qualification-mixture dimension, standard vs. special requests, already described before)
- the communication channel (synchronous vs. a-synchronous).

In table 2 one can find an overview of possible tasks of generalists and specialists in various call center configurations (two- or one-level and back-office).

Table 1: activities and agent groups



Group Activity	Generalist group		Specialist group	
	Classify	Handle	Classify	Handle
Two-level	All standard All special	All standard		All special
Back-office	Syn. Standard Syn. Special	Syn. standard	Asyn. standard Asyn. special	Asyn. Standard All special
One-level	All standard All special	All standard	All standard All special	All standard All special

syn.: Synchronous; asyn.: a-synchronous.

Following section 2.3, in figures 12 to 15 more possible types of configurations of a call center process can be found. A short description is provided with the figures.

Figure 12 represents the situation that no separate groups for asynchronous service requests are created. Generalists for classifying and handling standard requests that come in through any type of media. Specialists for handling special requests that come in through any type of media. The special requests are first classified by generalists.

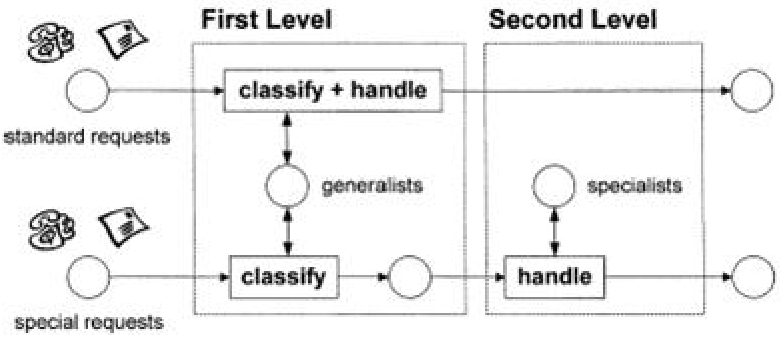


Figure 12: Two-level design

Figure 13 gives a representation of a call center with separated communication channels. The back-office is occupied by specialists. Requests that come in by (for example) mail are classified and handled by the back-office. Generalists only classify and handle standard requests and classify special requests that come in by telephone.

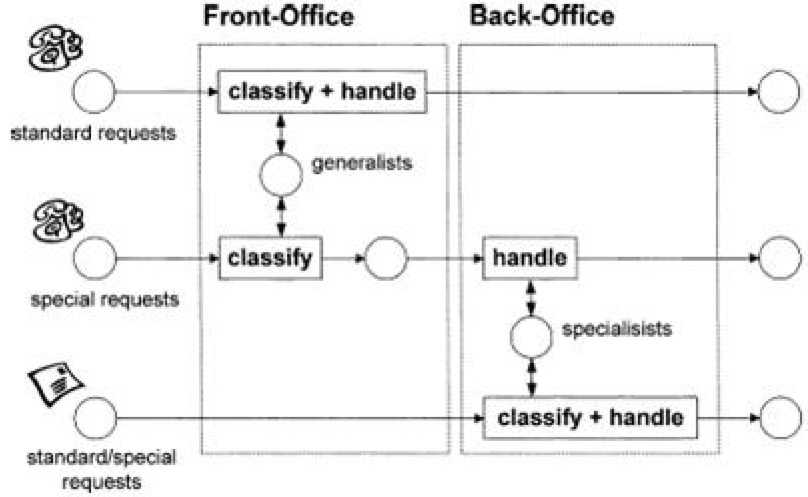


Figure 13: Back-office design

The strongest integration of qualification (standard and special service requests) groups is realized within the one-level design in figure 14. In this design first and second level (or front office and back-office) will not be distinguished. In this situation generalists and specialists are both able to classify and handle all types of requests. Since most specialists are more expensive than generalists, requests are primarily assigned to a free generalist (priority 1). Only if no generalist is available the request will be assigned to a specialist (priority 2).

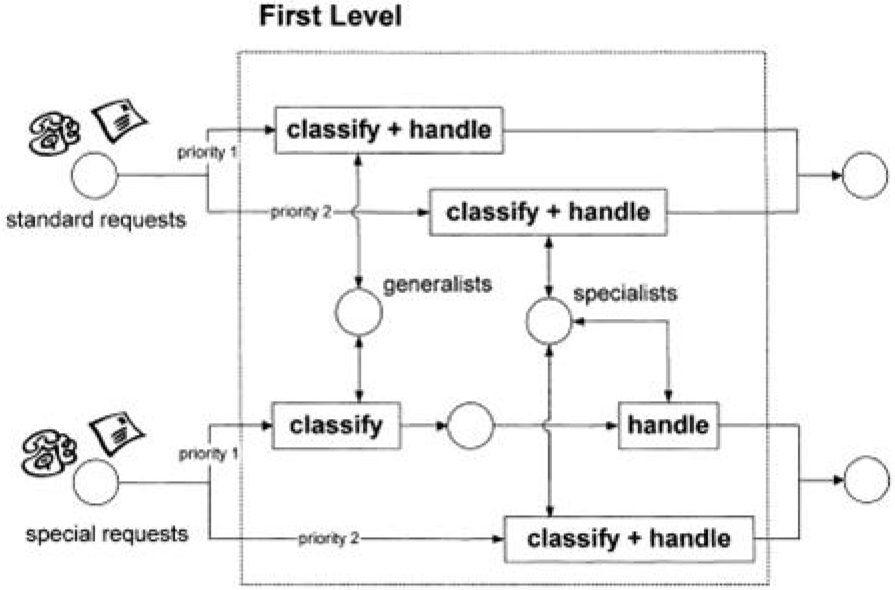


Figure 14: One-level design

In the two-level design (figure 15) two groups ‘‘generalists 1’’ and ‘‘specialists 1’’ will be established for synchronous requests and two additional groups ‘‘generalists 2’’ and ‘‘specialists 2’’ will be entrusted with asynchronous requests. Since in the back-office design generalists handle only synchronous requests there is no difference between integrated and separated channels in the front office (figure 13). In the back-office an additional group has to be defined for separated communication channels, so that synchronous requests are handled by the group ‘‘specialists 1’’ and a-synchronous requests by the group ‘‘specialists 2’’. For the separated version of the one-level design four groups (generalists 1, generalists 2, specialists 1, specialists 2) have to be built similar to the two-level pattern. The basic routing strategy remains the same as shown in figure 14.

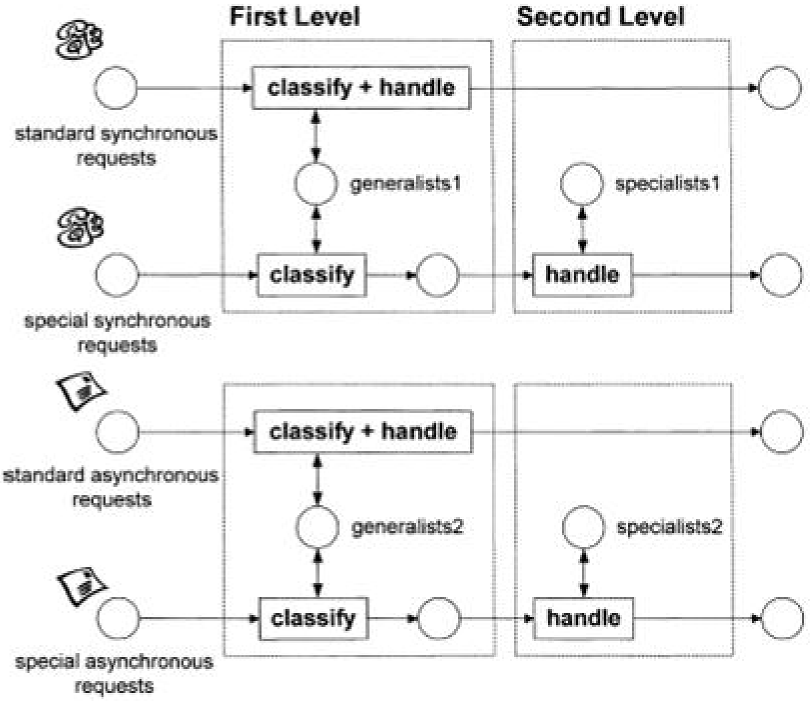


Figure 15: Two-level design with separated communication channels

The handling of a synchronous request is not finished with the completed call. Additional after-call
work has to be done afterwards by the agent (figure 16). This work comprises administrative tasks like data entry and necessary internal communication. In Zapf’s model only this part of the after-call work is included which is done immediately by the agent who has handled the call. After finishing this work the agent is available for accepting further requests.

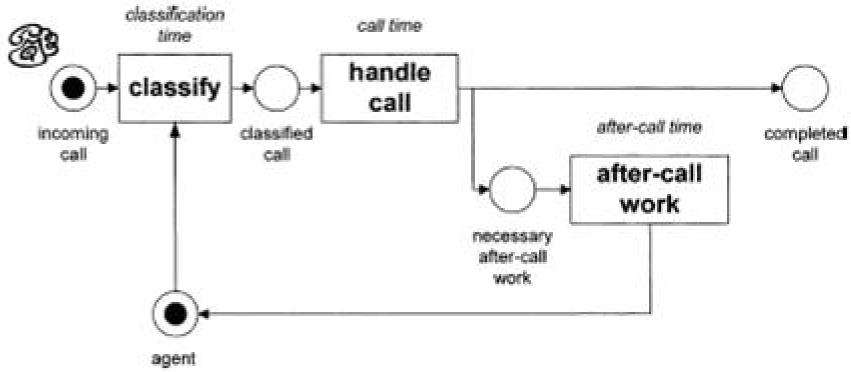


Figure 16: Request classification and handling for synchronous requests without forwarding

In the case of synchronous requests customers do not wait for an agent as long as you like but only for a particular time period. If calls are in the waiting queue for a longer time the customer hangs up (figure 17). We call this period the waiting tolerance which is determined by the customers preferences and his current situation. After hanging up some customers re-dial in order to get a free agent. The part of re-dialers is represented by the percentage of re-dialing. The parameter time between dial attempts defines the time interval between hanging up and re-dialing.

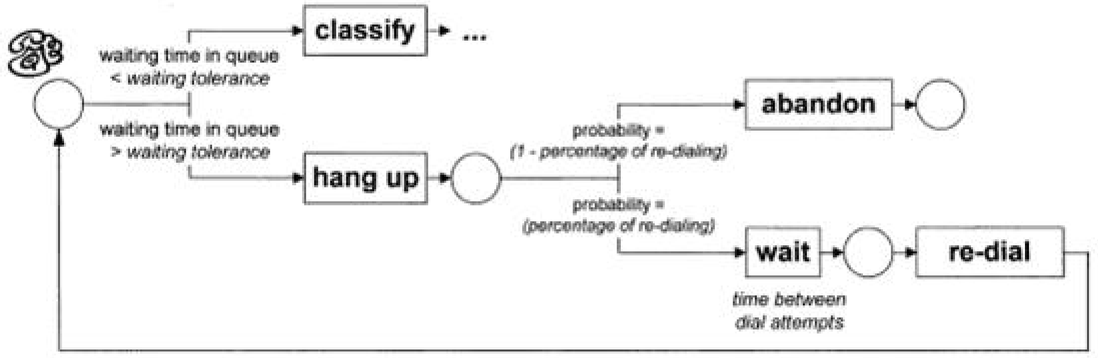
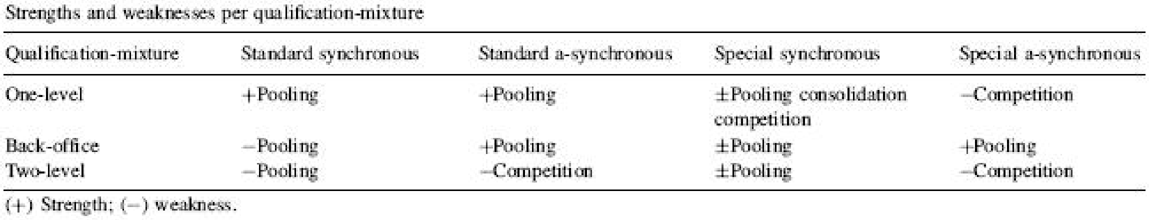


Figure 17: Call abandonment process for synchronous requests

Table 2 gives a rough summary of the identified strengths and weaknesses per qualification-mixture. The one-level design (figure 14) is very good for handling standard requests because of pooling generalists and specialists. The design has weaknesses for synchronous special requests in overload situations and for special a-synchronous requests since specialists are additional occupied with standard requests. The back-office design (figure 13) has strengths in handling asynchronous requests since specialists are reserved for these requests. Classify synchronous requests is the weakness of the back-office design because of a small generalist group. The two-level design (figure 12) does badly for synchronous and a-synchronous requests.

Table 2



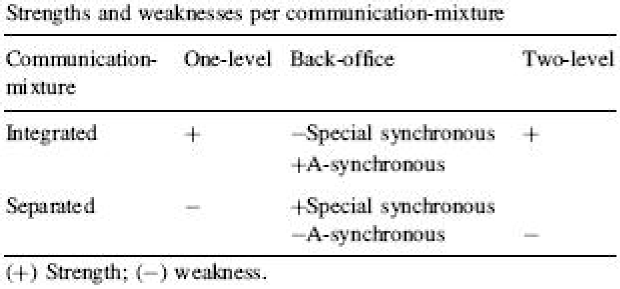
Strengths and weaknesses per qualification-mixture

Qualification-mixture	Standard synchronous	Standard a-synchronous	Special synchronous	Special a-synchronous
One-level	+Pooling	+Pooling	±Pooling consolidation competition	−Competition
Back-office	−Pooling	+Pooling	±Pooling	+Pooling
Two-level	−Pooling	−Competition	±Pooling	−Competition

(+) Strength; (−) weakness.

The integration of communication channels is in most cases more efficient than the separation (table 3). Only in the back-office design (figure 13) the integration of communication channels leads to worse performance for special synchronous requests which can be explained through task competition between handling synchronous and a-synchronous requests.

Table 3



Strengths and weaknesses per communication-mixture

Communication-mixture	One-level	Back-office	Two-level
Integrated	+	−Special synchronous +A-synchronous	+
Separated	−	+Special synchronous −A-synchronous	−

(+) Strength; (−) weakness.

Appendix C: Canonical designs for SBR

In section 2.2.2 the principles of SBR are introduced. To visualize the different possible routing policies this appendix provides some canonical designs, which are also meant for typology and control simplification in modeling call centers and routing policies.

Topology Simplification. The first means of reduction is to consider simple special network topologies such as those shown in Figure 18. These configurations represent building blocks for more complex systems. For example, in a "V" design a single pool of agents handles two (or more) types of calls. In a "W" design, two pools of agents cater to three types of calls: Pool 1 serves Types 1 and 2; Pool 2 serves Types 2 and 3.

The "X" design, in which two types of calls can be served by either of two pools of agents, represents full flexibility. It also reflects the fact that skill groups may be defined on a relative, rather than absolute, basis. For example, an X-design arises when CSR Pool 1 is assigned call Type 1 as a "primary skill," CSR Pool 2 is assigned call Type 2 as primary, and both pools have the other type of call assigned as secondary. A pool takes "secondary skill" calls only when deemed necessary: Say, only if it has idle CSRs and the other pool is congested. In this case, skills-based routing captures the fact that different type-to-pool assignments have differing (perhaps implicit) costs or rewards. It is also important to note that the same network topology can be used quite differently, given various levels of traffic and routing schemes. For example, an "N" design can be used when Type-1 customers are VIP but there are not enough specialized Pool-1 CSRs to serve them. In this case, Pool-2 CSRs can contribute to maintaining an adequate service level for Type 1s. Conversely, the same N-design can be used when Type-2 customers are VIP and Pool-2 capacity is in excess. Here, acceptable resource efficiency can be maintained by routing Type-1 calls to idle Pool-2 CSRs.

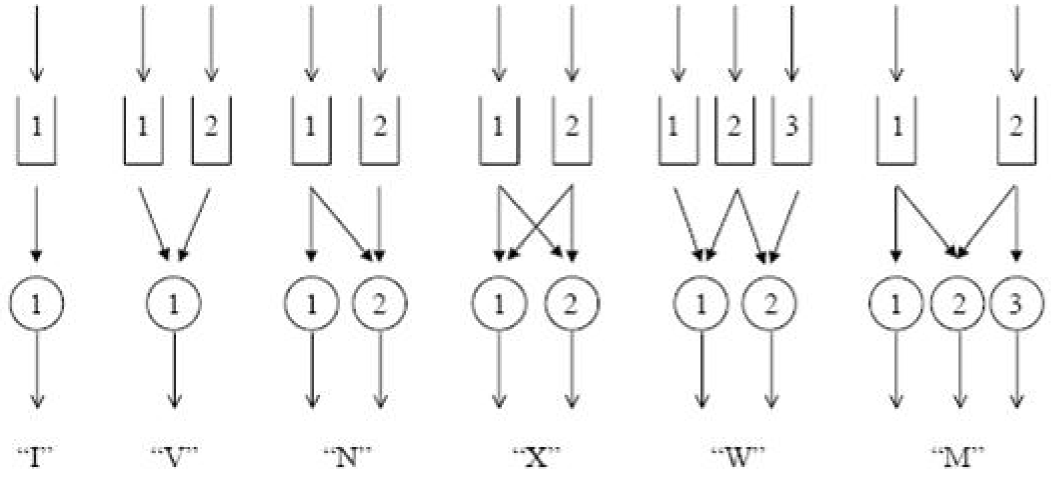


Figure 18: different canonical designs for SBR

Figure 19 is an example of a system with an elaborate skills-based routing structure. In it six types of calls are routed to five pools of agents. All agents in a pool can handle the same set of call types; equivalently it is said that the agents within a pool have the same skills. Arrows between call types and agent pools describe the various pools' skills. Dashed arrows at the sides of queues represent customer abandonment. (Note that the nomenclature for skills-based routing has not yet become standardized. For example, one ACD manufacturer refers to customer types as "skills" and to agents with the same skills as having the same "skill set.")

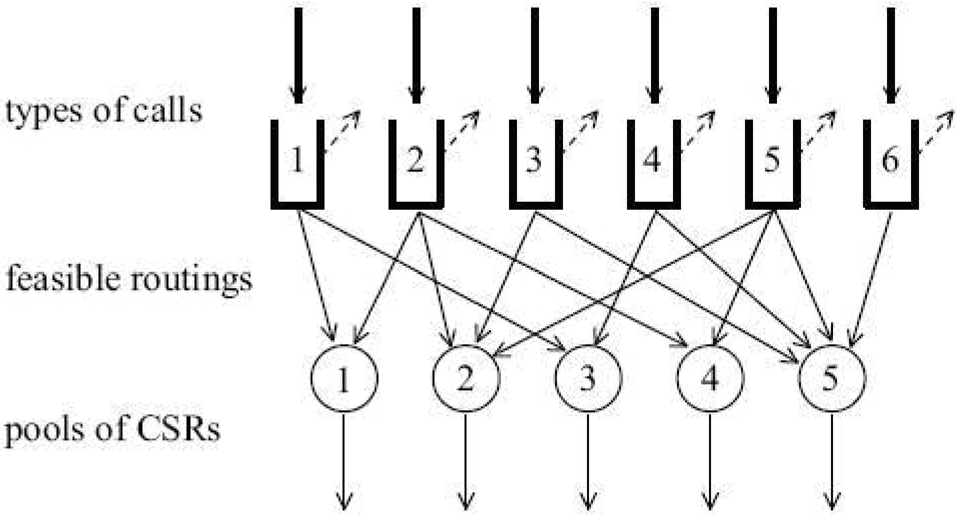


Figure 19: An Example of Skills-Based Routing

Appendix D: Hierarchical view of arrival rates

Over short periods of time, minute-by-minute for example, there is significant stochastic variability in the number of arriving calls. Over longer periods of time—the course of the day, the days of the week or month, the months of the year - there also can be predictable variability, such as the seasonal patterns that arriving calls follow (figure 20).

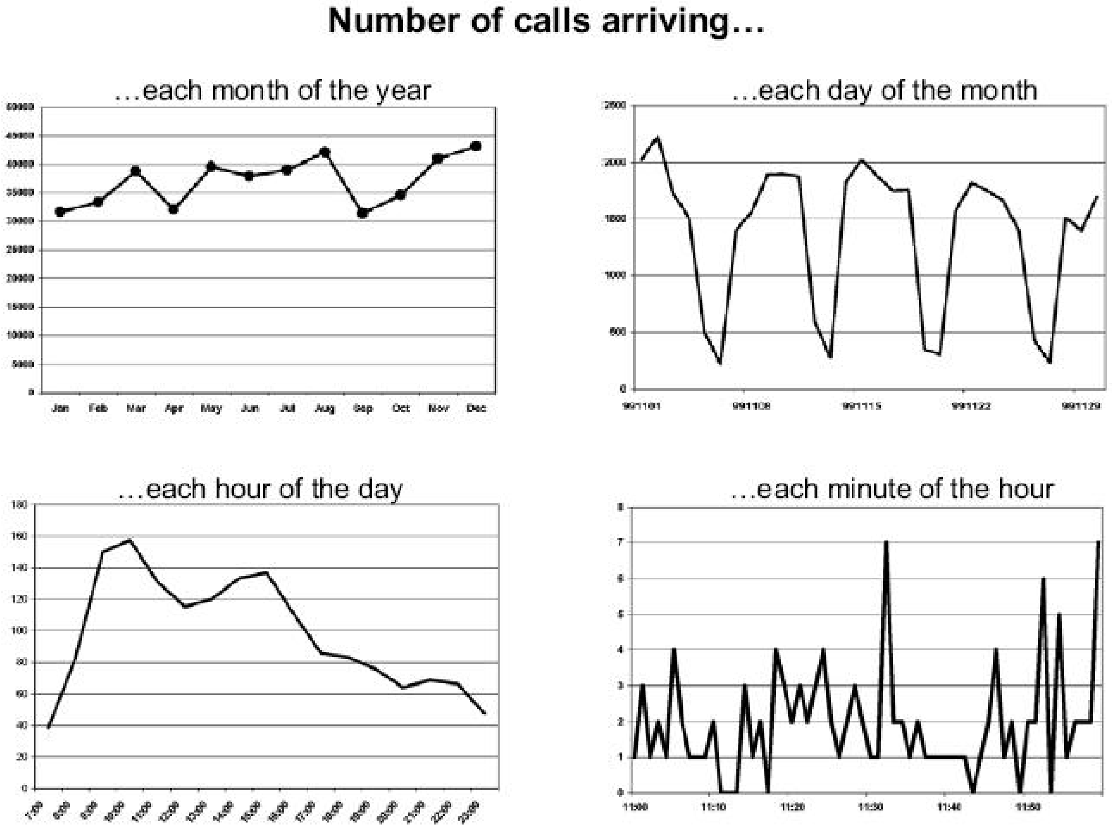


Figure 20: A Hierarchical View of Arrival Rates (Gans, Koole and Mandelbaum, 2003)

At the lowest level of the hierarchy, the arrival times of individual calls are not predictable (lower right panel of figure 20). Here, common practice uses the M|M|N (Erlang C) queueing model to estimate stationary system performance of short - half-hour or hour - intervals.

Appendix E: Mazzuchi and Wallace (2004) Performance measurements

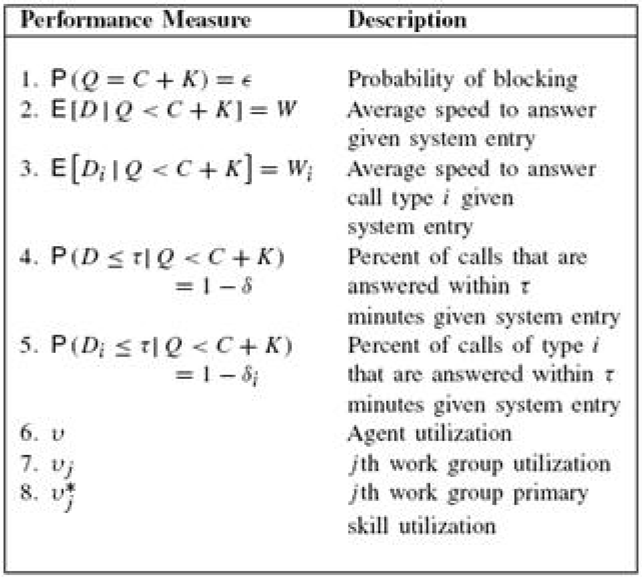
Mazzuchi and Wallace (2004) provide a list with some commonly used performance metrics.

- the probability that an arriving caller is blocked
- speed-to-answer performance measures
- tracking agent's utilization

The first two performance metrics are usually included in the Service Level Agreement. The last performance metric normally contains several sub-measurements. In this appendix some more detailed information is provided about the specific formulas of the performance metrics.

Q is number of callers in system
D is aggregate delay experienced by caller (calltype i)

Table 4: performance measurements according to Mazzuchi and Wallace (2004)



Performance Measure	Description
1. $P(Q = C + K) = \epsilon$	Probability of blocking
2. $E[D \mid Q < C + K] = W$	Average speed to answer given system entry
3. $E\left[D_i \mid Q < C + K\right] = W_i$	Average speed to answer call type i given system entry
4. $P(D \leq \tau \mid Q < C + K) = 1 - \delta$	Percent of calls that are answered within τ minutes given system entry
5. $P(D_i \leq \tau \mid Q < C + K) = 1 - \delta_i$	Percent of calls of type i that are answered within τ minutes given system entry
6. υ	Agent utilization
7. υ_j	jth work group utilization
8. υ_j^*	jth work group primary skill utilization

The first performance metric, the probability that an arriving caller is blocked, is a measure of the call center's availability and is sometimes apart of the service level agreements (SLAs). The second parameter E[D |Q < C + K] and the fourth P(D ≤ τ |Q < C + K) are speed-to-answer performance measures and are typically apart of the service levels as well. These two aggregate quantities are conditioned given admission or entry into the system. Usually, one of the two and not both speed-to-answer metrics is apart of the SLA. Average speed to answer (ASA) is the call center term reserved for E[D |Q < C + K]. Both Speed-to-answer and availability SLAs drive staffing and equipment (trunk lines) requirements. Massey and Wallace (2004) developed asymptotic-based algorithms to determine optimal (C,K) in a M|M|C|K queue while holding the SLAs for blocking and the conditional probability of delay fixed.

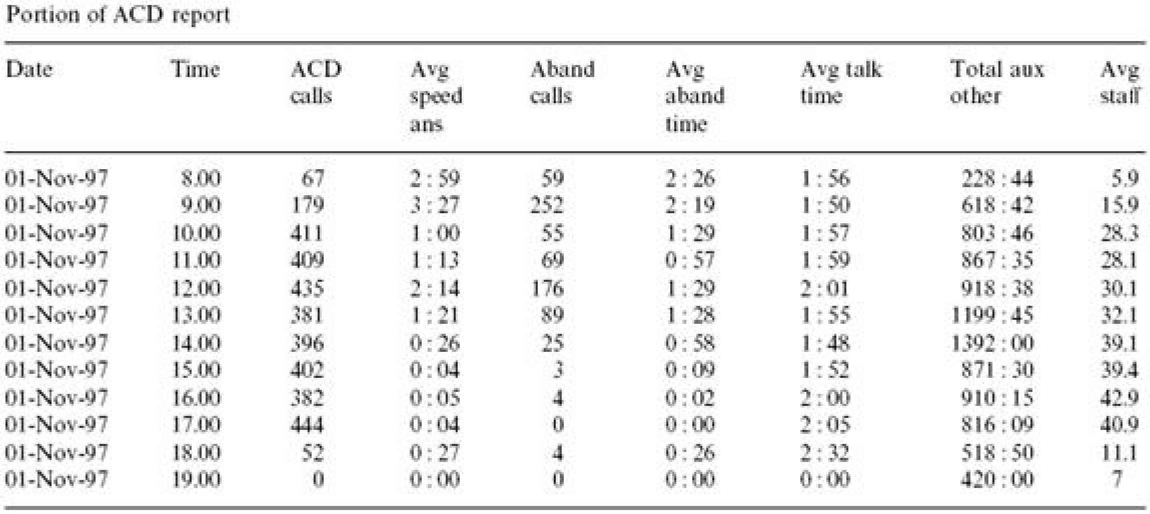
The last three measures of performance deal specifically with tracking agent's utilization. The average utilization for an agent is the percent of time that he/she is busy processing calls or one minus the fraction of time he/she is idle.

Appendix F: Lin, Lai and Hung (1998)

The authors' case deals with a 24H hotline service. The authors use an integrated approach, on a monthly basis, in which one of the final stages towards scheduling is the activity that implies the use of historical data, the choice for call center configuration and other information (figure 21). A regression model leads to a simulation model, with outcomes that can be used to make decisions for rostering and scheduling (heuristics).

Lin, Lai and Hung (1998) use a very simple performance measure. By using the ACD report (table 5) the system performance is measured by means of the call abandonment rate, defined by (aband calls)/(ACD calls + aband calls) , where aband calls and ACD calls denote the numbers of abandoned calls (after entering the queue) and completed calls, respectively.

Table 5: an example of an ACD report



Portion of ACD report

Date	Time	ACD calls	Avg speed ans	Aband calls	Avg aband time	Avg talk time	Total aux other	Avg staff
01-Nov-97	8.00	67	2 : 59	59	2 : 26	1 : 56	228 : 44	5.9
01-Nov-97	9.00	179	3 : 27	252	2 : 19	1 : 50	618 : 42	15.9
01-Nov-97	10.00	411	1 : 00	55	1 : 29	1 : 57	803 : 46	28.3
01-Nov-97	11.00	409	1 : 13	69	0 : 57	1 : 59	867 : 35	28.1
01-Nov-97	12.00	435	2 : 14	176	1 : 29	2 : 01	918 : 38	30.1
01-Nov-97	13.00	381	1 : 21	89	1 : 28	1 : 55	1199 : 45	32.1
01-Nov-97	14.00	396	0 : 26	25	0 : 58	1 : 48	1392 : 00	39.1
01-Nov-97	15.00	402	0 : 04	3	0 : 09	1 : 52	871 : 30	39.4
01-Nov-97	16.00	382	0 : 05	4	0 : 02	2 : 00	910 : 15	42.9
01-Nov-97	17.00	444	0 : 04	0	0 : 00	2 : 05	816 : 09	40.9
01-Nov-97	18.00	52	0 : 27	4	0 : 26	2 : 32	518 : 50	11.1
01-Nov-97	19.00	0	0 : 00	0	0 : 00	0 : 00	420 : 00	7

The total calls (= Aband calls + ACD calls) provide forecasts for future call volume. Based on discussion with management and supervisors, the forecast is made on an hourly basis for the 7 days in an average week, as call traffic and talk time reflect hourly and weekday differences. The forecasting model adopted is the simple but efficient 3-month moving average (for each hour in the week) as no particular trend was observed. In choosing an appropriate model to relate service level with the system parameters, the regression model of abandonment rate (y_{ih}) on workload (x_{ih}) was found to give high correlation (R^2 over 0.7) in more than 100 hrs (out of 24x7=168 hrs) in the week. These occur mostly in daytime and evening hours when call traffic is significant. Hence for each hour h (h=0,…, 23) on weekday i (i=1,…,7), we first examine the following linear relationship in the recent 3 months' data:

$$y_{ih} = m_{ih} x_{ih} + c_{ih} + \varepsilon \qquad (2)$$

where

$$y_{ih} = (aband\ calls)_{ih} / (total\ calls)_{ih}$$

$$x_{ih} = workload = demand\ volume / service\ capacity = \frac{(totals\ calls)_{ih} / 60\min}{(net\ staff\ level)_{ih} / (Avg\ talk\ time)_{ih}}$$

and m_{ih}, c_{ih} are regression coefficients estimated from the least-squares method. If this model is adequate for hour h on weekday i (R^2 over 0.7), we can set y_{ih} at the target abandonment rate (8% between 8:00-24:00, 24:00-1:00 and 15% between 1:00-8:00) and directly evaluate the net staff level required by

$$net\,staff\ level = \frac{m_h \overline{\left(total\ calls\right)_{ih}} / 60}{\left(y_{ih} - c_{ih}\right) / \overline{\left(Avg\ talk\ time\right)_{ih}}}, \qquad (3)$$

where $\overline{\left(total\ calls\right)_{ih}}$ and $\overline{\left(Avg\ talk\ time\right)_{ih}}$ represent the 3-month moving average for the specific hour and weekday of interest.

For those hours (about 50-60) showing poor correlation in Equation (2), or hours which have insufficient past data for constructing the regression model, queuing simulation model is applied to evaluate the net staff level. The present network structure approximates an M/G/c/K queuing model with abandoned calls. As the uncapacitated M/G/c model with no abandoned calls already involves complex integral equations, a single-stage queuing simulation model was adopted for this more complicated case. The random factors in the simulation model include Poisson call arrivals, normally distributed talk time and abandonment time. In estimating the input parameter of average abandonment time, it is observed that a positive linear relationship may sometimes exist between Avg aband time (Table) and workload as in Equation (2), but with y_{ih} denoting Avg aband time for hour h on weekday i. Hence, a linear model is first tested for its adequacy (R^2 over 0.7). If R^2 is less than 0.7, the 3-month moving average of Avg aband time from the ACD report is used for estimation. (Note that if the linear model is valid, the predicted average abandonment time is recalculated from Equation (2) when the net staff level changes in the simulation runs.) In the simulation model, an abandoned call is defined as one whose waiting time is longer than its generated abandonment time. The net staff level is initially taken as an integer smaller than the average over the recent 3 months (e.g. average net staff level - 5). The average call abandonment rate is recorded over 30 replications to be compared with the target service level. The net staff level would be incremented by one for every 30 replications until the service level is satisfied. The minimum net staff level is thus obtained.

From past observation, product promotion and public holidays would affect the call traffic. The system also allows user-input of a call adjusting factor (r) on particular hours or days specified. Accordingly, the minimum net staff level required would either be calculated directly from Equation (3) by multiplying with the factor r, or by performing simulation runs with the adjusted total calls for those hours or days of concern.

The outcomes of the methodology as mentioned above are hourly forecasts of call traffic and the minimum (net) manpower requirement on everyday of the month.

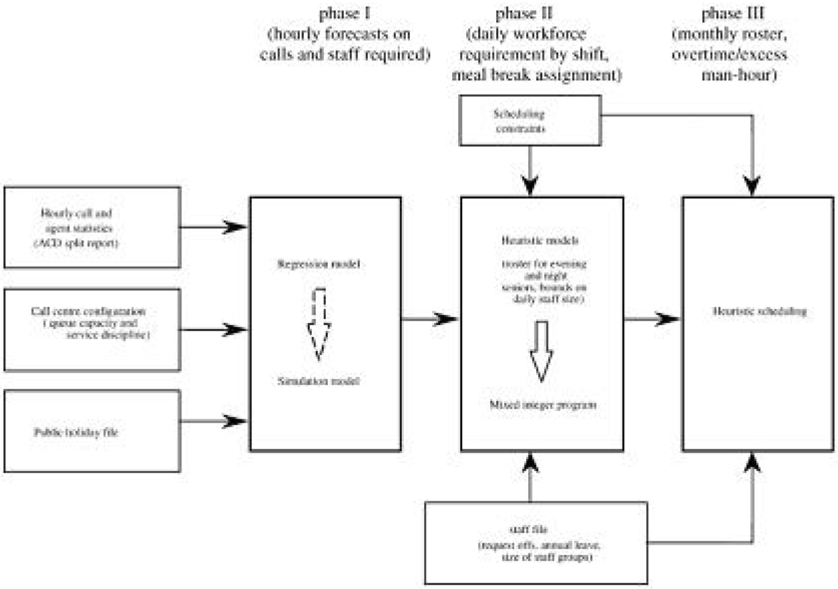


Figure 21: Logic flow of the workforce management system

Appendix G:Chan (2003)

Based on the identified Key Output Performance Variables (KOPV) and the Key Input Performance Variables (KIPV) by Chan (2003) a number of activities can be performed to design an effective workflow for call centers using simulation tools. Simulation enables dynamic analysis, which is necessary to analyze processes. A list of steps in the design process is provided:

The following are steps in designing an effective workflow using simulation.

1. Identify the key output performance variables (KOPV) of the process.
2. Identify the key input performance variables (KIPV) that affect the KOPV.
3. Determine the sequence of work and document it using a flowchart.
4. Collect and establish the process and resource parameters that will have impact on the KOPV defined such as inter-arrival time, cycle time, resource availability, effective operation hours, cost of operation and others.
5. Establish the constraints and limitations.
6. Make a model of the workflow and input all information above using a process-simulation software and conduct simulation on the model planned.
7. Validate the accuracy of the model with experts of the process or management and make modification on model till the model is validated.
8. Analyze the KOPV obtained from the simulation of the initial design to see if the model meets the requirement.
9. Make changes on the process parameters ("what-if' analysis) to find out if there are better the alternatives to meet the KOPV.

Select the most effective workflow from the various possible options. Effectiveness depends on the KOPVs identified. These are usually quantitative performance measurements.

Appendix H: Zapf (2004), validation and verification steps

The suitability of the simulation model can be checked in the following validation and verification steps:

1. Conceptual model validation: determine that the conceptual model is reasonable and correct for the intended application.
2. Computerized model verification: ensure that the computer programming and implementation of the model is correct.
3. Data validity: ensure that data is appropriate, accurate and sufficient.
4. Operational validity: determine that the results are sufficient accurate for the intended purpose over the application domain.

Ad 1. For the conceptual model validation the face validity technique can be used. The generic service process designs and the model details should be developed and discussed with communication center experts in order to ensure that the models are reasonable. For building the conceptual model also different real communication centers of the domains bank, book trade, car rental and energy industry have to be analyzed.

Ad 2. After face validation, which was mainly based on graphical process models and verbal process descriptions, the process logic has to be checked through trace technique. The different request types have to be tracked through every submodel to determine whether the logic is correct and the necessary accuracy is maintained.

Different dynamic testing techniques have to be applied for computerized model verification and the simulation models have to be executed under various conditions:

1. Fixed values: fixed values (constant factors) have to be defined for selected input variables (e.g. classification time, handling times, arrival rates) and the performance values have to be checked against hand calculated values.
2. Comparison to other models: sub-models have to be compared to analytical M/M/l and M/M/n queueing models.
3. Sensitivity analysis: selected input parameters (e.g. after-call time, percentage of re-dialing) have to be modified and the effect upon the results has to be determined.

Ad 3. To ensure data validity real data is used which should be collected automatically by the ACD system (Automatic call distribution) for the average request volume and average handling times of synchronous request. The stochastic distribution for the handling times has to be derived from the empirical data and statistically validated with the Chi-2-test and the test of Kolmogorov/Smirnov. The other data also has to be logged by the ACD system or provided by experts. Within the computerized model verification sensitivity analysis was used to ensure that small changes of these parameters are not critically for the performance measurement.

Ad 4. Regarding the operational validity 95% confidence intervals have to be calculated for every performance measure. In order to reflect the nature of a communication center the single experiments have to be performed in the form of multiple terminating simulation runs. For every experiment 30 independent replications have to be made according to a general rule. At the end of each replication it has to be checked that sufficient data has been collected and that the output data is not correlated.

Appendix I: Hasija et al. (2005), treatment threshold k_d

Hasija et al. (2005) researched sensitivity of the models and processes used. The cost-minimizing referral rate varies with changes in parameters related to queuing, i.e., arrival rates and service rates. When the arrival rate increases, the optimal referral rate converges to the optimal referral rate (for the deterministic case). For very large λ, waiting costs are relatively small, compared to the sum of staffing and mistreatment costs. Therefore it is optimal to use the treatment threshold from a deterministic model, which only considers staffing and mistreatment costs. The authors also provide a rule of thumb for choosing the optimal system. Treatment threshold k_d is the main parameter the rule is based on. This appendix includes a more broadly description of the rule and of the methodology followed by Hasija et al. (2005). The authors found that with a certain skill level of the gatekeeper a direct access system (only experts) is optimal. So, to justify gatekeepers (assumed high waiting costs) they should have high level skills.

The Queueing Network Model

An open queueing network model of a service center with gatekeepers is described. The 'network' is essentially two queues in series: n_g gatekeepers and n_e experts, with a staffing costs c_g and c_e per unit time, respectively (see figure). Customers (or 'calls') arrive to the gatekeepers according to a Poisson process with rate λ. To the gatekeepers, the calls vary in difficulty and complexity, and the difficulty of each call is represented with a random draw from a uniform distribution, U[0,1]. This random variable represents the call's percentile in a ranking of calls by treatment complexity. Given that a call has complexity x, the probability that the customer can be treated successfully by the gatekeeper is f(x): treatment function. Because complexity increases with x, we assume that f0(x) ≤ 0.

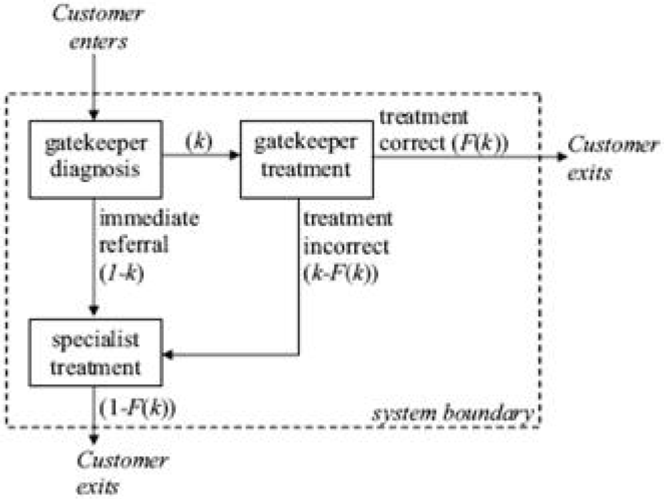


Figure 22: gatekeeper and referral configuration of a call center

With each new call, a gatekeeper spends time diagnosing the problem and determining the complexity (the value of x). The gatekeeper may then either send the call directly to the expert pool or attempt to solve the problem. If the gatekeeper successfully solves, or 'treats,' the problem, the call leaves the system. If the gatekeeper attempts to treat and the treatment fails, a cost m due to the inconvenience is assessed to the customer, and the call is sent to the expert pool. Once a call has reached an expert, it is served and leaves the system. Both server pools have unlimited waiting space, and there is a cost w for each unit of time spent waiting. The time required for an expert to treat a call averages 1/µ. The time for a gatekeeper to diagnose a call averages $1/\mu_d$, while the average time to diagnose and treat is $1/\mu_t > 1/\mu_d$. If the gatekeeper follows a static policy and treats a proportion k of calls, then the gatekeeper's service rate is,

$$\overline{\mu}(k)=\left(\frac{1-k}{\mu_d}+\frac{k}{\mu_t}\right)^{-1}$$

Assumption: service times are distributed as independent, exponential random variables, even when the gatekeeper only diagnoses some calls, and combines diagnosis with treatment in other calls. Given these assumptions, the gatekeeper and expert pools can each be modeled as M|M|N queueing systems, where the arrival rate to the expert pool is the sum of the rate of calls untreated by the gatekeeper and the rate of calls mistreated by the gatekeeper.

Minimizing costs is the objective. Some parameters can be fixed, others variable. Solving this problem can be done numerically (minimizing the objective function under certain restrictions) and is described in section 4.3, about mathematical models.

Considering different service times for experts and gatekeepers (diagnosing and may be treatment), a more accurate model would use a mixture of two exponential service times: a proportion k with mean $1/\mu_t$ and 1−k with mean $1/\mu_d$. Given that the gatekeeper's service times follow such a distribution, the gatekeeper pool is modeled as an $M|H_2|N$ queue and the expert pool as a G|M|N queue. This problem can easily be solved with square-root staffing, described in section 4.3.

The Deterministic Model

A lot of numerical experiments have been performed by the authors and in every case the heuristic and optimal solutions are nearly identical, and the difference in total cost when using each is negligible.

Consider a deterministic model of the two-tier system with no stochastic variability in the arrival or service rates, so that the capacity of the gatekeeper and expert pools are set equal to the load. Given the linear treatment function f (k), the total cost of this system is

$$C_2^d(k)=c_g\lambda\left(\frac{1-k}{\mu_d}+\frac{k}{\mu_t}\right)+c_e\frac{\lambda\left(1-bk+bk^2/2\right)}{\mu}+m\lambda\left(k-bk+bk^2/2\right)$$

and the optimal treatment threshold is,

$$k^d = \left[1 - \frac{1}{b} \frac{m + c_g \left(\frac{1}{\mu_t} - \frac{1}{\mu_d} \right)}{m + c_e \frac{1}{\mu}} \right]^+$$

A one-tier deterministic model has total cost,

$$C_1^d = c_e \frac{1}{\mu}$$

Rule of thumb for choosing optimal system

The choice of a one or two-tier system should be based on a cost comparison that takes optimal staffing and waiting costs into account. Therefore, the following rule of thumb is proposed for choosing the optimal system:

1. Calculate k^d using Equation 2.
2. Using k^d as the treatment threshold, use the square-root staffing rule to determine the number of gatekeepers and experts in a two-tier system. Given these staffing levels, calculate the total cost $\hat{C}_2(k^d)$. Also using the square-root staffing rule, determine the number of experts in the direct-access system and calculate the cost $\hat{C}_1$.
3. If $\hat{C}_2(k^d) < \hat{C}_1$, choose a two-tier system using k^d as the treatment threshold. Otherwise, choose a direct-access system.

This rule of thumb does not require managers to find $k*$ or k^h (approximations), both of which require significant computational effort compared to finding k^d.

Appendix J: Akşin and Harker (2003)

Revenue is generated by serving a customer. Each time a customer is lost, the system incurs a revenue loss. Thus, in order to relate staffing decisions to revenues, one needs to characterize the customer loss as a function of the number of servers. Given average revenue per customer served, and the proportion of customers that are lost, one can then determine the total revenues generated in the system. The analysis will assume that customer loss can occur in two different forms. The first one of these, labeled as blocked customers, occurs whenever the finite resources (servers and buffers) are all occupied by existing customers and the arriving customer leaves. The second type of loss occurs when all servers are busy, the customer is placed in one of the available buffer resources to wait, however loses patience and leaves. This latter type of loss will be labeled as a renege. The sizing problem considers a service system with k = 1,…,K types of customers. Each type of customer will be served by an access channel, alternatively called a department, with dedicated servers. Given the demand rates λ_k, the objective is to determine the number of service agents to be employed for the different access channels, such that systemwide revenues net of costs are maximized. More specifically, the model can be stated as:

$$\max_{s} \sum_{k=1}^{K} \left[v_k \lambda_k \left(1 - B_k(S,T)\right)\left(1 - R_k(S,T)\right) - C^s(S_k) \right]$$

where

- $S = (S_1,\ldots,S_K)$, server allocation vector;
- $T = (T_1,\ldots,T_K)$, buffer allocation vector;
- $B_k(S, T)$, blocking probability for type k customers;
- $R_k(S, T)$, renege probability for type k customers;
- v_k, revenue generated from type k customers;
- $C^s(S_k)$, cost associated with keeping Sk agents of specialization k for a single time period.

A sizing algorithm

If one can decompose the service system into single class systems in a way that captures the interaction between classes in the original system, one can then independently determine the optimal staffing levels in each single class subsystem. This idea is used in the algorithms in the article by Akşin and Harker (2003). In particular, the algorithms will start by decomposing the original system into individual single class systems, will determine staffing levels for each single class system, and then go back to the original system and iterate the same procedure. The algorithms will stop once equilibrium is attained. Variations in the way the system is decomposed and the individual class staffing is performed lead to two different versions of this basic algorithm. Numerical examples are used to generate guidelines for when one would prefer to use which algorithm. The interaction between classes can be captured as a change in the service rates. This is characterized by the state-dependent service rates μ(n) in a system with reneging. Thus, a multi-class system with reneging can be treated as K independent single class systems, where service rates μ_k in each class are modified such that they are as close as possible to the original rates:

$$\mu_k(n) = \frac{\mu_k \min(n_k, S_k)}{\sum_{i=1}^{K} \min(n_i, S_i)}$$

These rates change dynamically as the state vector n changes.

The performance model is based on queueing theory. For detailed information on the algorithms and the performance model, please check the article by Akşin and Harker (2003).

Appendix K: Table overview of existing approaches

In this appendix a summary/listing is provided from the results of chapters 2 to 6. First the authors that presented a (re)design approach towards determining staffing levels for call centers will be listed. For these authors the table represents the authors' name, the characteristics of the (re)design approach, the typology of the human resources occupying the call center, the type of measurements and the special characteristics (compared to the other approaches and authors). At the end of the table a number of authors (Akşin and Harker (2003), Pinker and Shumsky (2000), Mazzuchi and Wallace (2004) and Pichitlamkin et al. (2003)) is listed that do not present any particular (re)design approach, but provide interesting information on the other characteristics of determining optimal staffing levels.

Table 6: overview of existing approaches towards determining staffing levels, listed by author

Author(s)	**(Re)design Approach**	**Typology of Human Resources**	**Type of Performance Measurements**	**Special Characteristics**
Gans et al. (2003)	Capacity Management on different hierarchical levels: 1. low: queueing models (daily/weekly) 2. intermediate: math. progr models 3. high: long-term planning models (for hiring and training) forecasting and estimation (for performance) models to determine good input parameters	General typology/ Human Resources	Two regimes of performance Management are mentioned: Efficiency-Driven (ED) (quantitative) and Quality-Efficiency Driven (QED) (quantitative and qualitative)	Suggestion to use Dynamic Programming for the routing problem Typology Simplification by means of logical canonical designs for different SBR policies Square-root staffing rules for certain canonical designs for SBR policies
Zapf (2004)	Three design dimensions: 1. Task allocation to generalists and/or specialists 2. Front-Office and Back-Office Roles 3. Degree of integration of synchronous and a-synchronous service requests Using Queueing Theory and Simulation Modeling together to determine optimal staffing levels	Generalists Specialists The division of roles and tasks is not strict but variable (1st and 2nd line, FO And BO, integrated)	Mainly quantitative. Based on these quantitative some general qualitative conclusions are drawn (agent pooling and task competition)	Due to the use of Queueing Theory and Linear Programming, the modeling complexity can be quite high Nice and clear representations of different call center Structures and designs. Outcalls are also modeled Stochastic discrete event simulation (to overcome certain complexity restrictions of modeling) Agent pooling and task competition are explicitly dealt with
Hasija et al. (2005) & Shumsky and Pinker (2003)	Principal agent model is used as basis for determining the optimal referral rate and thus the division of workload over 1st and 2nd tier servers	Gatekeepers Experts	Mainly quantitative, but also Qualitative. Maximizing the firms profit is very important Cost minimizing referral rate	Referral rate based on principal agent model Square-root staffing rule for optimal staffing levels

Author(s)	(Re)design Approach	Typology of Human Resources	Type of Performance Measurements	Special Characteristics
Chan (2003)	Identification of Key Output Performance Variables (KOPV) and Key Input Performance Variables (KIPV) Leading to a simulation model	General typology/ Human Resources	Mainly quantitative. Depends on KOPVs identified Main objective is to grant effectiveness	Very clear and detailed approach towards developing a simulation model General approach which should be applicable to a wide Range of different call centers
Ernst et al. (2004)	Demand modeling based on forecasts on input parameters and satisfying service standards Using Queueing Theory and Simulation Modeling together for optimality and complex systems/processes	General typology/ Human Resources	Quantitative: minimizing costs Qualitative: customer and employee satisfaction	Proposed (re)design activities are part of a (practical) rostering tool
Mehrota (1997) & Mehrota and Fama (2003)	Workforce Management Software	Agents (with particular Skills) Also cross-trained agents is mentioned	Mehrota (1997) uses only quantitative measurements Mehrota and Fama (2003) use both quantitative and qualitative measurements. They also measure the well-being of call center agents	Quite an operational approach towards staffing call centers No mathematical foundations provided
Henderson and Mason (1998)	Queueing Theory and/or simulation modeling to determine staffing requirements for each period of the day Three techniques provided to do so	General typology/ Human Resources	Quantitative: CGOS (customer grade of service) depends on the customer's waiting time in queue Qualitative: customer's waiting time in queue and utility curves are used to obtain customer satisfaction	Three techniques to determine staff requirements: 1. Steady-state queueing models 2. Numerically calculate time varying distribution of GOS (grade of service) 3. Simulation RIIPS (Rostering by Iterating Integer Programming and Simulation) is a convergence tool that minimizes costs (linking between adjacent timeblocks is missing)

Author(s)	**(Re)design Approach**	**Typology of Human Resources**	**Type of Performance Measurements**	**Special Characteristics**
Lin, Lai and Hung (1998)	Regression model based on: - forecasts, - call center configuration - and the public holiday file simulation model delivers staff requirements (meant for scheduling and rostering)	Juniors Seniors (based on experience Level)	Quantitative, depends on the number of abandoned calls	Overlapping shifts, aimed at staffing a 24/24 hotline or call center
Akşin and Harker (2003)	-	Servers Commonly shared Resources (information System)	Mainly quantitative, but also qualitative (concluding from some quantitative measurements) Maximization of revenues is related to the negative effect of lost customers	Commonly shared resource (IS) as the bottleneck resource
Pinker and Shumsky (2000)	-	Flexible and Specialized workers	Mainly quantitative, but also qualitative (concluding from some quantitative measurements) The Quality of Service is Related to the fraction of customers served (which is positively related to the revenues of the system)	Queueing and waiting times ignored (because of the emphasis on simple quantitative measurements)
Mazzuchi and Wallace (2004)	-	Individual agents with a skills matrix	Quantitative: - speed to answer measurements - blocking probabilities - agent's utilization	The use of a skills matrix for the agents Server experience is not influencing the service time (per skill)
Pichitlamkin et al. (2003)	-	Inbound only agents Blend agents (handle Overflow from inbound)	Quantitative. Queueing model (CTMC) used to measure system performance	Outcalls as well (not related to incoming service requests) Use of Simulation and Queueing Theory together to obtain optimal results