A REVIEW OF WORKFORCE CROSS-TRAINING IN CALL CENTERS FROM AN OPERATIONS MANAGEMENT PERSPECTIVE

O. Zeynep Akşin[†]

Fikri Karaesmen[††]

and

E. Lerzan Örmeci[††]

[†] Graduate School of Business

Koç University

34450, Istanbul TURKEY

[††] Department of Industrial Engineering

Koç University

34450, Istanbul TURKEY

zaksin@ku.edu.tr, fkaraesmen@ku.edu.tr, lormeci@ku.edu.tr

July 2006, First version: November 2005

A REVIEW OF WORKFORCE CROSS-TRAINING IN CALL CENTERS FROM AN OPERATIONS MANAGEMENT PERSPECTIVE

Zeynep Akşin[†], Fikri Karaesmen[††], and Lerzan Örmeci[††]

[†] *Graduate School of Business, Koç University, Istanbul, Turkey, zaksin@ku.edu.tr*

[††] *Dept. of Ind. Eng., Koç University, Istanbul, Turkey, fkaraesmen@ku.edu.tr, lormeci@ku.edu.tr*

1 Introduction

Call centers, also known as telephone, customer service, technical support, or contact centers constitute a large industry worldwide, where the majority of customer-firm interactions take place. There are thousands of call centers in the world, with sizes in terms of full time employees ranging from a few to several thousand. Datamonitor estimates that the 2.5 million agent positions in the United States today will increase by more than 14% by 2005 (Datamonitor, 2002). Call centers in Europe, the Middle East and Africa (EMEA) are expected to swell to 45,000 by 2008, with 2.1 million agent positions (Datamonitor, 2004).

Modern call centers assume many different roles. While the telephone is the basic channel for call centers, contact centers incorporate contacts with customers via fax, e-mail, chat, and other web-based possibilities. Agents in multi-channel contact centers have the possibility of responding to customer requests via several different media (Armony and Maglaras, 2004a, 2004b). Inbound call centers answer customer calls, whereas outbound centers make telephone calls to customers or potential customers typically for telemarketing or data collection purposes. Many call centers combine these two features using what is known as call-blending, where agents that normally take inbound calls perform outbound calls during times of low call volume (Bhulai and Koole, 2003;

Gans and Zhou, 2003). As durable goods and technology companies globalize, technical support needs to be provided to customers that buy the same products around the globe. This support is given in several different languages in multi-language call centers located in hubs like Ireland, the Benelux countries and Eastern Europe. In these centers, agents that posess several technical skills or speak several different languages respond to customer queries (Akşin and Karaesmen, 2002). Today, call centers in mature industries like financial services are in the process of transforming into revenue or profit centers. To do this, these centers incorporate sales and cross-selling into their processes, thus requiring agents to become experts in both service and sales (Akşin and Harker, 1999; Güneş and Akşin, 2004). Many companies outsource their call center needs to third parties. An agent working in one of these outsourcing firms, may be responding to calls originating from customers of different firms, all clients of the outsourcing company. Such agents need to possess knowhow for products, promotions or practices of different companies (Wallace and Whitt, 2004). Multi-channel contact centers, call-blending call centers, multi-language technical support centers, service and sales centers, and call center outsourcing are all examples of the growing diversity and complexity of call center jobs. Workforce cross-training has emerged as an important practice for companies that need to deal with this diversification in call center jobs.

There is a delicate balance between quality and costs in the management of call centers. As a direct and important point of contact with customers, the call center needs to provide good call content and high accessibility. Call content quality hinges extensively on human resource practices like staff selection, training, and compensation, since it occurs at the point of interaction between agents and customers. High accessibility implies among other things that a calling customer will be answered with a minimum wait. Determining the appropriate number of agents to have in the presence of uncertain call volumes constitutes one of the most important challenges of call center management, since adding staff implies adding costs for a call center. Indeed, 60-70% of call center

costs are associated with staffing (Gans et al., 2003). The fact that better staffing practices are becoming a competitive need is supported by the prediction that call center investment in workforce optimization technologies will exceed $1 billion by 2006 (Datamonitor, 2005). To cope with the growing diversity of calls while keeping staffing costs to a minimum, a higher level of flexibility in answering calls is required. Structures with multi-skill agents are becoming more prevalent, as the following studies on multi-channel contact centers indicate: *"78% of call centers surveyed by ICMI used agents skilled in both phone and non-phone transactions to handle inbound calls, 17% used only agents dedicated to the phone channel, and 5% said they used some multiskilled agents"(ICMI 2002). "When handling transactions from multiple channels (both service level and response time objective transactions), just over half (52%) of call center managers participating in a recent ICMI Web-based seminar said they use blended agent groups for the different channels as the workload allows. Another 29% use separate agent groups for each channel, while 8% said they relied on multimedia queues and multiskilled agents who handled whatever type of transaction was next in queue" (ICMI, 2000).* Technology provides skills-based routing capability, enabling the routing of calls to the agents with the appropriate skills, thus allowing call centers to reap the benefits of flexibility. Through cross-training, call centers can increase the flexibility of their staff. IDC estimates revenue in the contact center training industry will grow from $415 million in 2001 to nearly $1 billion in 2006 (ICCM Weekly, 2002). Some of this growth will come from cross-training.

This chapter will review the relatively young however growing literature on workforce cross-training in call centers, making ties to related work in operations flexibility. In the subsequent section, we discuss the costs and benefits of cross-training that have been analyzed in the literature. This section also illustrates the multi-disciplinary nature of the issue, on which we take a predominantly operational view. Section 3 establishes the importance of flexibility in call center operations. Viewing workforce cross-training from this flexibility lens, we identify three basic

questions pertaining to the design and control problems, which are introduced in Section 2. The first design problem is the skill set design, which we also label the *scope* decision. In terms of cross-training, this refers to the questions: In which skills should servers be cross-trained? How many skills should they have? Literature related to this problem is presented in Section 3. Having decided on the scope, the next issue is to decide what proportion of the workforce should be cross-trained. We label this design problem the *scale* decision and review related work on it at the end of Section 3. The design problems are closely related to the control problems of staffing and routing. Indeed, the value of different designs will interact with the subsequent staffing and routing decisions. Papers that focus on the control part of the problem are reviewed in Section 4. We end the chapter with a discussion of future directions in the last section.

2 Costs and Benefits of Call Center Cross-Training

Like all human resource initiatives, there are benefits expected from cross-training practices. These are motivational benefits due to the enlargement or enrichment of jobs, cost benefits due to improved capacity utilization or improved speed, quality benefits including improved customer service. These benefits come along with costs like training costs, loss of expertise and job efficiency, and mental overload. In this section, we review selected articles from the organizational behavior and operations management literatures that identify different costs and benefits of cross-training, and then interpret these in the context of call centers.

There is a vast literature in organizational behavior that explores the relationship between job design and performance. Grebner et al. (2003) focus on this problem in the call center context. The authors provide examples of earlier studies supporting the argument that call center jobs are predominantly specialized and simplified (Isic et al., 1999; Taylor et al., 2002), and as a result require a relatively short period of training (Baumgartner et al. 2002). In their article, they

empirically test the premise that this division of labor and simplification, while reducing personnel costs, results in low variety-low job control which leads to lower well being and a higher intention to quit among call center workers. The authors argue that call center jobs need to be redesigned in order to improve autonomy, variety, and complexity. In general, it has been argued that cross-training improves job enlargement (variety) and job enrichment (autonomy), which in turn has a positive impact on performance (Hackman and Oldham, 1976; Ilgen and Hollenbeck, 1991; Xie and Johns, 1995). However, in Xie and Johns (1995), the authors demonstrate that there is a limit on the positive impact of job scope, and that beyond a threshold, job scope can become excessive and induce stress which is dysfunctional for the organization. Thus we find that the motivational benefits one expects from cross-training in call centers can turn into costs if the resulting job scope is too high. Combined with Grebner et al. (2003)'s argument, this suggests that from a motivational standpoint, moderate job scope is superior to no variety or excessive scope job designs. An empirical investigation that determines the ideal region for job scope in different types of call centers remains to be done.

Campion and McClelland (1991, 1993) explore both the costs and benefits of job enlargement in service jobs. Taking an interdisciplinary perspective, they summarize four different models of job design coming from different disciplines: A *motivational* design that argues for job enlargement and enrichment; a *mechanistic* design recommending simplification and specialization; a *biological* model that advocates reduced pysical stress and strain; and a *perceptual-motor* model recommending reduced attention and concentration requirements that lead to increased reliability. These models clearly point to some tradeoffs to be made between different benefits and costs involved. The authors identify the benefits of job enlargement as satisfaction (of individuals), mental underload, enhanced ability of catching errors, and improved customer service. The costs of job enlargement are stated to be higher mental overload, training requirements, higher basic skills requirements, more

chance to make errors, decline in job efficiency, more compensable factors like education and skills (leading to higher compensation). Job enlargement is classified into two types: *task enlargement* which involves adding new tasks to the same job, and *knowledge enlargement* which refers to adding requirements to the job that enhance understanding rules and procedures about other products of the organization. In their analysis, Campion and McClelland (1993) find that task enlargement mostly results in costs or negative benefits: more mental overload, greater chance of making errors, lower job efficiency, less satisfaction, less chance of catching errors, and worse customer service. On the other hand, knowledge enlargement is found to result in more satisfaction, less mental underload, greater chances of catching errors, better customer service, less mental overload, lesser chances of making errors, and higher efficiency.

Reinterpreting in a call center context, we could say that task enlargement involves completing more portions of a customer request. For example not just answering the initial part of a technical query, but completing the entire query, or not just attempting a sale but actually completing the entire sales transaction could be seen as examples of task enlargement. On the other hand answering calls pertaining to different products, different languages, or different regions could be seen as examples of knowledge enlargement. Though the results remain to be tested in a call center setting, the Campion McClelland results suggest that escalating calls to specialists in a technical support center, or separating service and sales roles may be desirable in order to avoid enlargement costs. Most call centers in the Evenson et al. (1999) study are reported to exhibit this type of service and sales separation, thus in a way providing a confirmation from practice. On the other hand, cross-training that enhances flexibility through call-blending of inbound and outbound, or developing an expertise in several products seems to be beneficial from a job design perspective.

The organizational benefit derived from the flexibility of staff in the presence of uncertainty, is not considered in the job design literature. Taking a modeling perspective and focusing on call

centers, Pinker and Shumsky (2000) ask the following question: "Does cross-training workers allow a firm to achieve economies of scale when there is variability in the content of work, or does it create a workforce that performs many tasks with consistent mediocrity?". The authors model the tradeoff between cost efficiency due to economies of scale resulting from cross-trained staff and quality benefits from experience-based learning in specialists. The authors find support for the well known fact that a system with flexible servers can achieve the same throughput with fewer servers, however also demonstrate that higher flexibility results in lower quality and less customer satisfaction due to a decrease in experience in any given skill. This type of a quality deterioration is also found to happen in a specialist system due to the possibility of low utilization, once again sacrificing experience related quality performance. The authors conclude that an ideal design will consist of a mixture of specialized and flexible servers.

In a general context, taking a predominantly operations management perspective, Hopp and Van Oyen (2004) develop a framework to assess the appropriateness of cross-trained workers in different manufacturing and service settings. Called agile workforce evaluation framework, this framework identifies, by surveying an extensive literature, the links between cross-training and organizational strategy, and superior architectures and worker coordination mechanisms for workforce agility implementations. The links to organizational strategy are classified in two groups, direct and indirect. The direct links are to cost in the form of labor cost reduction, time which captures efficiency, quality, and variety. Improvements in motivation, retention, ergonomic factors, experience based learning, communication, and problem solving are the indirect links. Of these, variety is one which has not been directly considered in the job design literature. Based on the brief review of the literature we have presented so far, we note that all of the direct and indirect links play a role in evaluating worker cross-training in call centers, emphasizing the importance of this practice for this industry.

The Hopp and Van Oyen paper classifies workforce agility architectures in terms of skill patterns, worker coordination policies, and team structure. We provide examples of the first two features in the context of call centers. Skill patterns relate to the questions that we refer to under the scope decision mentioned in the introduction. It also resembles the task versus knowledge enlargement of Campion and McClelland (1993). Skill patterns can be established taking skill types, entities (jobs or customers) or resources as a basis. In a call center setting, as mentioned by the authors, skill types can refer to calls pertaining to different products. Examples for entity based tasks are similar to those identified under task enlargement above. Resource based tasks could be for example cross-training in a contact center in a way that servers are dedicated to particular channels (resources) however are cross-trained to do all tasks within a channel.

Worker coordination policies determine how workers and tasks are matched over time. Call centers allow for a wide range of worker coordination policies: fully cross-trained generalists who respond to all types of queries all the time, partially cross-trained workers who can respond to queries about more than one type of product but not all, cross-trained workers who answer the same type of calls at certain times of the day (for example outbound calls after five, inbound before), cross-trained workers who help out different groups of specialists depending on system congestion (Örmeci, 2004), worker groups whose skill sets are nested such that calls are escalated from the simplest level to more complex as the nature of the problem is explored (Shumsky and Pinker, 2003; Das, 2003; Hasija et al., 2005). It is also possible to come across systems where each server has its own skill set and priorities defined over this set, such that calls are assigned to the next available worker posessing the required skills at the appropriate priority level. This occurs in systems that use skills based routing to coordinate workers (Wallace and Whitt, 2004; Mazzuchi and Wallace, 2004).

The appropriateness of different design choices along the dimensions of skill pattern and worker

coordination in different environments depends on *training efficiency* and *switching efficiency* (Hopp and Van Oyen, 2004). Of these, training efficiency is relevant for the skill pattern choice, and captures the ease with which workers can be trained in different skills. This factor covers similar costs as those considered in the job enlargement context before. Switching efficiency refers to the costs associated with changing tasks. These may be in the form of time lost between task changes, or in switching from one resource to another, or in terms of setup work that needs to be performed on each different entity (customer for example). Modern call center technology mitigates switching inefficiencies to some extent. Computer telephony integration and automatic call dispatching allow service representatives to view customer data and earlier history when calls need to be passed on. Since most call centers operate in a paperless environment, a physical switching time is non existent or minimal. Agents use the phone and their computers as the basic resource and this does not change from task to task. Since switching inefficiencies are relatively small, the main costs to be considered in cross-training design for call centers seem to be those mentioned under training efficiency.

Our review on the costs and benefits of cross-training suggests that apart from direct costs like the cost of training, compensation for additional skills, or quantifiable efficiency losses, costs associated with cross-training are typically beyond the scope of operations management models. The latter often focus on the benefits of cross-training, particularly from a flexibility standpoint. The organizational behavior literature that focuses on job design seems to recommend moderate scope in cross-training. In Section 3 we review articles that explore the job scope question from a flexibility perspective, thus ignoring some of the human resource related issues analyzed in the job design literature. Quality-efficiency tradeoffs modeled in Pinker and Shumsky (2000) imply that good designs mix specialized and flexible servers. Papers addressing the question of the appropriate mix of specialized and flexible servers are also reviewed in this section. Finally, we review related

literature on worker coordination mechanism in call centers as problems of staffing and routing in Section 4.

3 Cross-Training and its Impacts on Performance

This section focuses on the performance improvements brought about by cross-training. In Section 3.1, the cross-training issue is related to the framework of operational resource flexibility, and the related literature is reviewed. Section 3.2 outlines some of the general guidelines for the scope of cross-training suggested by the operational flexibility literature. Finally, Section 3.3 focuses on the issue of scale of cross-training. This section introduces and discusses the general principles but postpones a number of critical operational issues such as staffing and routing in a call center context to Section 4.

3.1 Cross-Training and Operational Flexibility Literature

In this section we present the impacts of cross-training structures on the performance of a call center focusing on operational issues. As already mentioned in Section 2, our discussion focuses on the benefits in terms of resource flexibility. In a survey article focusing on manufacturing applications, Sethi and Sethi (1990) describe several different dimensions of flexibility. Our discussion will be based on a particular type of flexibility most closely related to cross-training. Resource flexibility will be understood as the capability of a resource to perform several different tasks.

In the operations literature, resource flexibility has been studied extensively. A typical problem in that setting is to evaluate the performance of the system, such as a flexible manufacturing system, under a particular type of resource flexibility structure (see Sethi and Sethi, 1990). More recently, a number of papers have addressed operational issues in the presence of flexible resources. A relatively well-studied problem from that perspective is that of capacity investment. These papers

assume a certain form of flexibility and then explore the question of the ideal level of this flexibility and how it relates to value in terms of throughput or revenues under uncertain demand (see for example, Fine and Freund (1990) Netessine et al. (2002)). Van Mieghem (1998) provides a review of the literature on this problem.

Our focus as the main design issue in this section is cross-training or flexibility structure. There is a rich operations literature on this problem. Starting with the influential work of Jordan and Graves (1995), this stream of work keeps all other design parameters fixed and investigates the isolated effects of varying the flexibility structure. Motivated by an automative sector example, Jordan and Graves (1995) explore the problem of assigning multiple products to multiple plants. The demand for each product is random and the flexibility of the plants determines which products they can handle. Jordan and Graves develop general guidelines for this problem. One of their important findings is that well-designed limited flexibility is almost as good as full flexibility. We discuss some of these general guidelines in Section 3.2.

While the general flexibility principles of Jordan and Graves were established for a particular model, a number of subsequent papers have observed that these principles are robust to the assumptions of the particular model. For instance, the model of Jordan and Graves is a single-period model that ignores the dynamics of the system. Sheikzadeh et al. (1998) consider a flexible manufacturing system modeled by parallel queues and observe that the general flexibility principles hold in this case. Gurumurthi and Benjaafar (2004) also present a numerical investigation of the benefits of flexibility based on a queueing model. Jordan et al. (2004) also consider a queueing-based model and confirm and extend some of the general principles.

Another interesting observation is that even though the model in Jordan and Graves (1995) is a parallel resource type system, most general principles also hold for serial systems such as production lines or assembly lines. Graves and Tomlin (2003) perform a similar analysis for a

multi-stage version of the system in Jordan and Graves, and develop a flexibility measure for such systems. Hopp et al. (2004) explore the benefits of chaining in the context of cross-training for production lines. Inman et al. (2004) investigate different cross-training structures for workers in an assembly line.

Despite the existing flexibility principles, it remains difficult to say whether one particular flexibility structure is better than the other in terms of operational performance when cross-training costs are disregarded. An interesting recent development has been the work of Iravani et al. (2005) who develop a *structural flexibility index* that quantifies the structural flexibility in production and service systems, and allows a ranking based on performance of these systems. Numerical results confirm that the index is quite accurate for comparing system performance.

3.2 General Flexibility Guidelines

Let us go through an example in detail to demonstrate the impacts of different cross-training structures. Let us assume that there are three different types of calls A, B, and C each one corresponding to a particular skill. Let us also assume that the call center is organized in three departments corresponding to the main skills demanded (A,B, or C). Figure 1 depicts a sample of eight different cross-training structures. In this figure, for each structure, the nodes on the left represent the different call types, and the nodes on the right represent resources (servers) and their skills. Structure 1 in Figure 1 depicts a system where there is no cross-training and each group of servers can respond to a single type of call. While this system has obvious advantages in terms of hiring and training costs, it also has an important disadvantage. Since calls arrive randomly over time, it may be possible that one of the departments is completely busy while in another one there may be idle servers. This causes customer service levels to fall even though capacity (i.e. total number of servers) is sufficient.

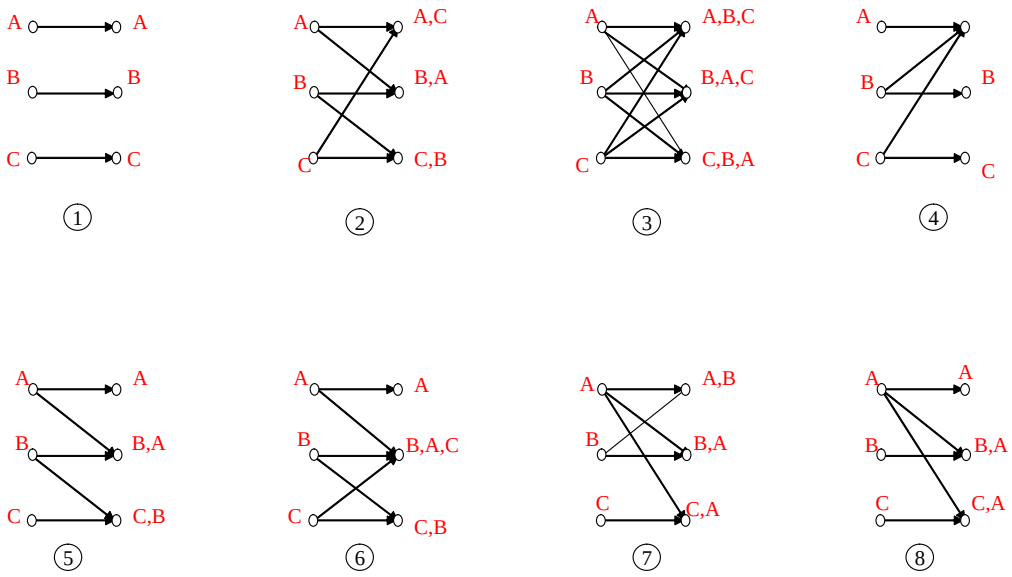

Figure 1: Different Cross-Training Structures

Let us now assume that the servers whose main skills are A are also trained in skill C, those whose main skills are B are trained in A, and those whose main skills are C are trained in B. The resulting structure is depicted as structure 2 in Figure 1. We can view this new structure as consisting of three service departments A, B and C distinguished by their main skills. In this case, if department A is busy, arriving type A calls can be answered by cross-trained servers in department B (if they are available). Similar assignments may also take place for calls of type B and C.

Structure 3 in Figure 1 represents the case of completely multi-skilled servers who are all able to respond to all three types of calls. In this structure, all servers are cross-trained to have all the three skills. As can be seen in the figure, in this structure, the notion of different departments loses its significance. In a way, there is a single multi-skilled pool of servers who do not differentiate between different types of calls. The system is as efficient as it gets in terms of customer service performance. In this structure, a customer can only wait because all servers are busy but not because of a mismatch between demand type and the skill-set of a server available. This is however

at the expense of increased cross-training costs and possibly higher burnout rates for servers due to increased task complexity.

The problem of designing the right cross-training structure now appears clearly. The system with dedicated departments in Structure 1 of Figure 1 is cost effective in terms of cross-training requirements but may suffer in customer performance, whereas the Structure 3 of the same figure is at the opposite extreme. Is there a cross-training structure somewhere in between the two extremes that is satisfactory in terms of customer service performance but not too costly in terms of cross-training expenses? In other words, what is the right scope of resource flexibility? The answer to this question in a given situation depends on a number of complicated aspects of the problem such as customer service performance objectives, the capacities of departments, the difficulty and the costs associated with cross-training the servers, and the operational costs of using cross-training effectively (i.e. by correctly routing the calls). On the other hand, there are certain general properties which provide a guideline to these issues. Starting with the work of Jordan and Graves (1995), this question has received a lot of attention. Below, we outline some of the general principles concerning cross-training structure.

We begin with a basic property, motivated by a comparison between two systems where one of them has additional cross-trained servers. An example would be a comparison of Structures 1, 5 and 2 in Figure 1. If we leave aside problems of call routing, answer quality and call duration, clearly the performance of structure 5 is at least as good as that of Structure 1 but cannot be better than the performance of structure 2. In fact additional cross-training can be seen as a possible call assignment option which cannot degrade performance.

Property 1 *Increased cross-training increases (in a non-strict sense) customer service performance.*

Despite its simplicity and generality, Property 1 is of limited use for cross-training design since

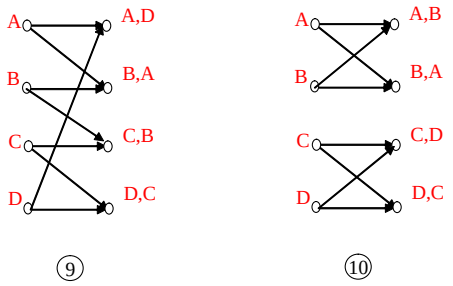

Figure 2: Two different structures: two chains versus a single chain

it does not enable us to make a useful statement on how much cross-training is required given that cross-training is costly. In order to obtain a better sense of the performance vs. cost tradeoff, more subtle results on flexibility structure are required. To this end, let us define the concept of chaining as introduced by Jordan and Graves (1995). A chain is a group of demand and resource types that are either directly or indirectly connected by demand-resource assignment decisions. Jordan and Graves argue that performance is enhanced by constructing chain structures in the demand-resource assignments. One of the main principles stated in Jordan and Graves is that longer and fewer chains are better than multiple shorter chains.

Property 2 *Longer and fewer chains in the cross-training structure lead to improved customer service performance.*

As an example of Property 2, consider the two structures in Figure 2 corresponding to 4 different demand types A, B, C, and D. Structure 10 has two chains whereas Structure 9 has a single but longer chain. The performance of Structure 9 turns out to be better than that Structure 10 in general. The intuition here is more subtle. Structure 10 can absorb peaks in the demand of a single class but cannot absorb peaks of both A and B or both C and D type demands whereas Structure 9 can actually handle peaks in A and B by shifting some of the demand to departments C and D.

Property 2 establishes that one should strive to construct long chains when designing the skill

structure. Structure 2 in Figure 1 comprises a single long chain that connects departments A, B, and C. On the other hand, this property does not enable a comparison of different chains of the same length. In fact, Structures 2-7 are all single long chains. It turns out however that the performance of the structure in Case 2 is, in general, better than all other structures except the fully flexible structure 3. Aksin and Karaesmen (2002, 2005) provide a formal justification of this general property for a particular model.

Property 3 *If the call rates for different call types are balanced and the number servers in each department is roughly identical, then balanced cross-training designs outperform less balanced ones.*

In order to interpret Property 3, note that all servers have 2 skills in Structure 2 of Figure 1 whereas in Structure 4, there are some servers with 3 skills and others with 1 skill. If demands and therefore staffing levels at different departments are roughly balanced, the three-skill department in Structure 4 may be heavily demanded and may not be able to absorb peaks from multiple demands as efficiently as in Structure 2.

By Property 3, Structure 2 is superior to structures 4, 5, 6, and 8. On the other hand, Structure 7 seems comparable to Structure 2 since all servers have two skills in both structures. It turns out that Structure 2 is in general superior to Structure 7 because in Structure 7 type C calls can only be handled by department C diminishing the effectiveness of cross-trained resources at that department (see Aksin and Karaesmen (2005)).

Property 4 *If the call rates for different call types are balanced, then cross-training designs enabling balanced routings outperform designs that allow less balanced routings.*

The final important design issue pertains to the question of combining Properties 1 which states that more flexibility is better for performance with Properties 2, 3 and 4 that highlight the

importance of balanced chain structures. Can we compare chain structures that may have different amounts of flexibility? One of the main results in flexibility design is that the performance Structure 2 is almost as good as Structure 3 (of Figure 1) under a variety of assumptions and for different models. Jordan and Graves (1995), Aksin and Karaesmen (2002), Gurumurthi and Benjaafar (2004), Jordan, Inman, and Blumenfeld (2004) all present numerical examples in different settings and Aksin and Karaesmen (2005) provide a theoretical justification for a particular model.

Property 5 *Well-designed limited resource flexibility is almost as good as full resource flexibility in terms of performance.*

It is interesting that Property 5, like Properties 1 to 4, is robust to modeling assumptions and stays consistent in different settings ranging from single-period models introduced by Jordan and Graves (1995), to queueing based models in manufacturing such as in Sheikzadeh et al. (1998), Gurumurthi and Benjaafar (2002), Jordan et al. (2004), and Inman et al. (2004). More support for these findings come from the work of Iravani et al. (2005) whose Structural Flexibility Index (SFI) is proposed as an indicator of the performance of the structure. First, SFI supports the above stated properties. For instance, the SFI indices for structures 2, 3, 5 and 8 of Figure 1 are respectively 6, 9, 3.73 and 4 implying that regardless of the setting Structure 2 is anticipated to be superior to Structures 5, and 8. Second, in a systematic study, Iravani et al. (2005) test the predictive performance of SFI on a variety of different models and find that the index is robust with respect to model structure and assumptions.

Most papers discussed so far proposing flexibility guidelines, describe models and applications in manufacturing settings. Even though the queueing literature is relatively rich in the investigation of pooling issues (see Buzacott (1996) and Mandelbaum and Reiman (1998) for example), there is less work on systematic investigation of these structural properties in a call center setting. Aksin and Karaesmen (2002) propose an approximation for a call center based on an upper bound on

system performance and show that this approximation possesses some of the general structural properties presented above.

A major challenge in the transition from the static Jordan and Graves (1995) framework to a queueing framework more appropriate for call centers is that the static framework assumes that calls (demand) will be allocated optimally to resources. In a queueing framework this is not a trivial issue since calls have to be routed dynamically. It is known that careless routing policies may have a negative effect on performance despite a correct flexibility structure (see Gurumurthi and Benjaafar (2004) and Jordan et al. (2004) for examples). The routing issue will be discussed in detailed in Section 4 which also presents a review of some recent call center research.

3.3 Scale issues in Cross-Training

The operational design issue discussed so far in Sections 3.1 and 3.2 is essentially the problem of skill set design: How many different skills should the servers have and which ones? This issue can also be called the *scope* decision (Aksin and Karaesmen (2002)). Another important question is with regards to the *scale* decision: what is the right proportion of servers to cross-train given desirable skill-set structures? Going back to Structure 2 of Figure 1, it is plausible that Department A could in fact consist of a mixture of specialists (A skills only) and cross-trained servers (A,B skills). The question is to find the right trade-off for the proportion of specialists and cross-trained servers. This question seems to have received less attention. Aksin et al. (2005) present some results for the Jordan and Graves framework and show that there is a diminishing returns property in terms of the scale of cross-training; the marginal value of an additional cross-trained server is decreasing in the number of existing cross-trained servers. This implies that finding the right scale is an important question since initial increases in scale improve performance significantly but the improvement drops as scale is added.

To our knowledge, only a few papers directly address cross-training scale design issues in call centers. Pinker and Shumsky (2000) investigate the mix of cross-trained workers versus specialists taking into account the quality tradeoff. Chevalier et al. (2004) find that, for a number of different cost structures, cross-training around 20% of the servers is the optimal tradeoff. Both of these papers investigate the scale issue for relatively simple cross-training structures and it would be interesting to verify whether such results are robust to the scope of cross-training. In a more recent paper, Jouini et al. (2004) describe a call center design problem where a transition occurs from a completely pooled structure to a dedicated team-based organization where each team is assigned to a group of customers. The dedicated organization is similar to the Structure 1 of Figure 1 and does not benefit from the demand pooling effect. The authors however find that this structure can be surprisingly efficient if the teams can accept only a small number of calls from other customer groups than their own. This suggests that a significant performance improvement can be obtained by cross-training a limited number of servers in each team.

4 Staffing and Routing

In the previous section, the aim has been to characterize preferable skill set designs in a multi-skill call center, which gives possible cross-training levels. In this context, these decisions correspond to strategic level decisions. In this section, on the other hand, we will consider the tactical and operational decisions regarding cross-training. More explicitly, we want to compute the number of operators from each skill set determined previously, such that the call center satisfies certain constraints due to quality of service and workforce scheduling, with a minimal cost. Hence, we start by describing the operations of a call center, which includes, among others, the quality of service criteria, certain staffing issues, and routing rules for calls of different kinds. Then, we will concentrate on the relations between cross-training and the operational characteristics of call

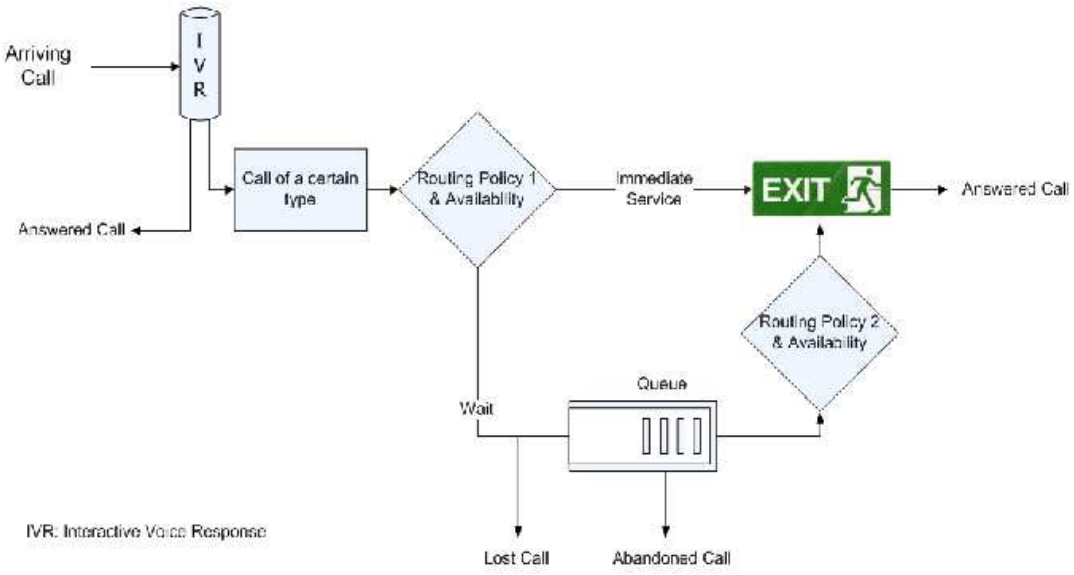

Figure 3: Call answering process

centers.

Figure 3 shows the process of answering an incoming call, from which we can identify a number of factors that affect the performance of a call center: arrival rates of calls referred to as call volumes, call types, routing policy, number of total trunk lines, abandonment behavior, number, capability (i.e., cross-training level) and service rates of operators. Among these terms, only routing policy needs further explanation: Routing policy 1 routes an incoming call to one of the available servers, if any, or delays it; whereas routing policy 2 assigns one of the calls waiting in the line, if any, to an agent who has just finished service, or keeps the agent idle. Call center management can completely control some of these factors such as the routing policy, number of trunk lines, number and capability of operators; partially control some such as service rates of operators, abandonment behavior; and finally cannot control some such as call volumes and call types. This section will take number and capability of operators and the routing policy as the control variables, with the objective of operating with a minimal cost (which typically consists of staffing and communication costs) while certain constraints regarding customer satisfaction, which we refer to as quality of

service (QoS) constraints, are satisfied. More explicitly, in section 4.1, we will assume that there is only one class of calls, so that we need to optimize only the number of operators. The aim of this section is to introduce the basic concepts on staffing call centers, so the issue of cross-training will not be mentioned here. The remaining sections, however, are devoted to discuss the effects of cross-training on staffing and routing in call centers. Therefore, in sections 4.2-4.4, we consider call centers which give different kinds of services. We assume that their skill set designs have already been determined, i.e., the possible cross-training levels are fixed. Section 4.2 discusses call routing policies given the number of agents in each skill set. Section 4.3 aims to find optimal staffing levels for each skill set present in the given design, while assuming that a routing policy is chosen. At this stage it is possible to assign no agents to a skill set, so that the given skill-set design may not be fully used. Hence, the capability of the operators (or the skill-set design) is fully determined only at this stage. Finally, in Section 4.4, we address the problem of determining the staffing levels and the routing policy together.

4.1 Staffing, shift scheduling and rostering

We first define the problems to be introduced: The *staffing problem* seeks to find a minimal workforce level for short time intervals to guarantee a certain service level during those intervals. Usually, it is assumed that each of these short time intervals behaves independently of all other intervals, although this is generally not true (see Avramidis, Deslauriers, and L'Ecuyer (2004), Brown et al. (2005), Steckley et al. (2004)). In the next level, a number of shifts are defined, such that the time of lunch, and sometimes coffee, breaks, are known as well as the starting and finishing time of the shift. Then, the *shift scheduling problem* is to determine the number of employees to schedule in each shift in order to meet the minimal workforce levels. As a result of the "independence" assumption, the solution of the staffing problem becomes an input to the

shift scheduling problem. In fact, this assumption can be relaxed so that the two problems can be solved together which will yield a better solution. However, due to its complexity, the combined problem has been considered only recently, as we will see below. Finally, a call center has a certain number of employees, where each employee has certain rights regarding the number of off-days a week, the number of weekends off, the number and sequence of day and night shifts, etc. The *rostering problem* aims to create a work schedule for each of these employees such that constraints due to staffing and shift scheduling problems are satisfied as well as these additional constraints. The rostering problem can either take the solution of the shift scheduling problem as an input, or solve the shift scheduling and rostering problems together. In order to solve the rostering problem, feasible schedules are defined over a longer time horizon, so that each schedule satisfies all the constraints due to employee rights and/or company policies. Then the problem is to assign the employees to these schedules in a such a way that the QoS constraints are also satisfied. Usually, similar methodologies are developed to solve both shift scheduling and rostering problems; so we will review both problems together. Each of these problems is difficult to solve in call centers, even if the additional complexity of cross-training is not present. Hence, in this subsection, we describe these problems without referring to cross-training issues. In other words, this section assumes a call center operating with only one class of calls.

Call centers usually operate 24 hours a day and 7 days a week. Figures 4, 5, and 6 present typical call volume patterns for a day, for 6 weeks and for a year, respectively. From these figures, we observe strong seasonality effects on both daily and weekly call volumes, which lead to significant differences on call volumes in relatively short intervals. This has a significant effect on the staffing policies. In order to convert the call volumes to staffing levels, we need to define the performance measures significant to call centers, and QoS constraints which build up on these measures. Here are the most common performance measures:

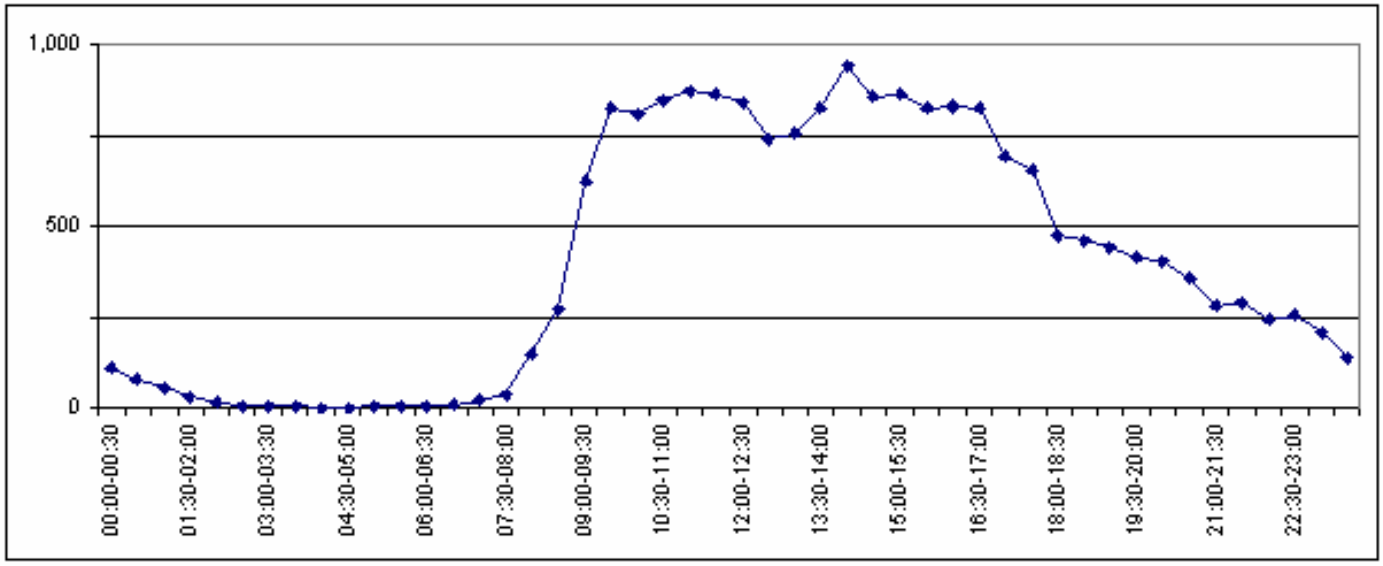

Figure 4: Half-hourly call volumes for a day

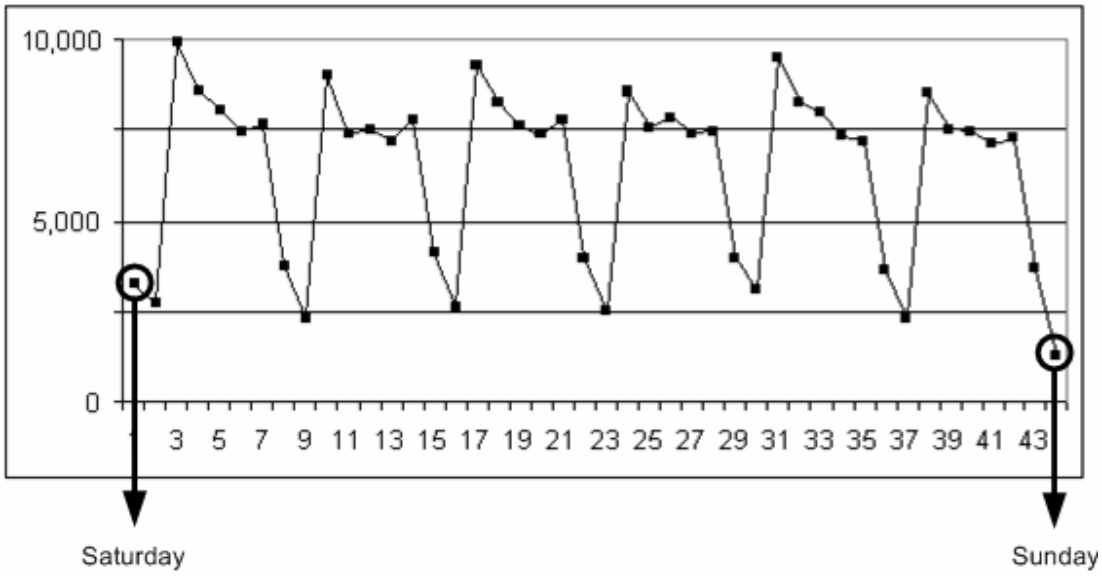

Figure 5: Daily call volumes for 6 weeks

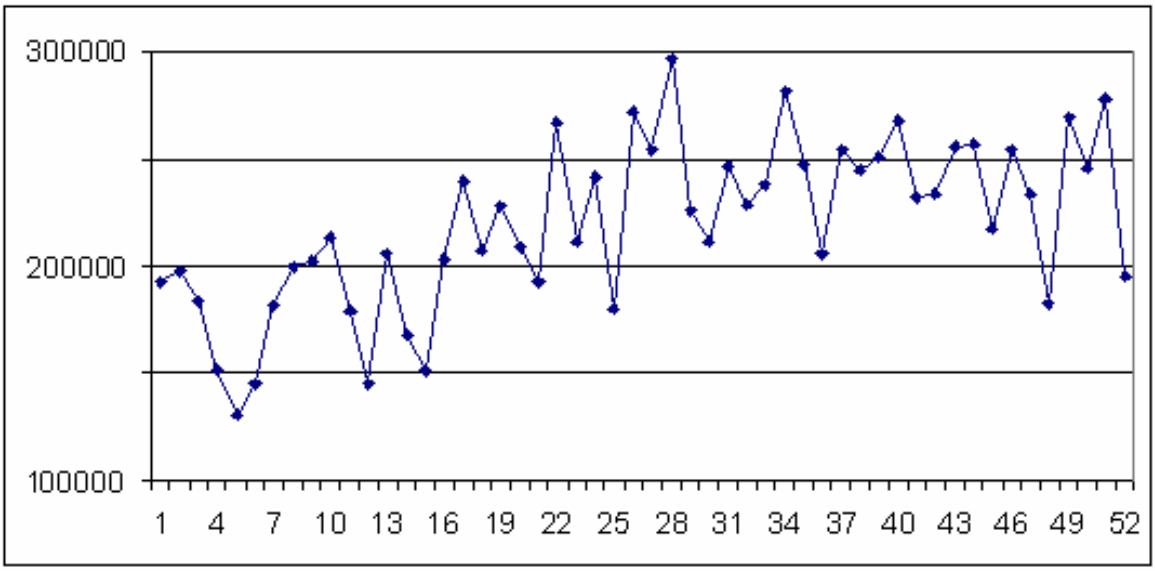

Figure 6: Weekly call volumes for a year

1. The percentage of calls that are answered within a specified amount of time, say w seconds

2. The percentage of abandonments (calls that leave the system before being answered)

3. The percentage of lost calls (calls that do not find an available trunk line to wait for an operator, and so cannot enter the system)

4. The average amount of waiting time in the queue for answered calls and/or for all calls that enter the system

Each call center specifies its own QoS constraints by setting a minimum or maximum level for one or more of these performance measures. Call centers, mostly, compute the minimum required staffing levels for each half-hour. Moreover, in practice, the primary QoS constraint builds on the first type of performance measure. Hence, the minimum required staffing levels are computed such that at least $p\%$ of all calls are answered within a specified amount of time, say w seconds, in each half-hour. The parameters p and w change with the type of the call center and/or with the type of calls to be answered. Much of the workforce scheduling software, commonly used in call centers,

determine the minimum staffing levels by a numerical procedure based on the well-known Erlang-C or Erlang-delay formula. This method has two main drawbacks. The first one is its underlying assumptions: Erlang-C formula gives the steady state behavior of a queueing system with Poisson arrivals and exponential service times, where the parameters are assumed to be constant over the half-hour being considered. However, half-hours are not always sufficiently long for a system to move into steady state, especially because the arrival rates are not constant during the half-hours. Another crucial assumption is the independence of intervals. Moreover, the effects of finite number of trunk lines (so no busy signals) and of call abandonments are completely ignored. The second drawback is due to the lack of intuition since this procedure, being numerical, does not provide any information about how changes in the parameters may affect the staffing levels. Hence, different staffing schemes are proposed in the literature. Here, we review only the most recent work, and refer the interested readers to the references given in section 4 of Gans et al. (2003).

The most commonly known methodology, both in academia and in practice, is the square-root safety staffing: If the arrivals were deterministic and constant over time, then it would be enough to have a number of servers exactly equal to the load of the system, say ρ. The square-root safety staffing rule proposes to have an additional number of servers, proportional to the square root of the arrival rate, $\sqrt{\rho}$, in order to compensate for the randomness in the arrival process. Then, the required number of servers will be $\rho + \beta\sqrt{\rho}$, where β can be considered as the target service level of the call center. There are a number of papers that analyze square-root safety staffing rule in different systems, such as Jennings et al. (1996); Garnett, Mandelbaum, and Reiman (2002); Borst, Mandelbaum and Reiman (2004); Feldman et al. (2006); Armony and Mandelbaum (2004); and Green, Kolesar, and Whitt (2006). More recent work models the effects of different factors on staffing, such as Whitt (2006a) considering uncertain arrival rate and absenteeism, and Steckley, Henderson, and Mehrotra (2004) considering random arrival rates.

In the staffing problem, the specified constraint(s) has to be satisfied in each half hour, as opposed to 8 hours or a day. Then, the minimum required staffing levels, regardless of the procedure used, follow the half-hourly call volume patterns closely. However, the actual staff levels cannot exactly match the minimum requirements exactly, since the operators work in shifts either full time being present for 6-9 hours with a one-hour of break, or part-time being present for a minimum of 3 hours. The shift scheduling problem, as mentioned earlier, aims to find the number of employees to work in each shift by satisfying the minimum staffing requirements with a minimal cost. This problem has been considered by many authors, starting with Dantzig, since the 1950's. Here, we will review recent work closely related to call centers. Henderson and Mason (1998) develop a new technique for rostering in a call center, which combines simulation with integer programming by relaxing the QoS constraints from half-hours to adjacent half-hours. Atlason et al. (2004) and Ingolfsson et al. (2005) use similar techniques to solve the staffing and rostering problems together. Koole and van der Sluis (2003), on the other hand, develop a local search algorithm for a call center staffing problem with a global service level constraint (as opposed to half-hourly constraints).

4.2 Skill-based routing in call centers with different kinds of calls

In this subsection, we consider a call center serving different types of calls with possibly cross-trained operators. We first consider the routing of calls in a simple system to show its possible effects on the performance of a call center. Assume that we have two types of calls $\{A, B\}$, and two service stations, one dedicated to B-calls, the other fully-flexible, meaning that it can serve both A and B calls. The routing policy for incoming calls is as follows: All calls start receiving service if there is one available agent capable of answering the incoming call. An incoming B-call randomly goes to one of the two stations whenever both have available servers. If there is no available server, then calls join the queue. The routing policy for the agent who just finishes a

service is straightforward for the dedicated station, since s/he just checks the queue for B-calls, and starts serving if there is at least one call in queue. On the other hand, an agent in the fully-flexible station chooses a call randomly from the queue, whenever there is at least one call. The problem with this policy is obvious: The fully-flexible station may choose to serve a B-call with a significant probability, while there are A-calls waiting in line; and afterwards an agent in the dedicated station may stay idle because s/he cannot serve A-calls. In other words, this routing policy does not use the flexible capacity wisely, as it increases the probability of having A-calls waiting in the queue while the flexible station is busy with B-calls and the dedicated station has idle capacity. In fact, the optimal policy for this type of a system is given by Xu, Righter and Shanthikumar (1992): B-calls are served in the dedicated station, while A-calls are served in the fully-flexible station. The fully-flexible station serves B-calls only if the number of B-calls in line exceeds a threshold. This shows that the routing policy should protect the flexible capacity for those who really need it.

Now, we will introduce and discuss a number of issues in a call center, which offers three different services. Then, the set of all possible kinds of calls is $\{A, B, C\}$. Moreover, we assume that the call center has three stations, where all operators in each station are capable of answering two types of calls, with the skill set design given by $\mathcal{S} = \Big\{\{A, B\}, \{B, C\}, \{C, A\}\Big\}$. Finally, we take the number of agents in each station fixed.

The first effect of having a number of services is on the description of the QoS constraints. Although it is possible to have one global QoS constraint for all calls, as in section 4.1, it is more common to set a different service level for each type of call, especially when the skills correspond to tasks bringing different returns, or to the capability of serving different customer segments. To describe explicitly, our example may have the following QoS goals: at least 90% of A calls are answered within 10 seconds, at least 80% of B calls are answered 20 seconds, and at least 80% of C calls are answered 30 seconds. This type of constraint is more difficult to evaluate, since their

evaluation drastically depends on the call routing policies, as we will discuss below.

If our call center has these QoS constraints, then the management should value A-calls the most and C-calls the least, while B-calls should be valued in between the two. We have seen above that the performance of a call center strongly depends on "routing policies". Moreover, in the simple example described above, we know that a policy of "threshold" type is optimal. Now we will describe a "threshold" routing policy for our call center, which preserves the priorities of the QoS constraints. As mentioned in the beginning of this section, the routing policy has two functions, one to direct the incoming calls to an agent or delay them, the other to assign a call waiting in line to an agent who has just become available, or keep her/him idle. Assume that our routing policy directs an incoming A-call to any available agent with the right skill; an incoming B or C call to any available agent with skills $\{B, C\}$; and delays them if all agents in pool $\{B, C\}$ are busy. An agent in pool $\{A, B\}$ or $\{C, A\}$ who just finishes his/her service takes an A-call if there is one in queue; if not, s/he serves a B or C call, respectively, only if the number of B or C calls is greater than a threshold. If an agent in pool $\{B, C\}$ becomes idle, s/he first checks the B-queue, and starts serving a B-call if any; if not, s/he checks the C-queue and serves a C-call if there is any. With such a routing policy, the QoS for A-calls will always be satisfied given that there is sufficient capacity; but QoS constraints for B and C calls may not be satisfied at all, as they have such a low priority in routing. If the routing policy is not modified, the staffing levels may be very high to achieve the specified QoS constraints. Hence, the optimal staffing levels from each skill set depend strongly on "routing policies". Here we also would like to note that the performance measures of a call center with a routing policy as described above do not have a closed form solution. In general, there are two methods to understand whether the QoS constraints are satisfied: to simulate the whole system, or to derive good approximations under general conditions, where the former is computationally intensive, and the latter is very difficult. There are several

papers that approximate certain performance measures (see e.g., Franx, Koole and Pot; 2006, Koole, Pot, and Talim; 2003, Shumsky; 2004), but usually an easier routing policy is assumed.

In skill-based routing (SBR) problems, the performance measures of the call center are not restricted to the ones described in Section 4.1. Some of the common objectives used in SBR are to minimize total waiting costs of all calls, to maximize the revenue generated by all calls, to maximize throughput of calls from a certain class while satisfying certain QoS constraints for other classes. Section 5 of Gans et al. (2003) has a comprehensive review of SBR. Here, we will consider only the most recent work.

One line of research concentrates on call centers which have one dedicated station for each call type, and one fully-flexible station with agents who can serve all call types. Örmeci (2004) considers optimal dynamic admission control of such a call center with no waiting room and two kinds of calls. The objective is to maximize the total revenue generated. It is shown that a call should be routed to a dedicated station if possible, and the optimal admission policy to the fully-flexible facility is of threshold type. Chevalier et al. (2004) generalizes the first part of this result to call centers with more than two classes of calls. Koole and Pot (2005), on the other hand, considers a call center with an infinite waiting room and possibly more than two classes of calls. They use the technique of approximate dynamic programming to find good call routing heuristics. Finally, Bhulai (2005) considers a call center with no waiting room and several stations having general sets of skills, as opposed to call centers with one fully-flexible and many dedicated stations. He uses approximate dynamic programming, as Koole and Pot (2005), to find dynamic call routing policies.

Nowadays, we observe that call centers are turning into contact centers. Hence, different types of services, such as email or call-back, have been emerging. As a result, agents are being cross-trained in these services in addition to the traditional job of answering incoming phone calls. This leads to *call blending* problems: Several papers analyze the effects of call blending on the performance of

call centers, and develop efficient blending policies (Gans and Zhou, 2003; Bhulai and Koole, 2003; Armony and Maglaras, 2004a, 2004b).

Another growing issue in the call center industry is outsourcing: according to Datamonitor, the total value for US outsourcing only will be almost $ 24 billion by 2008, compared to the current $ 19 billion (Datamonitor, 2004). In this case, routing of calls between the outsourcing firm and in-house call centers becomes an important issue, analyzed by Gans and Zhou (2004).

4.3 Staffing in call centers with cross-training

In this subsection, we still consider a call center serving different types of calls with a fixed skill set design, but now we take the routing policy given, and aim to find optimal staffing levels. We will continue using the same call center introduced in Section 4.2, so that our call center is offering three different services, i.e., the set of all possible kinds of calls is $\{A, B, C\}$. We have been using the term "skill set" to give the general idea, but in this section we need to be more precise. Hence, we explain it using the above example: The agents working in our call center may be trained for answering only one call type, or they may be cross-trained for two or three call types. We specify an agent's cross-training level with his/her skill set, where the skill set is defined as the set of call types that an agent can answer. For example, if an agent is trained to answer B and C calls, then his/her skill set is $\{B, C\}$. Hence, all possible skill sets in our example are $\{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{C, A\}, \{A, B, C\}$. In the previous section, we have identified methods to find a good skill set design. To illustrate the link of this section with the previous ones, we assume that we have analyzed this call center at the strategic level, and we have chosen the skill set as $\mathcal{S} = \Big\{\{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{C, A\}\Big\}$, i.e., all agents will have at most 2 skills. Therefore, when we compute the number of required agents from each skill set, we will never consider to have an agent with 3 skills.

In the remaining parts of this subsection, we assume that a skill set design, $\mathcal{S}$, is given. Let the number of elements in $\mathcal{S}$ be s, and $\mathcal{S}_k$ be the kth element of $\mathcal{S}$, where $k = 1, 2, ..., s$. Our aim in this subsection is to find the number of operators from each skill set $\mathcal{S}_k$ for short time intervals (e.g. for half-hours) to guarantee a certain service level during those intervals with a minimal cost.

Çezik and L'Ecuyer (2004) use the methodology of Atlason et al. (2004) to find optimal staffing levels for each half hour with the objective of minimizing the staffing costs of a multi-skill call center. The call center has an infinite waiting room, and the calls may or may not abandon the system. There is always a global QoS constraint so that at least $p\%$ of all calls have to be answered in w seconds; a QoS constraint for each different type of call can also be added. The routing policy is a non-idling policy with simple priorities: an incoming i-call checks for an available server in skill sets, $\{\mathcal{S}_k : i \in \mathcal{S}_k\}$, by a fixed numeric order; and when an agent in skill set $\mathcal{S}_k$ becomes idle, s/he checks the waiting line for calls that s/he can answer according to a given numeric order. As they note, this kind of routing policy makes the system higly unbalanced, and the service level of low-priority call types tends to be very low. The problem is solved by an iterative cutting-plane algorithm on an integer program, where QoS constraints are estimated by simulations. Due to computational complexities, especially in large problems, finding an optimal solution is not always possible, so they also propose practical heuristics.

Bhulai, Koole and Pot (2005) consider the same problem with Çezik and L'Ecuyer (2004). The main difference is that the QoS constraints are estimated by approximations based on steady state behavior of Markovian queueing systems. This reduces the computational time needed to solve the problem, which allows them to consider shift-scheduling problem in addition to staffing.

Wallace and Whitt (2005) use simulation to find optimal staffing levels as well as optimal number of trunk lines, where the objective is to minimize the total number of agents first and to minimize total trunk lines next while satisfying certain QoS constraints. The routing policy is

still a simple priority policy, but they also differentiate the skill level of agents. For example, an agent with skills $\{B, C\}$ has a primary skill of B and a secondary skill of C, whereas an agent with skills $\{C, B\}$ has a primary skill of C and a secondary skill of B. Hence, in general, they assume that each agent has a primary skill (a level of 1), and may have skills at levels 2, 3, etc. When we let the skill sets depend on the order of skills, our framework represents this situation as well. They use an Erlang formula to find an initial value for the total number of servers, then square root staffing to allocate these servers to different primary skills. Finally they describe a rule to add skills to these servers. After specifying the initial solution, they use simulation to improve the solution. They numerically show via simulation that a call center with all agents having at most two skills performs very closely to a call center with fully-flexible servers. Mazzuchi and Wallace (2004) conducts an experimental design on this call center, where they investigate the effects of call volume and of the number of skills of each agent on certain performance measures. They reach the same conclusion as Wallace and Whitt (2005). Sisselman and Whitt (2006), on the other hand, use this framework to incorporate the expected value of certain call types generated by certain kinds of agents, in particular the preference of agents to answer certain types of calls, in call routing.

The conclusions of these papers confirm the long-time observation of "little flexibility goes a long way", and what they suggest is to cross-train the agents in at most two skills. However, implementing this suggestion in all call centers may not be possible. Certain types of call centers, such as technical support call centers and call centers of certain banks, require the skill sets to be nested: The entry level agents know only the basics, and adding one more skill means "learning the subject one level deeper". Then, we can label the skills as 1,...,n, where level 1 is the entry level, and level n is the "guru" level. In this kind of a system, the agents cannot be cross-trained in at most two skills. Hence, we probably need different tools to analyze this kind of a system. Finally, we would like to note another issue that brings a nested skill set into picture: the necessity

of offering a career plan to the employees. When the cross-train levels are set to two, there is no career path for the agents, which will take them to a "better" position in the organization.

Chevalier et al. (2004), as mentioned above, consider a call center which has one dedicated station for each call type, and one fully-flexible station with agents who can serve all call types. They assume that an incoming call is served in its dedicated station if possible; if not, it is directed to an available agent in the full-flexible station; and if that station is also full, then the call is lost (due to no waiting room). We note that directing calls to the dedicated stations first is shown to be optimal. They show, through a numerical study, that spending 80% of the staffing budget on the dedicated agents (and so the remaining 20% spent on the fully-flexible station) works well over a wide range of parameters in systems with unlimited waiting space as well as in those with no waiting room.

4.4 Routing and staffing in call centers with cross-training

In sections 4.2 and 4.3, we have seen that the routing policies have a strong effect on the performance measures and so on the staffing levels. Hence, if the staffing problem is solved in combination with the optimal routing problem, significant improvements can be achieved. In this subsection, we still consider a call center serving different types of calls with a fixed skill set design, $\mathcal{S}$, but now we aim to find an optimal routing policy as well as optimal staffing levels. The combined problem of routing and staffing is considerably difficult, so the studies on this subject are based on fluid approximations, which ignores stochastic fluctuations observed in queueing systems.

Harrison and Zeevi (2005) and Bassamboo, Harrison and Zeevi (2004) formulate the problem as a two-stage stochastic linear program with recourse. At the first stage the staffing levels are determined, while in the second stage the routing problem is optimized by using a fluid approximation of the call center. The second stage observes the random features of the system, so that the

corresponding expected value is approximated via Monte Carlo simulation. In this way, the daily call patterns (see Figure 4) can be incorporated in the model. In the routing problem, the objective is to minimize total penalty due to abandonment, whereas the overall objective is to minimize the staffing costs and the expected abandonment penalty. Bassamboo et al. (2004), on the other hand, derive an asymptotic lower bound on the expected total cost for the same problem, and propose a staffing and routing method, which is shown to achieve this asymptotic lower bound. Whitt (2006b) is another study which proposes a fluid model to solve routing and staffing problems simultaneously, where two optimization problems are constructed based on the fluid approximations. He also discusses how to implement stochastic fluctuations and dynamic routing.

5 Future Directions

As described in Section 3, the literature exploring the skill set design question assumes given or identical capacity in the different skill types. It is described how properties of superior cross-training structure are quite robust to underlying system characteristics. So irrespective of the operating details of their systems, managers can follow some of these guidelines in determining cross-training policies. However, this analysis is performed taking a benefits perspective. Determining the appropriate structure for a given call center requires understanding costs in addition to benefits. The operational costs of a flexibility design however depend on capacity that is deployed within the designed skill-set structure. Section 4 illustrates that for a call center performance is closely tied to capacity and the way it is eventually exploited through routing decisions. Linking skill-set design to capacity choice in call centers or other systems is an important direction for future research, which would enable systematically making a tradeoff between the benefits from flexibility and the costs of staffing. For example the flexibility literature recommends structures having balanced flexibility with capacity pools that have the same number of skills. Are such structures still superior to others

when the cost of capacity is incorporated in the analysis? These costs will be characterized more precisely as research focusing on staffing and routing develops. In particular approaches that are capable of jointly optimizing staffing and routing will enable a better assessment of costs. This is another important area for future research.

Another direction for investigation is one that adopts an interdisciplinary approach to cross-training design and combines motivational, biological and perceptual-motor aspects of the issue with operational ones. Empirical research that explores the motivational impact of the superior structures in Section 3, or that helps define better routing policies as those that improve operational as well as motivational performance is needed for different call center environments. Other questions lying at the interface of operations and human resource management are: what is the relationship between these flexibility structures and career paths? How can career paths be formulated such that skill sets come closer to designs that are known to be operationally superior?

Acknowledgment: The authors would like to thank Tolga Çezik, Gül Gürkan, Ger Koole, Robert Shumsky and Ward Whitt for their comments on an earlier version.

References

Akşin O.Z. and Harker P.T., (1999). "To Sell or Not to Sell: Determining the Tradeoffs Between Service and Sales in Retail Banking Phone Centers", *Journal of Service Research*, 2:1 19-33.

Akşin, O.Z. and Karaesmen, F. (2002). "Designing Flexibility: Characterizing the Value of Cross-Training Practices", Working Paper, Koç University.

Akşin, O.Z. and Karaesmen, F. (2005). "Characterizing the Performance of Process Flexibility Structures", Working Paper, Koç University.

Akşin, O.Z., Karaesmen, F. and Örmeci, E.L. (2005). "On the Interaction Between Resource Flexibility and Flexibility Structures", In *Proceedings of the Fifth International Conference on*

'Analysis of Manufacturing Systems - Production Management'.

Armony, M. and Maglaras, C., (2004a). "On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules, and System Design", *Operations Research*, 52:2, 271-292.

Armony, M. and Maglaras, C., (2004b). "Contact Centers with a Call-Back Option and Real- Time Delay Information", *Operations Research*, 52:4, 527-545.

Armony, M., and Mandelbaum, A., (2004). "Design, Staffing and Control of Large Service Systems: The Case of a Single Customer Class and Multiple Server Types", Working paper, Israel Institute of Technology.

Atlason, J., Epelman, M.A., and Henderson, S.G. (2004). "Call Center Staffing with Simulation and Cutting Plane Methods", *Annals of Operations Research*, 127, 333-358.

Avramidis, A. N., Deslauriers, A. and P. LEcuyer, P., (2004). "Modeling daily arrivals to a telephone call center", *Management Science*, 50:7, 896908.

Bassamboo, A., Harrison, J. M., and Zeevi A. (2004). "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method", Technical Report.

Baumgartner, M., Good, K., and Udris, I. (2002). "Call Centers in der Schweitz", *P*sychologische Untersuchungen in 14 Organisationen, (Call Centers in Switzerland, Psychological Investigations in 14 Organisations) Zurich, Switzerland.

Bhulai, S. and Koole, G. (2003). "A Queueing Model for Call Blending in Call Centers", *IEEE Transactions on Automatic Control*, 48:8, 1434-1438.

Bhulai, S., Koole, G. and Pot A. (2005). "Simple Methods for Shift Scheduling in Multi-skill Call Centers", Working Paper, Free University, The Netherlands.

Bhulai, S. (2005). "Dynamic Routing Policies for Multi-Skill Call Centers", Working Paper, Free University, The Netherlands.

Borst, S., Mandelbaum, A., and Reiman M. (2004). "Dimensioning Large Call Centers", *Operations*

Research, 52:1, 17-34.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., (2005). "Statistical analysis of a telephone call center: A queueing-science perspective", *J*ournal of the American Statistical Association, 100, 3650.

Buzacott, J.A. (1996). "Commonalities in Reengineered Business Processes: Models and Issues", *Management Science*, 42:5, 768-782.

Campion, M.A. and McClelland, C.L. (1991). "Interdisciplinary Examination of the Costs and Benefits of Enlarged Jobs: A Job Design Quasi-experiment", *Journal of Applied Psychology*, 76, 186-198.

Campion, M.A. and McClelland, C.L. (1993). "Follow-up and Extension of the Interdisciplinary Costs and Benefits of Enlarged Jobs", *Journal of Applied Psychology*, 78:3, 339-351.

Çezik, T., and L'Écuyer P. (2004). "Staffing Multiskill Call Centers via Linear Programming and Simulation", Technical Report, Université de Montréal.

Chevalier P., Shumsky R.A., and Tabordon, N. (2004). "Routing and Staffing in Large Call Centers with Specialized and Fully Flexible Servers", Working paper.

Dantzig, G. B. (1954). "A Comment on Edie's 'Traffic Delays at Toll Booths' ", *Operations Research*, 2:3, 107-138.

Das, A. (2003). "Knowledge and Productivity in Technical Support Work", *Management Science*, 49:4, 416-431.

Datamonitor (2002), Cited in Call Center Management Review, May 2002, http://www.ccmreview.com

Datamonitor (2004), 4-26-2004, Cited in http://www.incoming.com/statistics

Datamonitor (2005), 4-14-05, Cited in http://www.incoming.com/statistics

Feldman, Z., Mandelbaum, A., Massey, W.A., and Whitt W. (2006). "Staffing of Time-Varying Queues to Achieve Time-Stable Performance", *Management Science*, forthcoming.

Fine, C.H. and Freund, R.M. (1990). "Optimal Investment in Product-Flexible Manufacturing Capacity", *Management Science*, 36:4, 449-466.

Evenson, A., Harker, P.T. and Frei, F.X. (1999). "Effective Call Center Management: Evidence from Financial Services". Working Paper 9925B, Wharton Financial Institutions Center.

Franx G. J., Koole G. M., and Pot S. A. (2006). "Approximating multi-skill blocking systems by hyperexponential decomposition", *Performance Evaluation*, forthcoming.

Gans, N., Koole, G. M. and Mandelbaum, A. (2003). "Telephone Call Centers: Tutorial, Review, and Research Prospects", *Manufacturing & Service Operations Management*, Vol. 5, 97-141.

Gans, N. and Zhou, Y. (2003). "A Call-Routing Problem with Service-Level Constraints". *Operations Research*, 51, 255-271.

Gans, N. and Zhou, Y. (2004). "Overflow Routing for Call Center Outsourcing", Working Paper, The Wharton School.

Garnett, O., Mandelbaum, A., and Reiman M. (2002). "Designing a Call Center with Impatient Customers", *Manufacturing and Service Operations Management*, 4:3, 208-227.

Graves, S. C. and Tomlin B.T. (2003). "Process Flexibility in Supply Chains", *Management Science*, 49, 907-919.

Grebner, S., Semmer, N.K., Lo Faso, L., Gut, S., Kalin, W., and Elfering, A. (2003). "Working Conditions, Well-Being, and Job-Related Attitudes Among Call Center Agents", *European Journal of Work and Organizational Psychology*, 12:4, 341-365.

Green L. V., Kolesar, P. J., and W. Whitt (2006). "Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System", *Production and Operations Management (POMS)*, forthcoming.

Güneş, E.D., and Akşin, O.Z. (2004). "Value Creation in Service Delivery: Relating Market Segmentation, Incentives, and Operational Performance", *Manufacturing and Service Operations*

Management, 6:4, 338-357.

Gurumurthi, S. and Benjaafar S. (2004). "Modeling and Analysis of Flexible Queueing Systems", *Naval Research Logistics*, 51, 755-782.

Hackman, J.R., and Oldham, G.R. (1976). "Motivation through the Design of Work: Test of a Theory", *Organizational Behavior and Human Performance*, 16, 250-279.

Harrison, J.M., and Zeevi A. (2005). "A Method for Staffing Large Call Centers Based on Stochastic Fluid Models," *Manufacturing & Service Operations Management*, 7:1, 20-36.

Hasija, S., Pinker, E. and Shumsky, R. (2005). "Staffing and Routing in a Two-Tier Call Center", forthcoming *International Journal of Operational Research*.

Henderson S., and Mason A. (1998). "Rostering by Iterating Integer Programming and Simulation". In *Proceedings of the 1998 Winter Simulation Conference*, 677-683.

Hopp, W.J., Tekin, E., and Van Oyen, M.P. (2004). "Benefits of Skill Chaining in Production Lines with Cross-Trained Workers", *Management Science*, 50, 83-98.

Hopp, W.J. and Van Oyen, M.P. (2004). "Agile Workforce Evaluation: a Framework for Cross-Training and Coordination", *IIE Transactions*, 36:10, 919-940.

ICCM Weekly (2002), 4-3-2002, Cited in http://www.incoming.com

Ilgen, D.R. and Hollenbeck, J.R. (1991). "The Structure of Work: Job design and Roles",. In M.D, Dunnette and L.M. Hough, editors, *Handbook of Industrial and Organizational Psychology*, 2, 165-207.

Incoming Calls Management Institute (ICMI, 2000), 9-1-2000, Cited in Call Center Management Review, September 2000, http://www.ccmreview.com

Incoming Calls Management Institute (ICMI, 2002), 7-1-2002 Multichannel Call Center Study Final Report, cited in Call Center Management Review, July 2002, http://www.ccmreview.com

Inman R.R., W.C. Jordan, and D. E. Blumenfeld (2004). "Chained Cross-Training of Assembly

Line Workers", *International Journal of Production Research*, 42:10, 1899-1910.

Iravani, S.M., Van Oyen, M.P., and K.T. Sims (2005). "Structural Flexibility: A New Perspective on the Design of Manufacturing and Service Operations", *Management Science*, 51, 151-166.

Isic, A., Dormann, C., and Zapf, D. (1999). "Belastungen und Resources an Call Center Arbeitsplatzen" (Job stressors and resources among call center employees), *Zeitschrift fur Arbeitswissenschaft*, 53, 202-208.

Jennings, O., Mandelbaum, A., Massey, W., and Whitt W. (1996). "Server Staffing to Meet Time-Varying Demand", *Management Science*, 42:10, 1383-1394.

Jordan, W.C. and Graves, S.C. (1995). "Principles on the Benefits of Manufacturing Process Flexibility", *Management Science*, 41:4, 577-594.

Jordan, W.C., R.R. Inman, and D. E. Blumenfeld (2004). "Chained Cross-Training of Workers for Robust Performance", *IIE Transactions*, 36, 953-967.

Jouini O., Dallery Y. and Nait- Abdallah R. (2004). "Analysis of the Impact of Team-Based Organizations in Call Center Management ", *Working Paper, Ecole Centrale Paris*.

Koole G. M., and Pot S. A. (2005). "Approximate Dynamic Programming in Multi-Skill Call Centers", In *Proceedings of the 2005 Winter Simulation Conference*.

Koole G. M., Pot S. A., and Talim J. (2003). "Routing Heuristics for Multi-Skill Call Centers", In *Proceedings of the Winter Simulation Conference*, 1813-1816.

Koole G., and van der Sluis, E. (2003). "Optimal Shift Scheduling with a Global Service Level Constraint", *IIE Transactions*, 35, 1049-1055.

Mandelbaum A., and Reiman, M. I. (1998). "On Pooling in Queueing Networks", *Management Science*, 44:7, 971-981.

Mazzuchi, T.A. and Wallace, R.B. (2004). "Analyzing Skill-Based Routing Call Centers Using Discrete-Event Simulation and Design Experiment", in *Proceedings of the 2004 Winter Simulation*

Conference, R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

Netessine, S., Dobson, G. and Shumsky, R. (2002). "Flexible Service Capacity: Optimal Investment and the Impact of Demand Correlation", *Operations Research*, 50:2, 375-389.

Örmeci, E.L. (2004). "Dynamic Admission Control in a Call Center with One Shared and Two Dedicated Service Facilities", *IEEE Transactions on Automatic Control*, 49:7, 1157-1161.

Pinker, E.J. and Shumsky, R.A. (2000). "The Efficiency-Quality Trade-Off of Cross-Trained Workers", *Manufacturing & Service Operations Management*, 2:1, 32-49.

Sethi, A.K. and Sethi, S.P. (1990). "Flexibility in Manufacturing: a Survey", *International Journal of Flexible Manufacturing Systems*, 2, 289-328.

Sheikzadeh, M., Benjaafar S., and Gupta, D. (1998). "Machine Sharing in Manufacturing Systems: Total Flexibility versus Chaining", *International Journal of Manufacturing Systems*, 10, 351-378.

Shumsky, R.A. (2004). "Approximation and Analysis of a Queueing System with Flexible and Specialized Servers", *OR Spektrum*, Special Issue on Call Center Management, Vol. 26, No. 3.

Shumsky, R.A. and Pinker, E.J. (2003). "Gatekeepers and Referrals in Services", *Management Science*, 49:7, 839-856.

Sisselman, M. E., and Whitt, W. (2006). "Value-Based Routing and Preference-Based Routing in Customer Contact Centers. *Production and Operations Management*, forthcoming.

Steckley, S.G., Henderson, S.G. and Mehrotra, V. (2004). "Service System Planning in the Presence of a Random Arrival Rate", Technical Report, Cornell University.

Ingolfsson, A., Cabral E., and Wu, X. (2005). "Combining Integer Programming and the Randomization Method to Schedule Employees", Technical Report, University of Alberta.

Taylor, P., Mulvey, G., Hyman, J., and Bain, P. (2002). "Work Organization, Control and the Experience of Work in Call Centers", *Work, Employment, and Society*, 16, 122-150.

Van Mieghem, J.A. (1998). "Investment Strategies for Flexible Resources", *Management Science*,

44, 1071-1078.

Van Oyen, M.P., Gel, E.G.S., and Hopp, W.J. (2001). "Performance Opportunity for Workforce Agility in Collaborative and Noncollaborative Work Systems", *IIE Transactions*, 33:9, 761-777.

Wallace R.B. and Whitt W. (2005). "A Staffing Algorithm for Call Centers with Skill-based Routing," *Manufacturing and Service Operations Management (M&SOM)*, 7, 276-294.

Whitt W. (2006a). "Staffing a Call Center with Uncertain Arrival Rate and Absenteeism," *Production and Operations Management*, 15:1, 88-102.

Whitt W. (2006b). "A Multi-Class Fluid Model for a Contact Center with Skill-Based Routing." *International Journal of Electronics and Communications (AEU)*, 60:2, 95-102.

Xie, J.L. and Johns, G. (1995). "Job Scope and Stress: Can Job Scope be too High?", *Academy of Management Journal*, 18, 1288-1309.

Xu, S. H., Righter R., and Shanthikumar J. G. (1992). "Optimal Dynamic Assignment of Customers to Heterogeneous Servers in Parallel", *Operations Research*, 40, 1126-1138.