PRODUCTION AND OPERATIONS MANAGEMENT POMS
Vol. 24, No. 7, July 2015, pp. 1086–1100
ISSN 1059-1478 | EISSN 1937-5956 | 15 | 2407 | 1086
DOI 10.1111/poms.12333

Flexibility Structure and Capacity Design with Human Resource Considerations

O. Zeynep Akşin
College of Administrative Sciences and Economics, Koç University, 34450, Sariyer, Istanbul Turkey, zaksin@ku.edu.tr

Nesrin Çakan, Fikri Karaesmen, E. Lerzan Örmeci
Department of Industrial Engineering, Koç University, 34450, Sariyer, Istanbul, Turkey
ncakan@gmail.com, fkaraesmen@ku.edu.tr, lormeci@ku.edu.tr

Most service systems consist of multidepartmental structures with multiskill agents that can deal with several types of service requests. The design of flexibility in terms of agents' skill sets and assignments of requests is a critical issue for such systems. The objective of this study was to identify preferred flexibility structures when demand is random and capacity is finite. We compare structures recommended by the flexibility literature to structures we observe in practice within call centers. To enable a comparison of flexibility structures under optimal capacity, the capacity optimization problem for this setting is formulated as a two-stage stochastic optimization problem. A simulation-based optimization procedure for this problem using sample-path gradient estimation is proposed and tested, and used in the subsequent comparison of the flexibility structures being studied. The analysis illustrates under what conditions on demand, cost, and human resource considerations, the structures found in practice are preferred.

Key words: flexibility; call centers; multidimensional newsvendor; gradient estimation via perturbation analysis
History: Received: October 2008; Accepted: October 2014 by Michael Pinedo, after 3 revisions.

1. Introduction

It is known that flexible resources, such as cross-trained servers or flexible equipment, mitigate the negative effect of demand uncertainty. This is because flexible resources can be allocated to different tasks, thereby adjusting supply to better meet uncertain demand. Especially for service systems like call centers, where production and consumption occur simultaneously, flexibility is essential in meeting and satisfying customer needs.

Demand uncertainty can come in different forms. Uncertain arrival rates imply unknown demand parameters for the underlying capacity models of service systems. Thus, forecasts and the associated forecast errors are necessary to measure demand parameter uncertainty. For a given arrival rate, there is also demand uncertainty due to stochastic variability, which is modeled by an arrival process to the service system. Both parameter uncertainty and stochastic variability affect capacity choices that lead to desired performance in terms of service levels. Both effects imply the need for safety capacity and flexibility. For many call centers, demand uncertainty in the form of unknown arrival rates dominates the uncertainty that results from stochastic variability. Motivated by this observation, this study focuses on flexibility design in settings with parameter uncertainty.

Through resource flexibility, demand is aggregated and resources are shared. Therefore, flexibility improves the utilization of a system. Nevertheless, it is usually expensive. Cross-training a server has a cost. Furthermore, additional skills typically require higher compensation. While in some settings broadening the scope of a server through cross-training can have positive effects on employee well-being and productivity, there are limits to this beyond which additional flexibility may be detrimental to individuals. This latter effect imposes constraints on the flexibility design problem which tries to determine the appropriate skill sets for employees, and the number of employees in each skill set, such that the benefits of flexibility are maximized while minimizing its direct and indirect costs.

The flexibility design problem for call centers can be viewed as a hierarchical planning problem. At a strategic level, managers will try to determine the type of flexibility. This will start out with job design which we label as skills' definition. Each skill consists of a set of tasks, and different skill definitions may group the tasks differently. A server who has a particular skill will be able to perform all tasks within that skill. Different skill definitions may be preferred, for example, to enable a natural career progression for agents, to manage the total call volume of a particular skill, or to group tasks by product or service being offered. Given a skill definition, the strategic level will then determine the skill sets for agents. This is where the type of flexibility in a system as a function of the skills that have been defined is determined. The tactical level problem determines the capacity levels for each skill set. Once flexible capacity is in place, the operational control problem which consists of skill-based routing will route incoming calls to the appropriate agent pools, thereby exploiting the flexible capacity. In this study, we will mostly focus on the tactical problem of capacity optimization for a given flexibility structure, which will enable us to make statements about preferred choices (from a profit standpoint) at the strategic level in terms of different flexibility structures. The operational problem is approximated as a linear program in the ensuing analysis, thus disregarding stochastic variability.

Our motivation to compare different flexibility structures in terms of their profit implications stems from observing different flexibility structures in real call centers. These have not all been analyzed from a profit perspective in the extant literature. In particular, the flexibility literature, building on the work by Jordan and Graves (1995), suggests that a two-skill complete chain, where each server has two skills, with its capacity appropriately optimized, would be an ideal structure for a call center. A chain is formally defined as a group of demand and resource types that are either directly or indirectly connected by demand-resource assignment decisions. A complete chain structure allows reallocation of demands within the resources of the whole system. The performance of the two-skill complete chain has been explored and confirmed in Wallace and Whitt (2005), where the capacity is determined taking staffing costs into account. In practice, flexibility structures are developed by also taking certain human resource-related issues into account, which are not explicitly considered in the flexibility literature.

A structure which we call the *nested structure* is typical in many banking and insurance contact centers. Nested structures are adopted in call centers where career planning is extremely important. The employees have a high profile and a high potential to learn different tasks. Their aim is to be promoted to higher positions after a certain amount of contact center experience. The nested structure implies a natural career progression from being inexperienced agents with limited skills to becoming experienced multiskill agents by learning additional skills over time. With an appropriate skill definition upfront, this progression can be made from simple or fundamental to more complex or advanced skills: The most standard operations can be taught easily to the beginners, thus agents start out with these skills. As their contact center experience grows, the agents are trained to respond to additional more complex queries. This type of a progressive training structure results in what we call a nested structure. Apart from the entry level agents, all agents in such a call center will be cross-trained in several skills.

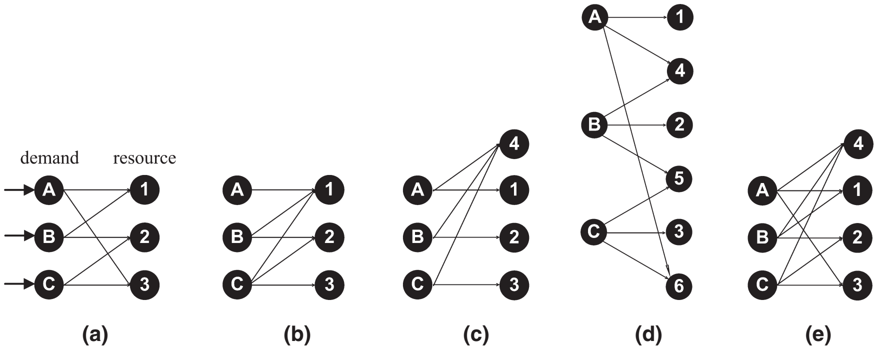
The second type of flexibility structure, observed in technical support providing call centers is labeled as the *overflow structure*. These centers have two kinds of employees, dedicated or non-flexible agents who focus on only one type of skill, and expert or flexible agents who can provide assistance requiring all or several types of skills. This structure seems to acknowledge the cost and difficulty of cross-training all employees and instead adopts a structure with a small proportion of so-called super-servers.

Finally, we also consider the *adjacent-level-of-flexibility* structures proposed in Bassamboo et al. (2010b). In that study, it is shown that "the optimal structure invests in at most two adjacent levels of flexibility" for systems with symmetric parameters. Here, we consider the two possible adjacent-level flexibility structures in a three-skill setting, one with pools of one-skill and two-skill agents (Adjacent level-12, A12), the other with pools of two-skill and three-skill agents (Adjacent level-23, A23). We note that A12 consists of pools of specialists and a two-skill complete chain, while A23 is a two-skill complete chain combined with a generalist pool as in the overflow structure, however, excluding the specialist pools we see in the overflow structure.

The question of whether one would rather have cross-trained servers throughout a service organization, or focus cross-training activities on an exclusive set of servers has been addressed from a human resource practice standpoint before. Hunter (1999) describes two prevailing models of work organization in retail banking (branch, call center), labeling one as the *inclusive* and the other as the *segmented* model. In terms of cross-training practice, the inclusive model implies cross-training for most employees throughout the organization, whereas the segmented model refers to systems with cross-training for a select few. Viewing the above flexibility structures from this perspective, we can consider the two chain, nested, and A23 structures as examples of what one might find in an organization with inclusive work practices, whereas an overflow structure or A12 would suggest segmented work practices.

Our research objective is to compare the two-skill chain structure, and the adjacent-level flexibility structures recommended by the literature with the nested and overflow structures found in practice. Figure 1 shows the five flexibility structures to be

Figure 1 Different Flexibility Structures (a) Two-Skill Complete Chain (b) Nested (c) Overflow (d) Adjacent Level A12 (e) Adjacent Level A23



analyzed in this study for the case when there are three different types of skills (used synonymously with call types or products throughout). We aim to answer the following research questions: When capacity is optimized to maximize profits, can the nested or overflow structures achieve the performance of the recommended structure of the flexibility design literature? Under what conditions might the former two structures that originate from some human resource-related concerns be preferred?

We first develop a methodology to solve the capacity optimization problem for any given flexibility structure. This methodology is then used to investigate the flexibility structures specified above. The remaining parts of this study are organized as follows: A review of the literature in section 2 is followed by a presentation of the model in section 3. Section 4 introduces the solution method and analyzes its theoretical background. Some benchmarks from the literature are used to numerically verify this solution method. Section 5 proceeds with a numerical study that compares the flexibility structures in different environments. Section 5.1 describes the experimental design and section 5.2 gives a detailed account for the comparison of the flexibility structures. Following a summary of the numerical results in section 5.3, section 5.4 presents a case study from a banking call center where the insights from the analysis have been used to propose alternative flexibility structures. Using demand data from this call center, we demonstrate that the proposed structures outperform the current flexibility structure in place at this call center. Analysis of the demand data also illustrates that this call center indeed operates in an uncertainty dominated regime, as assumed in the study. The numerical examples and case study illustrate how capacity optimization reduces the importance of flexibility design, and how specialist servers help to control costs. The study ends with concluding remarks in section 6.

2. Related Literature

The operations management literature has mostly focused on the benefits of different flexibility structures, not explicitly dealing with its costs. Within the spectrum of full-flexibility and full-specialization, a variety of limited-flexibility structures can be built. Jordan and Graves (1995) are the first to develop principles on the benefits of process flexibility, showing that well-designed limited flexibility can be as good as full flexibility. Later on, Akşin and Karaesmen (2002, 2007) and Iravani et al. (2005, 2007) analytically justify these principles, and propose methods to evaluate different flexibility structures.

From a throughput maximization perspective, a limited flexibility structure called a complete chain, first proposed by Jordan and Graves (1995), has been shown to perform almost as well as the fully flexible structure in a variety of different settings. The benefits of chaining have been explored by many: Sheikzadeh et al. (1998) and Jordan et al. (2004) analyze chaining within manufacturing systems, Graves and Tomlin (2003) within multistage systems, Inman et al. (2004) within assembly lines, Gurumurthi and Benjaafar (2004) within service systems, Hopp et al. (2004) and Van Oyen et al. (2001) within both service and manufacturing systems, Akşin and Karaesmen (2002, 2007) and Wallace and Whitt (2005) within contact centers, and Andradottir et al. (2013) within networks.

Both Jordan and Graves (1995) and Akşin and Karaesmen (2007) show the close relationship of flexibility to capacity design, suggesting that flexibility and capacity need to be jointly designed. Capacity optimization prevents holding unnecessary flexible capacity, which consequently decreases the cost of the system and prevents waste of highly qualified resources. While the literature focusing on the flexibility design problem typically assumes that capacity is fixed, capacity optimization problems have mostly focused on given, relatively simple flexibility structures (see, e.g., Chevalier et al. 2004, Fine and Freund 1990, Harrison and Zeevi 2005, van Mieghem 1998, Netessine et al. 2002). Exceptions are Chevalier and Van den Schrieck (2008), Bassamboo et al. (2012), and Bassamboo et al. (2010b).

In the setting being considered herein, capacity is provided by human servers. While we focus on flexibility benefits in terms of increased system throughput, and flexibility costs in terms of direct staffing costs for servers, actual benefits and costs may also include human resource-related ones like motivational effects, mental load implications, career paths, etc. These costs and benefits are reviewed in Akşin et al. (2007b). That study and Akşin et al. (2007a) provide detailed reviews of the call center flexibility design problem and illustrate its close ties to human resource management. While we do not consider any of these issues explicitly in our modeling, we consider some of them in developing our managerial implications at the end.

In this study, we formulate a two-stage stochastic optimization model, as, for example, in Fine and Freund (1990), van Mieghem (1998), Harrison and van Mieghem (1999), and propose a solution method based on the gradient estimation via perturbation analysis (GPA) technique. A detailed exposition of GPA can be found in Glasserman (1991). Some theoretical issues are investigated by Robbins and Monroe (1951), Talluri and van Ryzin (1999), and Karaesmen and van Ryzin (2004).

3. Model Formulation

Capacity in service systems is typically modeled by a queue. In a queueing model, assumptions are made about the arrival process and the service process, and the system's performance is analyzed under given parameters. In this framework, the arrival process is frequently modeled as a Poisson process with known arrival rates (λ). Capacity is determined based on a square-root safety principle that adds safety capacity which is proportional to the square root of this known arrival rate.

In practice, it has been documented that call centers face arrivals that are more variable than a Poisson process (see, e.g., Avramidis et al. 2004, Steckley et al. 2005). One way of modeling a process that is more variable than a Poisson process is to imagine that the arrival process is actually a Poisson process with an arrival rate that varies according to some random process. A typical interpretation is to view this randomness as a consequence of forecasting difficulties (Steckley et al. 2005). We model arrival rates as a random variable with a specified distribution herein.

Instead of a queuing model, we adopt a newsvendor-like capacity model that faces random arrival rates. Through this choice, we disregard the stochastic variability effect on capacity. As argued in Akşin et al. (2008), and later shown in Bassamboo et al. (2010a), this type of a newsvendor-like capacity model is appropriate and accurate when the uncertainty in the demand rate is more significant than the stochastic variability. In particular, Theorem 1 of Bassamboo et al. (2010a) shows that when the coefficient of variation of the random arrival rates is larger than $1/\sqrt{\epsilon}$, where $\varepsilon = \lambda/\mu$ is the offered load of the system with μ the average service rate, a capacity prescription that follows a newsvendor model is nearly optimal. In such settings, there is not much to be gained by adding a safety capacity as suggested by the square root staffing rule. On the other hand, if the condition does not hold, then the newsvendor prescription could be further refined through square root safety staffing. Our analysis focuses on such settings where uncertainty dominates stochastic variability. We present a case study from a call center in section 5.4, where analysis of real demand data demonstrates that uncertainty dominates stochastic variability. As shown in that example, the condition that verifies the dominance of the uncertainty over stochastic variability is only applicable to demand data that does not contain predictable variations, as these can be dealt with via staffing and scheduling.

The capacity design with flexible resources problem is thus modeled as a two-stage stochastic optimization problem. The capacities of the resources are determined in the first stage, prior to the realization of the demand. Realized demand is allocated to the resources in the second stage. The capacity optimization problem is a multidimensional newsvendor-like problem. All of the demand is assumed to be realized at the beginning of the period, and the demand that cannot be processed immediately due to the lack of capacity, is lost.

Consider a service or a manufacturing system with J parallel resources, which processes I different service types. The set of services that a resource can process will be referred to as the skill set of that resource. The skill sets of the resources are different from each other. We represent the flexibility structure of the system by the matrix K, where $k_{ij} = 1$ denotes that resource j has skill i and therefore demand i can be processed at resource j, otherwise $k_{ij} = 0$. The overall service capacity of resource j is denoted by c_j. The demand vector, $\mathbf{D} = (D_1, \ldots, D_I)$, is random with a joint probability density function $g_{\mathbf{D}}(\mathbf{d})$. The demand realization of service i is denoted by d_i. In the context of call centers, the demand realizations represent the mean demand rate. In a more general context, we assume that the demand represents the overall amount of required work, rather than the number of service requests for a service type i. Accordingly, they can have continuous values. We assume that each demand unit requires one unit of resource capacity. Finally, we let x_{ij} be the units of service i processed by resource j.

It is assumed that each unit of service i has an associated revenue p_i per unit, and each specialized resource j has an associated cost s_j per unit capacity. Similar to Chevalier et al. (2004), we assume that flexibility increases the cost of capacity in an amount proportional to the additional skills of the corresponding resource. Thereby, the unit cost of a flexible resource is denoted by the expression $s_j + f_j(\sum_i k_{ij} - 1)$, where f_j denotes the cost of flexibility at resource j for each additional skill. In the context of service systems, f_j represents the payment to resources for additional skills in terms of salary and benefits, which justifies this linear relationship. Then, we can formulate the problem as follows:

$$\text{Stage I: } \max_{\{c_j, x_{ij}\}} \Omega(\mathbf{c}) = \max_{\{c_j, x_{ij}\}} \left\{ E\left[\Phi(\mathbf{c}, \mathbf{D}) - \sum_j c_j s_j - \sum_j c_j \left(\sum_i k_{ij} - 1 \right) f_j \right] \right\}, \tag{1}$$

$$\text{Stage II: } \Phi(\mathbf{c}, \mathbf{d}) = \max_{\{x_{ij}\}} \sum_i \sum_j x_{ij} p_i, \tag{2}$$

$$\text{subject to: } \sum_i x_{ij} \le c_j \quad \forall j, \tag{3}$$

$$\sum_j x_{ij} \le d_i \quad \forall i, \tag{4}$$

$$0 \le x_{ij} \le M \times k_{ij} \quad \forall i, j, \tag{5}$$

where M is a large number, $\mathbf{c} = (c_1, \ldots, c_J)$ and k_{ij} is taken as a parameter. x_{ij} is a decision variable in both stages while c_j becomes a parameter in the second stage. The capacity should be decided at the beginning of the period so that the expected profit of the system, Ω, is maximized. The first term of Equation (1) represents the expected revenue for a given capacity $\mathbf{c}$. The second and the third terms represent the total cost of the capacity. Since the cost is constant for a given capacity value, the first-stage problem can be reformulated as follows:

$$\max_{\{c_j, x_{ij}\}} \Omega(\mathbf{c}) = \max_{\{c_j, x_{ij}\}} \left\{ E[\Phi(\mathbf{c}, \mathbf{D})] - \sum_j c_j s_j - \sum_j c_j \left(\sum_i k_{ij} - 1 \right) f_j \right\} \tag{6}$$

The second stage maximizes the revenue of the system for any demand realization, $\mathbf{d}$, and capacity level, $\mathbf{c}$. Inequality Equation (3) guarantees that the number of jobs handled by any resource is not more than its capacity, whereas inequality Equation (4) prevents the number of processed i jobs from exceeding the corresponding demand. Finally, Equation (5) ensures that the jobs are assigned to the capable resources.

4. The Solution Method

In this section, we propose a solution method to the capacity optimization problem, which is based on the gradient estimation via perturbation analysis technique. We then numerically compare capacity vectors obtained via this method to certain benchmark problems for which the optimal capacities are known.

4.1. GPA Combined with Stochastic Approximation

The GPA technique estimates the gradient of the objective function in consecutive experiments. The decision variable is then changed in the direction of the estimator with a certain step size.

Let $\mathbf{c} = [c_1, \ldots, c_J]$ denote the capacity vector, $\nabla = [(\partial\Omega/\partial c_1), \ldots, (\partial\Omega/\partial c_J)]$ denote the gradient vector of the expected profit with respect to the capacity, and $\tilde{\nabla}$ denote the gradient estimator. Beginning from an arbitrary initial capacity level, the method searches for an optimal level by successively perturbing $\mathbf{c}$ in the direction of $\tilde{\nabla}$ with a certain step size b_k, where k denotes the iteration number.

This approach is similar to the steepest descent algorithm. However, instead of the gradient, an estimator is employed. $\Omega(\mathbf{c})$ depends on the random demand distribution due to the first term of Equation (6), $E[\Phi(\ \mathbf{c},\mathbf{D})]$. However, stage 2 is solved for each realization of $\mathbf{d}$, so that it is not possible to derive the probability distribution or the expected value of $\Phi(\mathbf{c}, \mathbf{D})$. Hence, the exact gradient of $E[\Phi(\mathbf{c},\mathbf{D})]$, $\nabla_{\mathbf{c}} E[\Phi(\mathbf{c}, \mathbf{D})]$, and so of $\Omega(\mathbf{c})$, cannot be calculated. Thus, we estimate $\nabla_{\mathbf{c}} E[\Phi(\mathbf{c}, \mathbf{D})]$ by simulation. The second and third terms of Equation (6), on the other hand, have fixed values for a given $\mathbf{c}$, so their gradients are easily computed.

The estimation procedure begins with the solution of the second-stage problem for an initial capacity vector. At each iteration k, R realizations of the demand vector are generated in a common probability space, where we denote the rth realization of the demand vector at iteration k by $\mathbf{d}_r^k$. For each realization r, we solve stage 2, and calculate the shadow price of the capacity constraint j, $u_j(\mathbf{c}, \mathbf{d}_r^k)$, where $u_j(\mathbf{c}, \mathbf{d})$ denotes the partial derivative of the total profit with respect to the capacity of the jth resource for a given demand realization $\mathbf{d}$, that is, $u_j(\mathbf{c}, \mathbf{d}) =$

$\partial\Phi(\mathbf{c},\mathbf{d})/\partial c_j$. Let **u(c,d)** be the vector of shadow prices associated with the capacity constraints. We use the average shadow price of R experiments as the estimator of the gradient. Therefore, $\frac{1}{R}\sum_{r=1}^{R} u_j(\mathbf{c}, \mathbf{d}_r^k)$ represents the estimator for the partial derivative of the total profit with respect to the capacity of the j^{th} resource in iteration k. Then, the gradient estimator of $E[\Phi(\mathbf{c},\mathbf{D})]$ at iteration k is given by:

$$\tilde{\nabla}_{\mathbf{c}}^{k} E[\Phi(\mathbf{c},\mathbf{D})] = \frac{1}{R}\sum_{r=1}^{R} \mathbf{u}(\mathbf{c}, \mathbf{d}_r^k).$$

By adding this term to the gradient of the second and third terms with respect to **c**, $\tilde{\nabla}$ is estimated. Then, the new capacity vector is found by:

$$\mathbf{c} \rightarrow \mathbf{c} + b_k \frac{\tilde{\nabla}}{||\tilde{\nabla}||},$$

where b_k is the step size chosen in iteration k. The algorithm stops when the magnitude of the gradient estimator ($\tilde{\nabla}$) becomes smaller than a specified $\varepsilon > 0$, or when a specified number of iterations, say N, is exceeded.

In our numerical experiments, we set $R = 20$, $\epsilon = 10^{-6}$ and $N = 10{,}000$. In most of our numerical results, the number of iterations stopping criterion is used. This is in line with what is found in the literature (see, e.g., Karaesmen and van Ryzin 2004). Moreover, we use a fixed step size of 1 when $k \leq 5000$, and decrease it after that point by setting $b_k = 1/(k - t)$ for $k > 5000$. This rule satisfies a certain convergence criterion as discussed in the Appendix.

There are some technical issues related to the convergence of the algorithm. First, the steepest descent approach leads to a globally optimal solution only if the objective function is jointly concave in the decision variables. This is easy to verify since the c_j variables appear on the right-hand side of the constraints in Stage 2 which is a linear program. This implies that **Φ(c,d)** is jointly concave in **c** for a fixed demand vector **d**. Since the expected value operator preserves concavity, the objective function in Equation (6) is also jointly concave in **c**. The unbiasedness of the gradient estimator as well as the details of the step-size selection rule are discussed in the Appendix.

4.2. Verifying GPA Results

The verification process consists of ensuring that the method based on GPA converges to the optimal capacity levels. To test the accuracy of the method and to observe the effects of some problem parameters on the performance of the method, we benchmark our results to those from the existing literature. The criterion of the relative percentage error is used in this evaluation, where we set the relative percentage error as:

$$\%\,\text{error} = 100\times \frac{|\text{capacity by our method} - \text{capacity by existing result}|}{\text{capacity by existing result}}.$$

We first compare the results of our method with the optimal newsvendor problem results for fully specialized and fully flexible structures in a set of experiments, and then run another set of experiments for the two structures analyzed by Netessine et al. (2002). Figure 2 presents all the structures used in these comparisons.

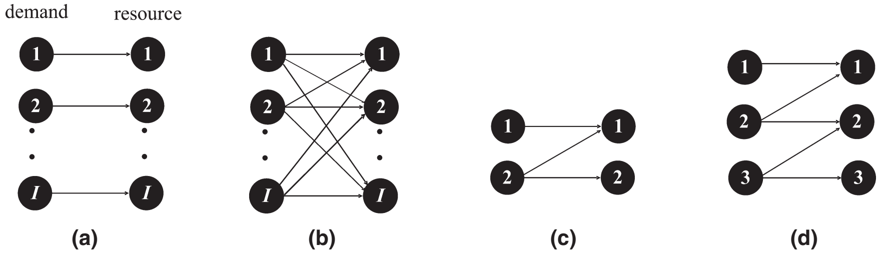
We focus on the two flexibility structures demonstrated in Figure 2a and b as bipartite graphs with nodes standing for the demand types and the resources; and arcs standing for the skills. The first structure represents a fully specialized system, where each resource has only one skill, and the second structure represents a fully flexible system, where each resource has all the skills. In the classical newsvendor problem, the optimal capacity of a resource is given by the formula $G_D^{-1}[(p - v)/p]$, where G_D^{-1} is the inverse of the cumulative demand distribution, G_D, v is the unit cost, and p is the unit price. To find the optimal newsvendor solution to the fully specialized structure, each resource-demand pair is treated independently. In the fully flexible case, the resource capacities are aggregated and the whole system is treated like a single resource-demand pair.

For simplicity, the unit specialized capacity cost, s, of the resources is assumed to be identical. Also, the cost of unit flexible capacity for any additional skill is assumed to be the same for each resource, so it is set to f. Hence, $v = s$ in fully specialized systems, and $v = s + f\times(I - 1)$ in fully flexible systems, where I is the total number of the demand types. We also note that $p = p_i$ for fully specialized resource i, and $p = (\sum_i E[D_i]p_i)/(\sum_i E[D_i])$ in fully flexible systems.

For each combination of the system parameters given in Table 1, four different types of problems are solved; fully flexible with 2 and 3 resources, and fully specialized with 2 and 3 resources. The number of demand types is taken to be equal to the number of resources in each case. The demand scenarios are created using a truncated normal distribution, $N(\lambda,\sigma)$, to ensure positive demand values. We set $\lambda = 50$ in all experiments, whereas σ has two levels: $\sigma = 5$ and $\sigma = 10$.

As a result, each of the fully flexible structures is evaluated in a total of 2 × 3 × 3 × 3 × 2 = 108 experiments, and each of the fully specialized is evaluated in 2 × 3 × 3 × 3 = 54 experiments. For the

Figure 2 (a) Fully Specialized Structure (b) Fully Flexible Structure (c) Structure 1 in Netessine et al. (2002) (d) Structure 2 in Netessine et al. (2002)



specialized systems, each capacity value is evaluated as a separate instance, such that there are 2 × 54 instances for the two-class specialized and 3 × 54 instances for the three-class specialized systems. Table 2 presents the summary of percentage errors for capacity: the second column shows the average relative error in capacity, while the third column shows the proportion of examples where the error in capacity values was <2%.

Next, we implement our method to find optimal capacity levels for the resources in the structures of Netessine et al. (2002) (see Figure 2) which we label as Netessine Structure 1 and Netessine Structure 2. We slightly modified the objective function of our model to incorporate the different pricing scheme of Netessine et al. (2002). The demands of all types in both systems arrive according to truncated normal distributions. The numerical results from our method are compared to the optimal capacity levels found for these structures by the algorithm proposed in Netessine et al. (2002). We first consider 16 instances from Table 1 (p = 30 or 50; s = 15 or 5; f = 5 or 2; initial capacity = 50) and compare capacity values under our method and the algorithm from Netessine et al. (2002) for Structure 1 and Structure 2. The demand scenarios are created using a truncated normal distribution, $N(\lambda, \sigma)$, with $\lambda = 50$ and $\sigma = 5$ or $\sigma = 10$. The errors in capacity are tabulated in Table 3. Some capacity pools exhibit higher errors compared to the specialized and fully flexible structure examples; nevertheless, simulations of profit under the capacity vectors determined by the GPA method and the algorithm from Netessine et al. (2002) reveal that their profits are close. In particular, there is only one case where the profit difference exceeds 5% and on average this difference is <2%. Finally, we consider one of the numerical examples from Netessine et al. (2002) and provide detailed results. For this example, Table 4 shows that the capacity levels found by our algorithm are close to those found to be optimal in Netessine et al. (2002). A simulation under both capacity vectors shows that profits under these capacity vectors are practically identical, thus demonstrating the flatness of the profit function around the optimal capacity values.

The above experiments provide evidence that solving the two-stage stochastic optimization problem via the GPA combined with stochastic approximation yields reliable solutions in a range of different flexibility designs.

Table 1 System Parameters

Initial capacity ($\mathbf{c}_0$)	Price (p)	Specialized capacity cost (s)	Additional skill cost (f)
100	50	15	5
50	40	10	2
0	30	5	

Table 2 Summary of Errors for Capacity Values

Problem	Average relative error	Less than 2% difference
3-3 flexible	0.67%	93%
3-3 specialized	0.75%	89%
2-2 flexible	0.84%	88%
2-2 specialized	1.0%	88%

5. Performance of the Flexibility Structures

This section compares the profits of the five flexibility structures shown in Figure 1 (two-skill complete chain, 2C; nested, N; Overflow, O; Adjacent level-12, A12; Adjacent level-23, A23) in different environments. We focus on the setting with three skills and explore the role of symmetry in demand between different skills, and variability in the demand rates, as well as differences in agent/server costs. To keep the distinction between structures, our implementation

Table 3 Summary of Errors for Capacity Values

Problem	Capacity 1	Capacity 2	Capacity 3
Netessine Structure 1	1.12%	3.84%	–
Netessine Structure 2	3.95%	8.37%	12.47%

Table 4 Comparison of the Results for Problems in Netessine et al. (2002)

Problem	Method	Pool size	Error in capacity	Profits
1	Netessine Structure 1	(138, 168)	–	3238
	GPA	(135, 169)	(2.25%, 0.54%)	3238
2	Netessine Structure 2	(127, 145, 165)	–	3105
	GPA	(124, 146, 166)	(2.12%, 0.96%, 0.73%)	3106

ensured that structures with specialist pools were unable to set the capacity of these pools to zero (the lower bound was set to one), thereby eliminating the possibility of two structures (e.g., 2 chain and A12) to become identical.

5.1. Experimental Design and Evaluation

All demand rates in the numerical experiments are assumed to have a truncated normal distribution. The examples are constructed such that the total mean demand is held constant at 120, while its distribution to the three skills is varied to ensure examples with a range of different levels of symmetry–asymmetry in mean demand rates. Without loss of generality, assume that the call types are ordered such that skill C is learned first and skill A is the last skill learned if agents follow a career path as implied by a nested structure. Another way of interpreting this ordering is to consider type A calls to be the most complex when call types are ordered by complexity. The demand rates in the experiments have one of the following orderings: (D1) $\lambda_A < \lambda_B < \lambda_C$; (D2) $\lambda_A > \lambda_B > \lambda_C$; (D3) $\lambda_A = \lambda_B = \lambda_C$. We will equivalently label D1 as the unbalanced-Pareto ordering, D2 as the unbalanced ordering, and D3 as the balanced ordering. The unbalanced-Pareto ordering represents a setting where the most complex calls, or the calls that only the most experienced agents can answer in a nested structure, have the lowest volume of calls. This type of inverse relationship between call complexity and call volume is a common situation in practice. The unbalanced ordering is the reverse of this case. The balanced ordering represents the setting analyzed in most studies in the flexibility literature. There are six asymmetric average demand vectors (see Table 5) and one symmetric (40,40,40). To analyze the role of variability in the demand rates, two different coefficient of variation (CV) values are considered. CV = 1 represents the high variability case, while CV = 0.3 represents the low variability cases. With average demand rates that vary between 10 and 90, we see that $1/\sqrt{\epsilon}$ varies between 0.31 and 0.10, thus ensuring that for the most part $CV > 1/\sqrt{\epsilon}$ and we are in an uncertainty dominated regime.

We set p to 100, while letting (s,f) take different values as specified in Table 6, which leads to a total of eight different cost configurations. These configurations will enable evaluation of the effect of decreasing s/p, specialist resource capacity costs relative to the revenue (going from the first column to the last), and the effect of s/f flexibility costs relative to the specialist resource capacity costs. Note that the total cost of a flexible server with two skills (s + f) and flexible servers with three skills (s + 2f) are decreasing as we go from configuration 1 to 8. Thus, there are 6 (asymmetric demand vectors) × 2 (CV values) × 8 (cost configurations) = 96 instances for D1 and D2 and 1 (symmetric demand vector) × 2 (CV values) × 8 (cost configurations) = 16 instances for D3.

The optimal capacity allocations for each flexibility structure are computed with an iteration limit of 10,000. We set the average demand as the initial capacity for specialist pools, and set capacity equal to 5 in all additional pools. For the A12 structure, we also consider the optimal capacity for the corresponding 2C structure as an initial capacity point along with capacity equal to 5 in the additional pools. To compare the overall system performances, the experiments are then repeated 10,000 times when optimal capacity levels are implemented in each problem. The best profit is reported whenever there are multiple initial capacity points. We compute the 95% confidence intervals on the difference of the mean profit values for each pair of structures. Whenever these confidence intervals include 0, it is not possible to conclude that one structure performs statistically better than the other, even though the average profits may be different.

5.2. Comparison of Flexibility Structures

We first focus on results for D1 (unbalanced-Pareto) and D3 (balanced). D2 only matters for the nested structure and we report on this difference at the end of the comparisons. To compare the performance of the flexibility structures, we first need to define

Table 5 Mean Demand Rates, λ_A, λ_B, λ_C

a	$\lambda_A < \lambda_B < \lambda_C$						$\lambda_A > \lambda_B > \lambda_C$						$\lambda_A = \lambda_B = \lambda_C$
Set	1	2	3	4	5	6	7	8	9	10	11	12	13
A	10	10	10	10	20	30	90	80	70	60	60	50	40
B	20	30	40	50	40	40	20	30	40	50	40	40	40
C	90	80	70	60	60	50	10	10	10	10	20	30	40

Table 6 Cost Configurations

	1	2	3	4	5	6	7	8
s	60	60	30	30	10	10	5	5
f	6	2	6	2	3	1	3	1

certain performance measures: Let Ω^S denote the profit of structure S, where S $\in$ {2C, N, O, A23, A12}. We report the number of instances where structure S_2 has a significantly higher profit than structure S_1. We define the relative profit of structure S_1 with respect to structure S_2 by the ratio $\Omega^{S_1}/\Omega^{S_2}$, and denote it by $RP(S_1|S_2)$. We report the average relative profit ($\bar{R}P(S_1|S_2)$) of a structure S_1 with respect to structure S_2 over all the instances where structure S_1 has a significantly lower profit than S_2; accordingly $0 \leq \bar{R}P(S_1|S_2) \leq 1$. Now let $\Omega^* = \max\{\Omega^S | S \in \{2C, N, O, A23, A12\}\}$. Then, the relative profit of a structure S is defined as the ratio Ω^S/Ω^*, denoted by $RP(S)$, which always satisfies $0 \leq RP(S) \leq 1$. Finally, we let $\bar{R}P(S)$ be the average of $RP(S)$ over all instances (considering both significant and not significant cases). Note that the comparisons in $\bar{R}P(S)$ are always with respect to the best-performing structure.

Table 7 shows pairwise comparison results for each structure under D1 and D3. In this Table, as well as in the other tables that follow, the first number reported in each cell represents the number of instances where the structure in the row has a significantly higher profit than the structure in the column. The second number is the average of the relative profits of the structure in the column (S_1) with respect to the structure in the row (S_2) among the significantly different cases ($\bar{R}P(S_1|S_2)$). The percentages in the last row represent $\bar{R}P(S)$, the average relative profit of the structure in the column with respect to the best-performing structure over all instances. For example, consider the first column: Structure N performs significantly better than structure 2C in two instances, where the average relative profit of structure 2C with respect to structure N is 97% over these two instances, that is, $\bar{R}P(2C|N) = 97\%$. On the other hand, the overall performance of structure 2C is given by the average relative profit as $\bar{R}P(2C) = 97\%$.

We can rank the structures with respect to the number of instances when each structure is significantly better than another and the average relative profits in this table. Then A12 is the best structure, closely followed by O, while 2C is the third, N the fourth and A23 the last. In fact, structure A12 performs as the best structure in almost all cases, which is reflected by 100% average relative profit. However, the average relative profits of all structures are quite high with a minimum of 95%. Hence, while the number of significantly different instances shows a clear ordering between structures, in terms of relative profit, the differences are low.

We then focus on the role of asymmetry in demand rates. Tables 8 and 9 show the results for 16 instances each for the most unbalanced demand vector (10, 20, 90) and the balanced demand vector, respectively. From these Tables, we observe that the 2C and N structures perform better for symmetric (balanced) demand, while O performs relatively better when demand is asymmetric. A12 consistently performs as the best both under symmetric and asymmetric demand; however, relative average profits with O under asymmetric demands and with 2C, N, and O under symmetric demands are very small (1–2%).

The role of balance in the demand rates is further explored in Table 10. In this analysis, the best structure under each of the six unbalanced demand vectors is compared to the best structure under symmetric demand. We let $\Omega^{(\ell)}$ be the average profit of the best performing structure when the demand rates are given by Set (ℓ), where $\ell = \{1, 2, 3, 4, 5, 6\}$, and $\Omega^{(13)}$ is the profit of the balanced demand case (see Table 5 for the labels). Since all demand rates sum up to 120, the differences can be attributed to the way the demands are distributed between classes. The third and fourth columns of Table 10 report the average and the minimum of $\Omega^{(\ell)}/\Omega^{(13)}$ over all instances. The differences between the average profits are not always statistically significant, so we statistically compare the average performances of unbalanced structures ($\Omega^{(\ell)}$) with that of the balanced structure ($\Omega^{(13)}$) using a significance level of 5%. In 48 of 96 instances, balanced demand structures perform significantly better, whereas there are no instances in which they perform

Table 7 Performance of all 112 Instances under Demand Patterns D1 and D3

S	2C		N		O		A23		A12	
2C	–		21	98%	19	92%	31	95%	0	–
N	2	97%	–		16	92%	18	93%	3	96%
O	66	97%	71	96%	–		72	96%	0	–
A23	0	–	6	99%	4	94%	–		0	–
A12	71	96%	81	95%	45	95%	95	95%	–	
$\bar{R}P(S)$	97%		96%		98%		95%		100%	

Table 8 Performance of 16 Instances under Unbalanced Demand Vector (10,20,90)

S	2C		N		O		A23		A12	
2C	–		2	95%	0	–	2	94%	0	–
N	0	–	–		0	–	0	–	0	–
O	11	95%	11	95%	–		12	95%	0	–
A23	0	–	0	–	0	–	–		0	–
A12	12	95%	13	94%	3	95%	14	94%	–	
$\bar{R}P(S)$	96%		95%		99%		95%		100%	

Table 9 Performance of 16 Instances under Balanced Demand Vector (40,40,40)

S	2C		N		O		A23		A12	
2C	–		3	98%	4	94%	2	96%	0	–
N	1	96%	–		3	91%	2	94%	1	96%
O	9	97%	9	97%	–		9	97%	0	–
A23	0	–	2	99%	2	94%	–		0	–
A12	9	97%	11	96%	8	96%	12	96%	–	
$\overline{RP}(S)$	98%		97%		97%		97%		100%	

Table 10 Performance Comparison by Demand Pattern

Set	Demand	Overall average (%)	Worst (%)	Significant differences	Average over significant cases (%)
(1)	10-20-90	94	72	13 out of 16	93
(2)	10-30-80	96	80	10 out of 16	93
(3)	10-40-70	97	86	10 out of 16	95
(4)	10-50-60	97	90	8 out of 16	95
(5)	20-40-60	99	93	7 out of 16	97
(6)	30-40-50	100	98	0 out of 16	–

significantly worse. The fifth column of Table 10 shows the distribution of the statistically significant instances over different demand structures. Finally, we report the average performance of the unbalanced structures over only the significant instances (last column of Table 10). We observe that more balance (lower rows of Table 10) leads to higher profits and better performance consistently under all these metrics.

In addition to the demand pattern, we observe that variability of demand rates, as measured by the CV, has an effect on the comparisons between the structures. A comparison of Tables 11 and 12 demonstrates that under low variability, A12 and O clearly outperform the others. While A12 is the better of the two, the difference between these two structures is negligible under low variability. As variability is increased, dominance of O and A12 over 2C and N becomes less distinct.

Tables 13 and 14 present the instances with respect to low and high s/p ratios, respectively. Looking closer to Tables 11–14, we see that N outperforms O when CV = 1 is coupled with high s/p and s/f values, in other words when specialist pools become relatively more expensive. Similarly, 2C outperforms O when CV = 1 is coupled with higher relative specialist costs as well as more balanced demand patterns. A23 is the worst structure and outperforms O and N only in several instances. It never outperforms A12 or 2C. The role specialist costs play in the performance of 2C and N is further demonstrated by Tables 13 and 14.

Table 11 Performance of 56 Instances with CV = 0.3

S	2C		N		O		A23		A12	
2C	–		12	99%	0	–	7	99%	0	–
N	0	–	–		0	–	2	99%	0	–
O	55	97%	56	96%	–		54	96%	0	–
A23	0	–	6	99%	0	–	–		0	–
A12	56	96%	56	95%	21	98%	56	96%	–	
$\overline{RP}(S)$	96%		95%		99%		96%		100%	

Table 12 Performance of 56 Instances with CV = 1

S	2C		N		O		A23		A12	
2C	–		9	96%	19	92%	24	93%	0	–
N	2	97%	–		16	92%	16	93%	3	96%
O	11	97%	15	96%	–		18	96%	0	–
A23	0	–	0	–	4	94%	–		0	–
A12	15	96%	25	95%	24	93%	39	94%	–	
$\overline{RP}(S)$	98%		97%		96%		95%		100%	

Table 13 Performance of 56 Instances with Low *s*/*p* Ratios

S	2C		N		O		A23		A12	
2C	–		0	–	0	–	2	98%	0	–
N	0	–	–		0	–	0	–	0	–
O	36	98%	40	98%	–		39	97%	0	–
A23	0	–	0	–	0	–	–		0	–
A12	39	98%	42	97%	1	–	42	97%	–	
$\overline{RP}(S)$	98%		98%		100%		98%		100%	

Table 14 Performance of 56 Instances with High *s*/*p* Ratios

S	2C		N		O		A23		A12	
2C	–		21	98%	19	92%	29	94%	0	–
N	2	97%	–		16	92%	18	93%	3	96%
O	30	95%	31	95%	–		33	95%	0	–
A23	0	–	6	99%	4	94%	–		0	–
A12	32	94%	39	93%	44	95%	53	93%	–	
$\overline{RP}(S)$	96%		95%		96%		93%		100%	

The D2 demand pattern only makes a difference for the nested structure. We summarize our observations for these 96 instances without the details for brevity. By construction, we expect the nested structure to perform better under D1. This is because we have a higher volume of calls for the specialist pool and a lower volume of calls for pools with higher skill numbers. This pattern is reversed in D2. When we compare the performance of N under the two demand patterns, we see that indeed N is better than the other structures in five fewer instances under D2 and is significantly worse than other structures in eight more instances compared to the D1 setting. Nevertheless, the average profit performance differences between the D1 and D2 settings are very small (<1%).

5.3. Summary of Results

Earlier literature on flexibility design has focused on the value of flexibility for a fixed capacity level. In our analysis, we compared the value of flexibility when capacity is optimized under each flexibility structure. The most interesting observation from our analysis is that when capacity is properly optimized, the profit difference induced by different flexibility structures is small. This observation is particularly applicable in the case of services where adjusting capacity is much easier than adjusting capacity in a manufacturing setting. Given that different flexibility structures have different human resource implications (as elaborated on in the following section), and may not be easily changed, the result on the importance of capacity optimization is significant. It is possible to view the five structures considered herein as combinations of basic structures: Overflow = generalist pool + specialists; 2 chain = 2-skill pools (no specialists); Nested = 1,2,3 skill pools (just one specialist pool); A12 = 2 chain + specialists; A23 = generalist pool + 2 chain (no specialists).

We then note that in the overall performance ranking A12 > O > 2C > N > A23, A12, and O are the two structures which have specialist pools. Allowing for specialist pools enables control of capacity costs while the two-skill pools and the generalist pool in A12 and O, respectively, allow to provide the little flexibility that is needed. On the other hand, N and A23 are two structures with three skill pools. These pools are expensive and unlike the O structure, the lack of specialist pools does not enable these structures to control costs. As the cost of flexible capacity increases and the need for it is increased through a higher CV, the performance of the O structure deteriorates relative to A12 and 2C due to the presence of flexibility through an expensive three-skill pool as opposed to two-skill pools in the former two structures. The 2C structure, which is the one recommended by the flexibility design literature, is indeed the best when specialist pools are also allowed as in A12. Analysis of the results for A12 and A23 shows that the optimal structure for symmetric systems, determined as systems that only invest in adjacent levels of flexibility (Bassamboo et al. 2010b), also appears to be the best when systems are not symmetric (as demonstrated by A12 in our numerical analysis). However, at the same time, A23 turned out to be the worst structure despite its focus on adjacent levels of flexibility. This contrast underlines the importance of choosing the right adjacent level structure. Combining our observation regarding specialist pools with the adjacent level result seems to suggest that A12, which is basically a two-chain plus specialists, would give excellent (if not the best) results in most settings.

Our recommendations are made under the assumption that flexibility in the form of additional skills makes a server more expensive. While we have assumed linear costs, one could also have concave costs. Furthermore, one can envisage settings where the costly flexibility assumption may not be true, thus removing the additional cost of flexible servers we have assumed. In such a setting, one would compare the flexibility structures in terms of benefits and possibly other criteria, which may change the suggestions made herein. Among other criteria, one can envisage the effect of different flexibility structures on call routing and agent occupancy or call routing and fairness. The study of such criteria would require a different modeling approach that enables an analysis of call routing in continuous time.

5.4. Managerial Insights on Flexibility Design and a Case Study

For a new system that is being established, flexibility design starts with the strategic issue of skills definition. Skills are defined by combining and grouping tasks to form a skill. Our results suggest that the best strategy is to define the skills such that the resulting demand rates would be balanced. This is in line with the intuition that the benefits of flexibility are highest in settings with high symmetry, and is supported by our numerical analysis.

Certain settings may not allow for free regrouping of tasks in skills to obtain the desired balanced structure for demand rates. An extreme example for this inflexibility at the skill definition level is the one offered by multilanguage call centers, where each language has to constitute a different skill. Systems which cannot change their skill definitions need to choose server skill sets given skill definitions. Allowing for specialist pools to control capacity costs and combining these with a two-chain or a generalist pool will result in consistently high profit performance in such systems.

Next, consider a system where skill sets have already been established. Suppose we have a call center that offers its employees a career path based on a progressive expansion of their skills. Such a system is organized according to a nested structure, which also becomes its major constraint. Our results suggest that building career paths differently, possibly a progression from specialist pools to multiskill pools and then onward to different managerial positions like supervisor or team leader is more appropriate than career paths that just build on the number of skills, particularly when additional skill costs are significant. However when nested structures that build on the number of skills cannot be avoided, managers need to pay attention to the demand arrival rates. Indeed, unless the demand rates are ordered as $\lambda_A < \lambda_B < \lambda_C$, the nested structure cannot provide a career path under optimal capacity. This is because when demand rates are ordered as in D2, optimal capacity is set such that there are few servers in the specialized pools, and the pool with three skills has the highest capacity. This implies that the idea of starting many servers as specialists and having some progress to two skills and some of those progress to three skills does not work in this instance. When D2 prevails, the nested structure does not provide the desired career path option.

Frequently, service organizations rely on generalist pools that can handle all service requests. Our results demonstrate that adding specialist pools and optimizing capacity can lead to superior profit performance for such systems, as demonstrated by the overflow structure. An advantage of the overflow structure relative to A12 is that as the number of skills increases, the number of agent pools to manage increases significantly with an A12 type structure, while for the O structure it remains as the number of skills plus one more pool. While the overflow structure comes out as a robust flexibility structure, as the number of skills increases, job scope for the flexible agents might become excessive. In general, job scope has been shown to have positive effects on performance (Hackman and Oldham 1976, Ilgen and Hollenbeck 1991). Xie and Johns (1995) 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. If this is the case, the two-chain structure with additional specialist pools (A12) may be preferred.

To illustrate the application of our analysis, we focus on a case study from a retail banking call center. This call center has predefined skills and our focus is on flexibility structure design under capacity optimization. The existing situation is described next: The call center has seven defined skills (1, …, 7) with agents grouped in five pools according to their skill sets. Current skill sets (pools) are as follows: Pool S = Skill 5; Pool AS = Skills 2,4,5; Pool A = Skills 1,3,2,4,5; Pool B = 1,3,5; Pool Q = 5,6,7; From S to AS to A, we see a nested structure with 2 and 4 grouped (24) and 1 and 3 grouped (13) together as a single skill. From S to B to A, we see another nested structure, again with (13) and (24) bundled together. Furthermore the presence of two nested structures suggests that learning (13) first or (24) first does not seem to matter. Pool Q is separate from the other pools. From S to Q, we can see another path for agents. The existing flexibility structure is shown in Figure 3.

Recommending an adjacent level flexibility structure as in A12, our results would suggest the following pools. In line with current practice, we assume that (13) and (24) are grouped and treated as a single skill. Specialist Pool 1 = 5 (current S); Specialist Pool 2 = (13) (new); Specialist Pool 3 = (24) (new); two-skill pool 1 = (13),5 (current B); two-skill pool 2 = (24),5 (current AS); two-skill pool 3 = (13),(24) (remove 5 from current A, labeled as A-). So we would just add two specialist pools and remove skill 5 from A to reduce the number of skills in this pool from 3 to 2. This structure would address career path concerns to some extent by offering several alternatives: Paths 1 and 2: 5 to 5,(13) or 5 to 5,(24); Paths 3;4;5;6: (13) to (13),(24); (24) to (24),(13); 13 to (13),5; (24) to (24),5.

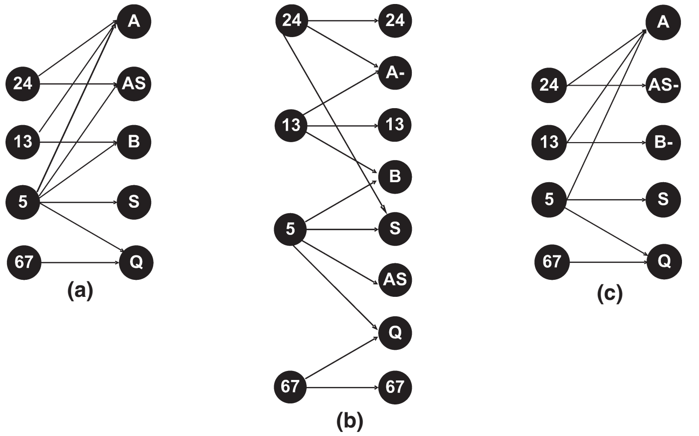
Recommending an overflow structure as in O, our results would suggest the following pools. Specialist Pool 1: 5 (current S), Specialist Pool 2: (13) (remove 5 from current B, labeled as B-), Specialist Pool 3: (24) (remove 5 from current AS, labeled as AS-), three-skill pool 5,(12),(34) current pool A. One advantage of this structure would be the fewer agent pools to manage.

The recommended structures are also shown in Figure 3. As this example demonstrates, in practice, we do not expect to find one of these structures in isolation but frequently find several of them combined. Thus, the banking example demonstrates a case where we have nested structures combined with other pools like pool Q. Some of these choices may be driven by concerns not addressed in our analysis. Nevertheless, insights from our analysis can guide flexibility structure redesign.

To further test the value of our recommendations, we perform a numerical analysis for this banking call center. We use a data set with the number of call arrivals and average call handle times in 30-minute intervals in the month of April, 2008 for each defined skill. An analysis of arrival rates indicates that weekdays are different from weekends and that the period between 13:00 and 15:30 represents the peak period of each day. There does not seem to be a consistently peak day among the weekdays for all groups. We thus focus our analysis of demand on weekdays during the identified peak period to avoid any time and day dependent effects on arrival rates. Average service times are found by taking the average call handle times for each skill. Whenever skills are combined, we assume that they are independent and the arrival rate and offered load (ε) of the combined skills can be obtained by summing those for each skill.

Table 15 tabulates the mean number of calls for 30 minutes, the standard deviation of the mean number of calls for 30 minutes, the corresponding empirical coefficient of variation CV_{emp}, the coefficient of variation if arrivals were assumed to be Poisson $CV_{Poisson}$, and the value of $1/\sqrt{\epsilon}$ for each of the four demand types (2 + 4, 1 + 3, 5, 6 + 7). We observe that all arrivals are overdispersed relative to a Poisson process. CV_{emp} is larger than the value of $1/\sqrt{\epsilon}$ for

Figure 3 (a) Existing Structure (b) Recommended A12 Structure (c) Recommended Overflow Structure



demand types 2 + 4, 1 + 3, 5, and is slightly less for demand type 6 + 7, thus demonstrating that the call center is mostly operating in an uncertainty dominated regime.

We use the GPA-based method to determine optimal capacity levels for the existing and recommended structures shown in Figure 3, under the given demand data. Since we do not have cost-related data, we use p = 100 and cost configurations 1 and 8 from Table 6. Profits are simulated for implemented capacities under 500,000 repetitions in these examples, and are tabulated in Table 16. For both cost configurations, we find that the two recommended structures are significantly better than the existing structure in terms of profit performance at a 95% confidence level. Furthermore, the overflow and A12-based recommended structures lead to profit performance that is very similar. In the first instance with higher costs s and f, the profit under the two structures is statistically indistinguishable. In the second instance, the A12-based structure results in slightly better performance (4% better). While the profit differences between the proposed structures and the existing one are low for the case with a higher s/p ratio, the proposed structures outperform the existing structure by 25−29% for the case when this ratio is low.

Table 16 Average Profits (Standard Error) for the Existing and Proposed Structures of the Banking Call Center

Costs	Existing	Overflow	A12
p = 100; s = 30; f = 6	19,893 (50)	20,311 (50)	20,332 (50)
p = 100; s = 5; f = 1	49,420 (24)	66,382 (27)	69,434 (27)

6. Concluding Remarks

We analyzed the capacity decision of multiresource service or production systems considering the flexibility structure and the cost of capacity. A solution method that determines the capacities of the resources under uncertain demand is proposed. The modeling approach we select allows us to capture random demand rates found in call centers.

Table 15 Demand Data from the Banking Call Center

Demand type	Mean	SD	CV_{emp} (%)	$CV_{Poisson}$ (%)	$1/\sqrt{\epsilon}$ (%)
2 + 4	176.54	73.10	41.41	7.53	19.63
1 + 3	1449.16	356.84	24.62	2.63	7.28
5	182.19	79.54	43.66	7.41	20.85
6 + 7	386.37	62.77	16.25	5.09	19.16

The numerical comparison of the five flexibility structures, found in practice or recommended by the literature, illustrates the importance of optimizing capacity given a flexibility structure. We confirm earlier statements from the literature that for services a little flexibility is important and valuable, and additionally show that capacity optimization coupled with some no-flexibility (specialist) resources is even better. Defining skills to minimize demand asymmetry is worth exploring where feasible.

The result that systems that can combine a high capacity in specialized resources with a lower level of flexible capacity (as found in the overflow structure) provide superior profit performance, is consistent with the 80-20 rule observed in Chevalier et al. (2004), as well as with the related results in Chevalier and Van den Schrieck (2008) and Bassamboo et al. (2012). Analyzing queueing systems in heavy traffic, with their capacity optimized, under the assumption of balanced demand rates, Bassamboo et al. (2012) show the optimality of adjacent-level flexibility structures. The fact that different modeling approaches, like the loss systems analyzed via approximations in Chevalier and Van den Schrieck (2008), or the queueing systems under heavy traffic in Bassamboo et al. (2012) lead to consistent results with our observations based on newsvendor-type models lends further support to the robustness of these with respect to modeling specifics.

Acknowledgments

This research was partially supported by The Scientific & Technological Research Council of Turkey, TÜBİTAK and the Turkish Academy of Sciences, TÜBA-GEBİP program.

Appendix

A. Unbiasedness and Convergence

Glasserman (1991) draws attention to the importance of two theoretical issues concerning the validity of the GPA method, unbiasedness and convergence. We investigate these issues in detail below.

First, consider the unbiasedness of the estimator. Let $X(\theta)$ be a random function of parameter θ. If $\nabla_\theta X(\theta)$ is an unbiased estimator of $\nabla_\theta E[X(\theta)]$, the following is true by definition:

$$E[\nabla_\theta X(\theta)] = \nabla_\theta E[X(\theta)]. \tag{A1}$$

In our setting, this condition can be written as:

$$E\left[\frac{1}{R}\sum_{r=1}^{R}\mathbf{u}(\mathbf{c},\mathbf{D}_r^k)\right] = \nabla_{\mathbf{c}} E[\Phi(\mathbf{c},\mathbf{D})].$$

It is known that (Glasserman 1994) if $X(\theta)$ is almost surely (a.s.) differentiable at θ, and $X(\theta)$ satisfies the Lipschitz condition, then the estimator is unbiased. We first show that $\Phi(\mathbf{c},\mathbf{d})$ is a.s. differentiable with respect to $\mathbf{c}$ for any given realization $\mathbf{d}$. $\Phi(\mathbf{c},\mathbf{d})$ is piecewise linear and concave with respect to $\mathbf{c}$, since it is the objective function of a linear maximization problem and $\mathbf{c}$ is the right-hand side of the constraints. Clearly, $\Phi(\mathbf{c},\mathbf{d})$ fails to be differentiable at a finite number of points. But since the average of R paths is taken and the demand is continuous, the non-differentiable points smooth-out (see Glasserman 1994). Now, we are ready to prove the following lemma.

Lemma 1. $\frac{1}{R}\sum_{r=1}^{R}\mathbf{u}(\mathbf{c},\mathbf{d}_r^k)$ *is an unbiased estimator of* $\nabla_{\mathbf{c}} E[\Phi(\mathbf{c},\mathbf{D})]$.

Proof. We know from the above discussion that $\Phi(\mathbf{c},\mathbf{d})$ is almost surely differentiable with respect to $\mathbf{c}$ for any given realization $\mathbf{d}$. Then, we only need to show that $\Phi(\mathbf{c},\mathbf{d})$ satisfies the Lipschitz condition for each c_j and for all $\mathbf{d}$.

Let $\mathbf{c}$ be a point at which $\Phi(\mathbf{c},\mathbf{d})$ is differentiable. At point $\mathbf{c}$, the effect of a small increase in the capacity cannot be more than the profit gained by the same amount of increase in the throughput of the most expensive job in the skill set of the corresponding resource. Moreover, this effect is also bounded by the unit capacity cost of resource j.

Let ε be the amount of increase in the capacity of resource j, $\bar{p}_j$ be the highest contribution margin among all jobs that the additional capacity can process, and $f_j + t_j$ be the unit capacity cost of resource j. Then, we have the following:

$$|\Phi(\mathbf{c}+\varepsilon e_j,\mathbf{d}) - \Phi(\mathbf{c},\mathbf{d})| \le \varepsilon \min\{\bar{p}_j, f_j + t_j\}, \tag{A2}$$

where e_j is the jth unit vector. Consequently, $\Phi(\mathbf{c},\mathbf{d})$ satisfies the Lipschitz condition.

The second important theoretical issue is related to the convergence of the method closely related to the step-size selection rule. Considering the possibility that the initial capacity is too far from the optimal, we implement the method by starting with a big step size to accelerate the convergence. For the first half of the iterations, a fixed step size of 1 is used. After that point, we begin decreasing the step size following the rule $b_k = 1/(k-t)$, satisfying the conditions given below:

$$\sum_{k=1}^{\infty} b_k = \infty \quad \sum_{k=1}^{\infty} b_k^2 < +\infty, \tag{A3}$$

which were established by Robbins and Monroe (1951) to guarantee convergence.

References

Akşin, O. Z., F. Karaesmen. 2002. Designing flexibility: Characterizing the value of cross-training practices. Working paper, INSEAD.

Akşin, O. Z., F. Karaesmen. 2007. Characterizing the performance of process flexibility structures. *Oper. Res. Lett.* **35**: 477–484.

Akşin, O. Z., M. Armony, V. Mehrotra. 2007a. The modern call-center: A multi-disciplinary perspective on operations management research. *Prod. Oper. Manag.* **16**(6): 665–688.

Akşin, O. Z., F. Karaesmen, L. Örmeci. 2007b. A review of workforce cross-training in call centers from an operations management perspective. D. Nembhard, ed. *Workforce Cross Training Handbook*. CRC Press LLC. Boca Raton, FL, 211–240.

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

Andradottir, S., H. Ayhan, D. G. Down. 2013. Design principles for flexible systems. *Prod. Oper. Manag.* **22**(5): 1144–1156.

Avramidis A. N., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Manage. Sci.* **50**(7): 896–908.

Bassamboo, A., R. S. Randhawa, A. Zeevi. 2010a. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Manage. Sci.* **56**(10): 1668–1686.

Bassamboo, A., R. S. Randhawa, J. A. Van Mieghem. 2010b. Optimal flexibility configurations in newsvendor networks:

Going beyond chaining and pairing. *Manage. Sci.* **56**(8): 1285–1303.

Bassamboo, A., R. S. Randhawa, J. A. Van Mieghem. 2012. A little flexibility is all you need: On the asymptotic value of flexible capacity in parallel queuing systems. *Oper. Res.* **60**(6): 1423–1435.

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

Chevalier, P., R. A. Shumsky, N. Tabordon. 2004. Routing and staffing in large call centers with specialized and fully flexible servers. Working paper. Tuck School of Business, Hanover, NH.

Fine, C. H. and R. M. Freund. 1990. Optimal investment in product-flexible manufacturing capacity. *Manage. Sci.* **36**(4): 449–466.

Glasserman P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer, Boston.

Glasserman P. 1994. Perturbation analysis of production networks. D. Yao, ed. *Stochastic Modeling and Analysis of Manufacturing Systems*. Springer-Verlag, New York, 233–280.

Graves, S. C., B. T. Tomlin. 2003. Process flexibility in supply chains. *Manage. Sci.* **49**(7): 907–919.

Gurumurthi, S., S. Benjaafar. 2004. Modeling and analysis of flexible queueing systems. *Nav. Res. Log.* **51**: 755–782.

Hackman, J. R., G. R. Oldham. 1976. Motivation through the design of work: Test of a theory. *Org. Behav. Hum. Perform.* **16**: 250–279.

Harrison, J. M., J. A. van Mieghem. 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *Eur. J. Oper. Res.* **113**: 17–29.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manuf. Serv. Oper. Manag.* **7**(1): 20–36.

Hopp, W. J., E. Tekin, M. P. VanOyen. 2004. Benefits of skill chaining in production lines with cross-trained workers. *Manage. Sci.* **50**(1): 83–98.

Hunter, L. W. 1999. Transforming retail banking: Inclusion and segmentation in service work. P. Cappelli, ed. *Employment Practices and Business Strategy*. Oxford University Press, New York, 153–192.

Ilgen, D. R., J. R. Hollenbeck. 1991. The structure of work: Job design and roles. M. D. Dunnette, L. M. Hough, ed. *Handbook of Industrial and Organizational Psychology 2*. Consulting Psychologists Press, Palo Alto, CA, 165–207.

Inman, R. R., W. C. Jordan and D. E. Blumenfeld. 2004. Chained cross-training of assembly line workers. *Int. J. Prod. Res.* **42** (10): 1899–1910.

Iravani, S. M., K. T. Sims, M. P. Van Oyen. 2005. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Manage. Sci.* **51**(2): 151–166.

Iravani, S. M. R., B. Kolfal, M. P. Van Oyen. 2007. Call center labor cross-training: It's a small world after all. *Manage. Sci.* **53**(7): 1102–1112.

Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Manage. Sci.* **41**(4): 577–594.

Jordan, W. C., R. R. Inman, D. E., Blumenfeld. 2004. Chained cross-training of workers for robust performance. *IIE Trans.* **36**: 953–967.

Karaesmen, I., G. van Ryzin. 2004. Overbooking with substitutable inventory classes. *Oper. Res.* **52**(1): 83–104.

van Mieghem, J. A. 1998. Investment strategies for flexible resources. *Manage. Sci.* **44**(8): 1071–1078.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* **50**(2): 375–388.

Robbins, H., S. Monroe. 1951. A stochastic approximation method. *Ann. Math. Stat.* **22**(3): 400–407.

Sheikzadeh, M., S. Benjaafar, D. Gupta. 1998. Machine sharing in manufacturing systems: Total flexibility versus chaining. *Int. J. Manuf. Syst.* **10**: 351–378.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2005. Performance measures for service systems with a random arrival rate. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, J. A. Joines, eds. *Proceedings of the 2005 Winter Simulation Conference*, IEEE, Piscataway, NJ.

Talluri, K., G. van Ryzin. 1999. A randomized linear programming method for computing network bid prices. *Transport. Sci.* **33**(2): 207–216.

Van Oyen, M. P., E. G. S. Gel, W. J. Hopp. 2001. Performance opportunity for workforce agility in collaborative and noncollaborative work systems. *IIE Trans.* **33** (9): 761–777.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manuf. Serv. Oper. Manag.* **7**(4): 276–294.

Xie, J. L., G. Johns. 1995. Job scope and stress: Can job scope be too high? *Acad. Manage. J.* **18**: 1288–1309.