Flexibility Structure and Capacity Design with Human Resource Considerations

O. Zeynep Akşin [1], **Nesrin Çakan [2], Fikri Karaesmen [3], E. Lerzan Örmeci [3]**

[1] College of Administrative Sciences and Economics
Koç University
34450, Sarıyer, Istanbul, Turkey
zaksin@ku.edu.tr

[2] Accenture, Istanbul, Turkey
nesrin.cakan@accenture.com

[3] Department of Industrial Engineering
Koç University
34450, Sarıyer, Istanbul, Turkey
lormeci@ku.edu.tr, fkaraesmen@ku.edu.tr,

Flexibility Structure and Capacity Design with Human Resource Considerations

O. Zeynep Akşin[1], **Nesrin Çakan** [2], **Fikri Karaesmen** [3], **E. Lerzan Örmeci** [3]

[1] College of Administrative Sciences and Economics, Koç University, Istanbul, Turkey
zaksin@ku.edu.tr

[2] Accenture, Istanbul, Turkey
nesrin.cakan@accenture.com

[3] Department of Industrial Engineering, Koç University, Istanbul, Turkey
fkaraesmen@ku.edu.tr, lormeci@ku.edu.tr

Abstract

Most service systems consist of multi-departmental structures corresponding to multiple types of service requests, with possibly multi-skill 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 paper is to identify preferred flexibility structures when demand is random and capacity is finite. We compare a structure recommended by the flexibility literature to structures we observe in practice within call centers. In order 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.

Keywords: flexibility, call centers, multidimensional newsvendor, gradient estimation via perturbation analysis

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.

Through resource flexibility, demand is aggregated and resources are shared. Therefore, flexibility improves both the utilization and the throughput level 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 paper 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.

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 multi-skill 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.

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 nested structure as an example of what one might find in an organization with inclusive work practices, whereas an overflow structure would suggest segmented work practices.

Our research objective is to compare the two-skill chain structure recommended by the literature with the nested and overflow structures found in practice. Figure 1 shows the three flexibility structures to be analyzed in this paper for the case when there are three different types of skills (used synonymously with call types or products throughout) and the flexible agents in the overflow structure have been assumed to be fully flexible. 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?

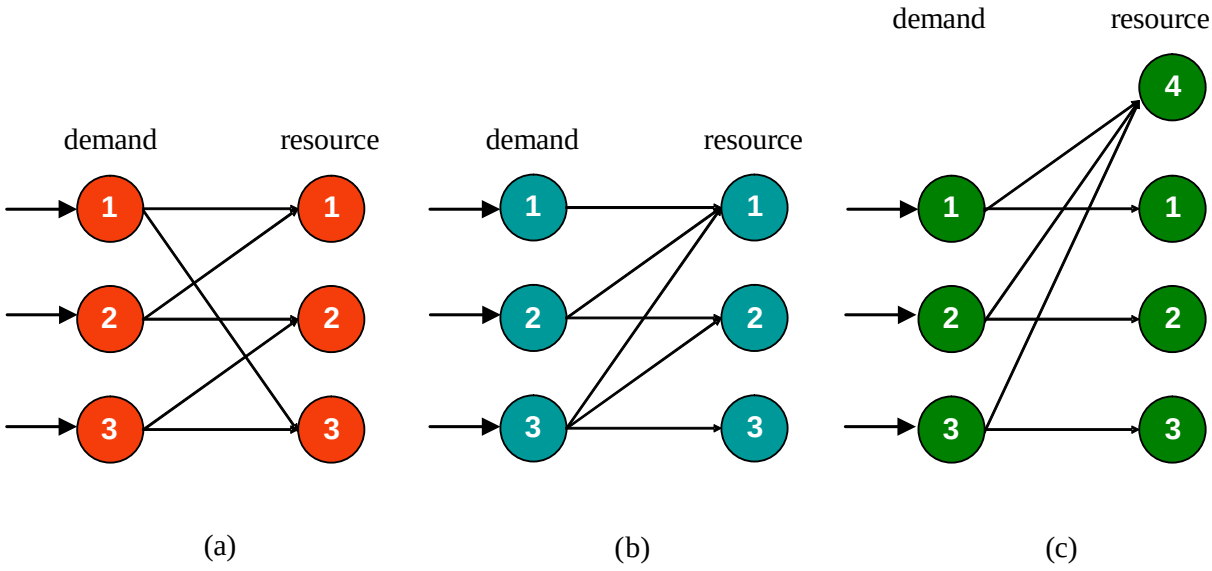

Figure 1: Different Flexibility Structures (a) Two-skill complete chain (b) Nested (c) Overflow

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 paper are organized as follows: A review of the literature in Section 2 is followed by a presentation of the model in Section 3. The discussion in Section 4 illustrates how time varying arrivals and abandonment from queues can be treated within the same modeling framework. Section 5 introduces the solution method and analyzes its theoretical background. Some benchmarks from the literature are used to numerically verify this solution method. Section 6 proceeds with a numerical study that compares the three flexibility structures under different assumptions on the number of customer types, demand, costs, and correlation structure. Section 6.1 describes the experimental design, while Section 6.2 gives a detailed account for the comparison of these structures. Section 6.3 presents our managerial insights based on this comparison. Conditions are identified under which each structure may be preferred by managers. We illustrate that capacity optimization enables the nested and overflow structures to overcome their throughput disadvantage vis-a-vis the two-skill chain structure in some settings. The paper ends with concluding remarks in Section 7.

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), Akşin and Karaesmen (2007), Iravani et al. (2005), and Iravani et al. (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 multi-stage 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, and Akşin and Karaesmen (2002), Akşin and Karaesmen (2007) and Wallace and Whitt (2005) within contact centers.

Both Jordan and Graves (1995) and Akşin and Karaesmen (2007) show the close relationship 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 capacity is fixed, capacity optimization problems have mostly focused on given, relatively simple flexibility structures (see for example Fine and Freund (1990), van Mieghem (1998), Netessine et al. (2002), Harrison and Zeevi (2005), Chevalier et al. (2004)). Two exceptions are Chevalier and Van den Schrieck (2008) and Bassamboo et al. (2008).

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 paper 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 paper, 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

We model the capacity design with flexible resources problem 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 multi-dimensional 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 types of jobs. The set of jobs that a resource can process will be referred to as the skill-set of that resource. The skill-sets of the resources can be 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 amount of capacity available in resource j is denoted by c_j and the amount of

job i processed by resource j is denoted by x_{ij}. 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 job i is denoted by d_i.

It is assumed that each job 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 $s_j + f_j \left(\sum_i k_{ij} - 1\right)$ where f_j denotes the cost of flexibility at resource j for each additional skill. Then, we can formulate the problem as follows:

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

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

$$\text{subject to}: \sum_i x_{ij} \leq c_j \quad \forall j, \quad (3)$$

$$\sum_j x_{ij} \leq d_i \quad \forall i, \quad (4)$$

$$0 \leq x_{ij} \leq M \times k_{ij} \quad \forall i, j, \quad (5)$$

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 (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:

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

The second stage maximizes the revenue of the system for any demand realization, $\mathbf{d}$, and capacity level, $\mathbf{c}$. Inequality (3) guarantees that the number of jobs handled by any resource is not more than its capacity, whereas inequality (4) prevents the number of processed i jobs

from exceeding the corresponding demand. Finally, (5) ensures that the jobs are assigned to the capable resources.

4. Time Varying Demands and Abandonment

In reality, call centers are queueing systems with abandonments, that experience time varying demand rates. In this section we illustrate to what extent the proposed model and analysis can accommodate these features.

Implicit in our formulation is the assumption that capacity is set once at the beginning of a time period. This suggests that time periods should be viewed as short durations, for example one day. During a day, demand arrival rates change, exhibiting a pattern with respect to time. Paralleling Harrison and Zeevi (2005), one can define a mean arrival rate vector $\mathbf{\Lambda} = (\mathbf{\Lambda}(t) : 0 \leq t \leq T)$ with $\mathbf{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_I(t))$. Note that $t = 0$ represents the beginning and $t = T$ the end of the time period in question. It is then possible to define a cumulative demand distribution at a demand level $\lambda = (\lambda_1, \ldots, \lambda_I)$ as $F(\lambda)$, representing the percentage of time that demand is less than or equal to λ as:

$$F(\lambda) = \frac{1}{T} \int_0^T P\{\mathbf{\Lambda}(t) \leq \lambda\} dt \qquad for\ \lambda \in R_+^m.$$

Defined this way, $F(\lambda)$ corresponds to $G_{\mathbf{D}}(\mathbf{d})$, the cumulative demand distribution, in our setting. This equivalence implies that as long as it is meaningful to consider capacity optimization once at the beginning of a time period as in Harrison and Zeevi (2005), it is possible to consider time varying demand rates in our model. Our modeling framework thus allows us to capture either temporal or stochastic variability in demand. In fact, even both types of variability can be captured at the expense of defining more complicated joint random variables.

A second assumption we make is the loss of all calls upon arrival, if capacity is not available. The model under this assumption ignores the queueing effects, and related phenomena such as abandonments and retrials but retains the essence of the capacity design problem. This type of model may be motivated by a fluid-approximation of the queueing system as in Harrison and Zeevi (2005). In the fluid-approximation, all calls that exceed the capacity of

the system abandon. Thus associating a revenue loss with lost demand in our setting, that is equal to the abandonment penalty in Harrison and Zeevi (2005), allows the possibility of building an equivalent objective function to theirs (see Çakan (2006)). Numerical comparisons between the fluid approximation in Harrison and Zeevi (2005) and a simulation of the original queueing system illustrate that this approximation works quite well for a capacity optimization objective. Given the possibility to construct an equivalent objective function in our setting, these comparisons also lend support to the model proposed herein for capacity optimization. As argued in Akşin et al. (2008), this type of a model is appropriate when the uncertainty in the demand rate is more significant than the short term queueing fluctuations.

5. 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.

5.1 The GPA technique

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 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 (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 $E[\Phi(\mathbf{c},\mathbf{D})]$, 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 (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 $\mathbf{d}$, i.e., $u_j(\mathbf{c}, \mathbf{d}) = \partial\Phi(\mathbf{c}, \mathbf{d})/\partial c_j$. Let $\mathbf{u}(\mathbf{c}, \mathbf{d})$ 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 $\mathbf{c}$, $\tilde{\nabla}$ is estimated.

The algorithm stops when the magnitude of the step size times the gradient estimator ($||b_k \times \tilde{\nabla}||$) becomes smaller than a specified $\epsilon > 0$, or when a specified number of iterations, say N, is exceeded. The details of the step-size selection rule are explained in Appendix B. The procedure is summarized in Figure 2.

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 $\Phi(\mathbf{c}, \mathbf{d})$ is jointly concave in $\mathbf{c}$ for a fixed demand vector $\mathbf{d}$. Since the expected value operator preserves convexity, the objective function in (6) is also jointly concave in $\mathbf{c}$. The other technical issue is related to the validity of the GPA method, which is discussed in

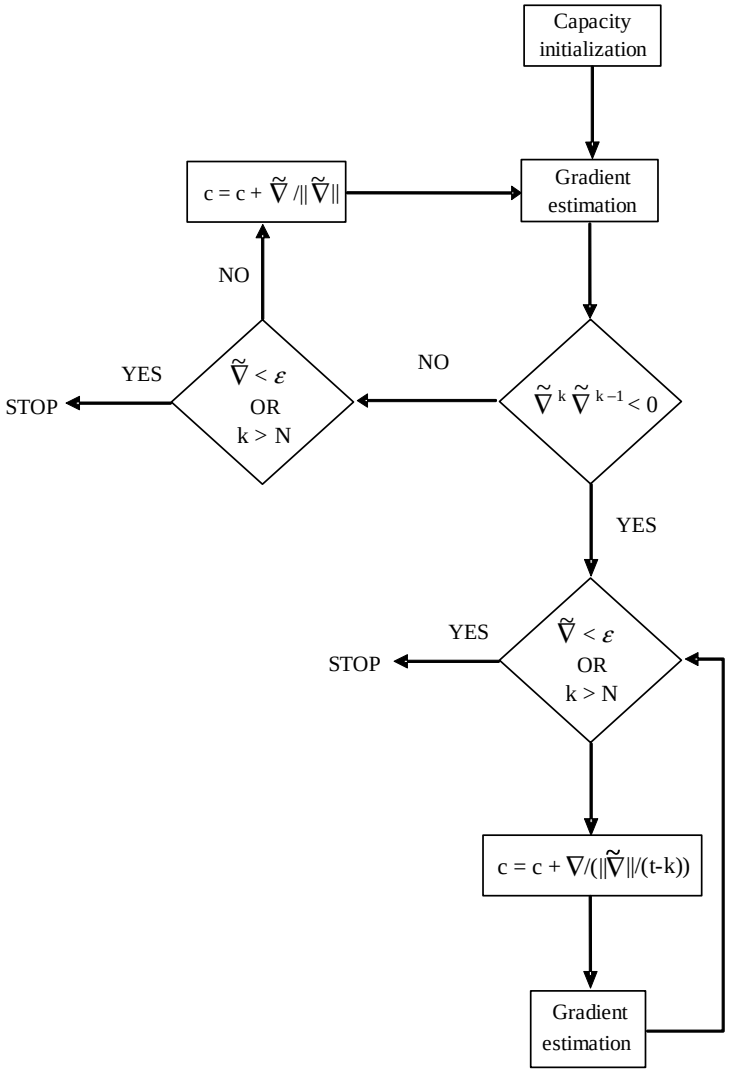

Figure 2: Summary of the procedure

Appendix B.

5.2 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 absolute percentage error is used

in this evaluation, where we set the absolute 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 in a set of experiments and then solve two benchmark problems defined by Netessine et al. (2002).

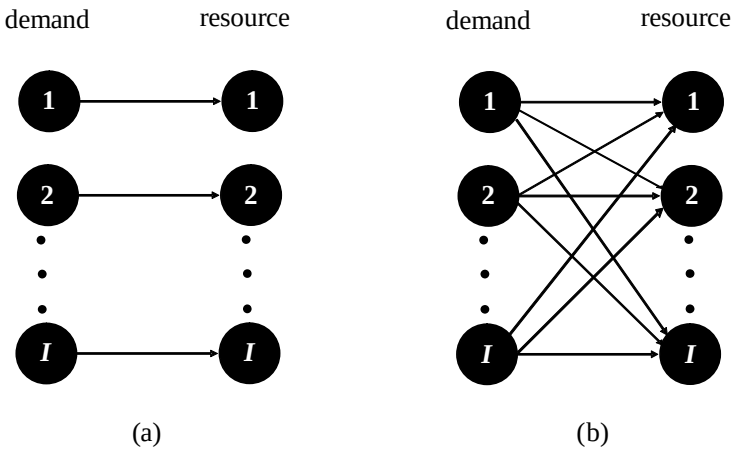

Figure 3: (a) Fully-specialized structure (b) Fully-flexible structure

We focus on the two flexibility structures demonstrated in Figure 3 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 are 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, $N(\mu, \sigma)$, to ensure positive demand values. We set $\mu = 50$ in all experiments, whereas σ has two levels: $\sigma = 5$ and $\sigma = 10$.

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 1: System parameters

As a result, each of the fully-flexible structures is evaluated in a total of $2 \times 3 \times 3 \times 3 \times 2 = 108$ experiments, and each of the fully-specialized is evaluated in $2 \times 3 \times 3 \times 3 = 54$ experiments. Table 2 presents the summary of percentage errors. The maximum percentage error is 3.48%, and all but 3 of the problem instances have less than 2% of percentage error. We conclude that our method converges to the right values.

Problem	Minimum	Maximum	Average
3-3 flexible	0.00	3.48	0.37
3-3 specialized	0.00	2.27	0.80
2-2 flexible	0.00	1.97	0.49
2-2 specialized	0.00	1.83	0.83

Table 2: Summary of percentage errors

Next, we implement our method to find optimal capacity levels for the resources in the structures of Netessine et al. (2002) (see Figure 4). We slightly modified the objective function of our model in order 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 analytically optimal capacity levels found for these structures in Netessine et al. (2002) . The results in Table 3 show that the capacity levels found by our algorithm are close to those found to be optimal in Netessine et al. (2002). In the second problem, the initial capacity choice affected the performance of the algorithm. Various experiments with different initial capacity values revealed that setting initial capacities equal to the expected demand of the corresponding demand type improves the convergence and the quality of the solution.

Problem	Method	Pool size	% error
1	Netessine	(138, 168)	–
	GPA	(137, 171)	(0.629, 1.661)
2	Netessine	(127, 145, 165)	–
	GPA	(125, 156, 166)	(1.118, 7.428, 0.745)

Table 3: Comparison of the results for problems in Netessine et al. (2002)

In Table 3, the maximum error in capacity levels is over 7%. However, further experiments revealed that the objective function is fairly flat in that region. Therefore, this difference in capacity levels translates into a very small difference in terms of the resulting objective function. The above experiments provide evidence that solving the two-stage stochastic optimization problem via the GPA method yields reliable solutions in a range of different flexibility designs.

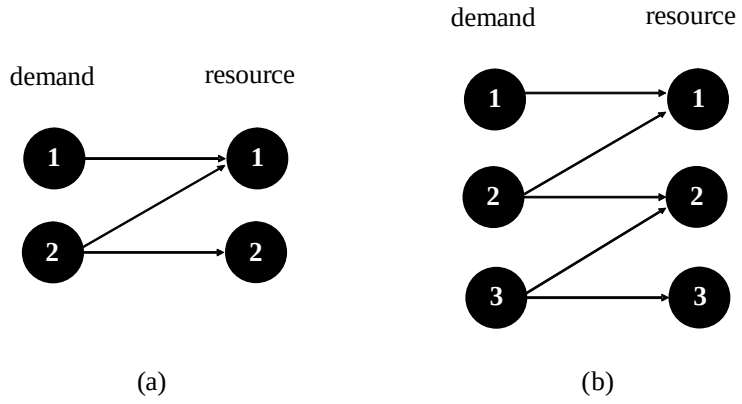

Figure 4: (a) Structure 1 in Netessine et al. (2002) (b) Structure 2 in Netessine et al. (2002)

6. Evaluating the Performance of the Flexibility Structures

This section compares the profits of the three flexibility structures (two-skill complete chain, C, nested, N, and overflow, O) in different environments. We explore the effects of the number of skills, as well as the intensity, variability and correlation of the demands for different skills on these performances.

6.1 Experimental design and evaluation

All demands in the numerical experiments are assumed to be normally distributed. We construct two sets of inputs representing systems with low and high number of skills: (I) a system with three demand types $\{A, B, C\}$, and (II) a system with eight demand types $\{A, B, \ldots, H\}$. We denote by μ_i the mean arrival rate of demand type i. In order to observe the effect of demand variability, two levels of the coefficient of variation, $C_v = 1$ or $C_v = 2$, are considered in both data sets.

Without loss of generality, assume that the call types are ordered such that skill C in the three skill case, and skill H in the eight skill case are learned first, and skill A is the last skill learned in both systems 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 arrival rates in the experiments have one of the following orderings: (O1) $\mu_A < \mu_B < \ldots$; (O2) $\mu_A > \mu_B > \ldots$; (O3) $\mu_A = \mu_B = \ldots$. We will equivalently label O1 as the unbalanced-Pareto ordering, O2 as the unbalanced ordering, and O3 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 papers in the flexibility literature.

In both data sets, we use 10 different cost configurations, where we specify different values

for (p, s, f) as shown in Table 4. These configurations will enable evaluation of the effect of high versus low resource capacity costs relative to the revenue, where configurations 1-5 correspond to high and the rest to low.

	1	2	3	4	5	6	7	8	9	10
p	50	50	50	60	50	50	50	100	100	100
s	15	15	15	20	30	5	5	5	5	5
f	5	10	15	10	5	1	4	1	4	5

Table 4: Cost configurations

In data set I, designed for a system with three demand types, there are 6 different demand volume scenarios for each ordering of the mean arrival rates, so a total of 18 scenarios (see Table 9 in the Appendix). This way $18 \times 2 \times 10 = 360$ experiments are set up for each flexibility structure, in which all the demands are independent.

We would further like to investigate the effect of correlation within the demands, if any. In data set I, we generate correlated random demands when (μ_A, μ_B, μ_C) is equal to $(10, 30, 90)$, $(30, 30, 30)$ and $(90, 30, 10)$. It is assumed that the demand for the simple calls is independent from the other demand types. Hence the influence of the correlation between the demand types A and B, ρ_{AB} is investigated. We let $\rho_{AB} \in \{-1, -0.5, 0.5, 1\}$, so that the effect of perfect positive and negative correlations, as well as any trends as the correlation changes from -1 to 1 can be observed. The number of experiments with correlations is $3 \times 2 \times 4 \times 10 = 240$ for each flexibility structure.

In data set II, designed for a system with eight demand types, there is one scenario for each ordering of the mean demand rates (see Table 10 in the Appendix). All demands are assumed to be independent, so a total of $3 \times 2 \times 10 = 60$ experiments are conducted in this set for each flexibility structure.

In each experiment, the optimal capacity allocations for each flexibility structure are computed. In order to compare the overall system performances, the experiments are repeated 1000 times when optimal capacity levels are implemented in each problem. These experiments provide 90 % confidence intervals on the real mean value of the profits as well as the real mean of the throughputs. Whenever the confidence intervals on the mean profit (throughput) value of different structures overlap, it is not possible to conclude that

one structure performs statistically better than the other, even though the average profits (throughputs) may be different.

6.2 Comparison of flexibility structures

The design of the experiments allows us to evaluate the effects of demand rate orderings, variability, correlation, different ratios of capacity cost to price, and the number of demand types. Within these factors, we identify the ordering of the demand rates as the one influencing the performances the most, while the number of demand types is a close follower. The performance rankings of the structures do not change with the correlation and the variance of demands. We observe only that high variability in the demands increases the capacity levels in all structures. The effect of the cost-price ratios, on the other hand, is significant in only one situation, as mentioned below. Consequently, we organize the presentation of the results according to the demand rate orderings.

Table 5 gives the number of the experiments in which the average profit of each structure ranks as the 1^{st}, 2^{nd} and 3^{rd} for data set I with independent demands. Each cell in the table contains two entries, where the first one counts the experiments with non-overlapping confidence intervals (these are the statistically significant differences) and the second simply counts the ranks according to the average profits. So, for instance, for the ordering $\mu_A = \mu_B = \ldots$, all confidence intervals around the average profits overlap with each other. Hence, even though the overflow structure performs the best in all 120 experiments in terms of the average profit, it is safer to conclude that the performance of this structure is not significantly better than the other two. Table 6 presents the same information as Table 5 for data set II.

The secondary criterion we employ to investigate the performances of different flexibility structures is the average throughput (i.e. total flow) of the corresponding systems, for which the rankings of the structures for data set I and II is given in Table 7 and Table 8, respectively. Note that these are not the optimal throughput levels but the throughputs that result from the capacity levels that optimize profit in each system.

Now we are ready to comment on how different parameters affect the structures and the capacity levels in detail. We do so for each demand rate ordering separately.

	$\mu_A < \mu_B < \ldots$			$\mu_A > \mu_B > \ldots$			$\mu_A = \mu_B = \ldots$		
Rank	O	N	C	O	N	C	O	N	C
# 1	5/20	50/100	0/0	16/25	0/0	83/95	0/120	0/0	0/0
# 2	4/5	5/20	46/95	83/95	0/0	16/25	0/0	0/109	0/11
# 3	46/95	0/0	9/25	0/0	99/120	0/0	0/0	0/11	0/109
Total	55/120	55/120	55/120	99/120	99/120	99/120	0/120	0/120	0/120

Table 5: Data set I: The average profit rankings for O(verflow), N(ested), (2-)C(hain)

	$\mu_A < \mu_B < \ldots$			$\mu_A > \mu_B > \ldots$			$\mu_A = \mu_B = \ldots$		
Rank	O	N	C	O	N	C	O	N	C
# 1	1/1	14/14	5/5	11/11	0/0	8/9	9/18	0/0	2/2
# 2	14/14	0/0	6/6	8/9	0/0	11/11	2/2	0/0	9/18
# 3	5/5	6/6	9/9	0/0	19/20	0/0	0/0	11/20	0/0
Total	20/20	20/20	20/20	19/20	19/20	19/20	11/20	11/20	11/20

Table 6: Data set II: The average profit rankings for O(verflow), N(ested), (2-)C(hain)

	$\mu_A < \mu_B < \ldots$			$\mu_A > \mu_B > \ldots$			$\mu_A = \mu_B = \ldots$		
Rank	O	N	C	O	N	C	O	N	C
# 1	0/14	29/64	3/42	0/14	0/0	101/106	3/70	0/6	5/44
# 2	1/6	2/41	29/73	101/106	0/0	0/14	4/15	1/47	3/58
# 3	31/100	1/15	0/5	0/0	101/120	0/0	1/35	7/67	0/18
Total	32/120	32/120	32/120	101/120	101/120	101/120	8/120	8/120	8/120

Table 7: Data set I: The average throughput rankings for O(verflow), N(ested), (2-)C(hain)

	$\mu_A < \mu_B < \ldots$			$\mu_A > \mu_B > \ldots$			$\mu_A = \mu_B = \ldots$		
Rank	O	N	C	O	N	C	O	N	C
# 1	0/0	14/14	6/6	5/5	0/0	14/15	7/13	0/0	2/7
# 2	12/12	0/0	8/8	14/15	0/0	5/5	2/2	0/5	7/13
# 3	8/8	6/6	6/6	0/0	19/20	0/0	0/5	9/15	0/0
Total	20/20	20/20	20/20	19/20	19/20	19/20	9/20	9/20	9/20

Table 8: Data set II: The average throughput rankings for O(verflow), N(ested), (2-)C(hain)

6.2.1 Unbalanced-Pareto Ordering: $\mu_A < \mu_B < \ldots$

Given the earlier stated match between this demand ordering and the career paths implied by the nested structure, it is not surprising to see that this structure performs the best in terms of profit for this demand order, as confirmed by both Tables 5 and 6. The main reason is that the capacity required in each resource decreases with the level of flexibility. As a result, optimal resource planning for the nested structure allocates the highest capacity to the resource with one skill and the least capacity to the one having all skills. Since the capacity costs, skill sets and demand rates are all aligned, unbalanced-Pareto ordering, i.e., $\mu_A < \mu_B < \ldots$ is the ideal case for the nested structure. A closer look at the details of the experiments summarized in Table 5 reveals that in all 20 cases where the nested structure is not the best (overflow is better) the demand rates are $(30, 40, 50)$, the most balanced demand rate configuration of this ordering. Hence, we can state that the nested structure is better when the demand rate asymmetry is higher.

Now let us consider the effect of the number of skills. In systems with three demand types, the nested structure never performs the worst, whereas with eight demand types it performs the worst 6 times. In all of these 6 problem instances the capacity costs are high. When the capacity cost is high in systems with many demand types, flexibility becomes very expensive for the nested structure since these systems have pools of agents ranging from one skill to eight skills. The capacity cost is increased for all but one type of resource in this structure.

Our other criterion is the average throughput obtained when capacities are set to profit maximizing levels. The nested structure performs the best under this criterion as well in both systems with 3 and 8 demand types (see Table 7 and 8). Now we can state our observation:

Observation 1 *When the demands are ordered as $\mu_A < \mu_B < \ldots$, the nested structure performs the best in terms of average profit and resulting average throughput in all systems except for those with high number of demand types, high capacity costs, and relatively balanced demand rates.*

The performances of the other two structures under this ordering can be summarized as follows: In systems with 3 demand types, the average profit of the 2-chain structure is

higher than that of the overflow, whereas this is reversed in systems with 8 types. The main reason for this change is the ability of the overflow structure to control the flexible capacity levels and so the associated costs. In the 2-chain structure, any capacity increase in one of the resources is actually an increase in the flexible resources. In the overflow structure, on the other hand, it is possible to increase the capacity of the specialized resources separately. In terms of throughput, the 2-chain structure performs better than the overflow in both systems with 3 and 8 demand types, confirming our capacity cost interpretation.

6.2.2 Unbalanced Ordering: $\mu_A > \mu_B > \ldots$:

For the 2-chain and overflow structures, this ordering is in fact equivalent to the unbalanced-Pareto ordering, since relabeling the demand types does not affect these structures. For the nested structure, on the other hand, this ordering is completely the reverse of the unbalanced-Pareto ordering, and consequently the nested structure performs the worst both in the average profit and the resulting average throughput criteria when $\mu_A > \mu_B > \ldots$.

As a natural consequence of the above discussion, comparing the performance of the 2-chain and overflow structures in systems with the unbalanced demand rate ordering results in the same observations with those under the unbalanced-Pareto ordering: The 2-chain structure achieves higher throughput in both systems with 3 and 8 demand types with higher flexible capacity cost. In terms of the average profit the 2-chain structure is better when the number of demand types is 3, and the overflow structure dominates in systems with 8 demand types. We summarize our observations as follows:

Observation 2 *When the demands are ordered as $\mu_A > \mu_B > \ldots$, the 2-chain structure performs the best in terms of average profit when the number of demand types is low, whereas the overflow structure achieves higher average profits when the number of demand types is high.*

6.2.3 Balanced Ordering: $\mu_A = \mu_B = \ldots$:

This ordering is the one most commonly studied in the flexibility literature. The symmetric demand structure implied by the balanced ordering is also where one expects the benefits

of flexibility to be the highest. It presents an interesting situation for systems with three demand types: Although the overflow structure generates the highest profit in all experiments, we cannot claim that it performs the best statistically. In fact, none of the rankings are statistically significant; in other words all three structures perform equivalently in terms of average profit.

When the number of demand types increases, the average profit of the overflow structure becomes significantly the highest, followed by the 2-chain. The performance of the nested structure deteriorates with higher number of demand types in terms of the average profit. These observations can be explained by our earlier discussion. The overflow structure has the highest capability to control the flexible capacity, and so the flexibility costs. The unit cost of all resources in the 2-chain structure remains the same, although the associated capacities may need to be large. Since all resources are flexible having 2 skills, this creates an unnecessarily high flexibility cost. Finally, in the nested structure the unit flexibility costs for almost all resources increases with the number of demand types. Then, except for the case when $\mu_A < \mu_B < \ldots$, it cannot balance this increase by adjusting the capacities according to the flexibility levels. As a result, the nested structure suffers the most from the increasing number of demand types, followed by the 2-chain. The relative performance of the overflow structure increases considerably with the number of demand types.

The above discussion highlights the performance of the overflow structure, which can be stated as:

Observation 3 *When the demand rates satisfy $\mu_A = \mu_B = \ldots$, the overflow structure performs the best in terms of average profit.*

6.3 Managerial insights on flexibility design

In practice, each system brings with it certain constraints. Sometimes the skill definitions are strictly protected. Sometimes the organizational structure of the system is established and cannot be changed. Driven by these constraints, for each system we need to identify a different managerial view and an opportunity to apply our results.

When the system has no constraints, flexibility design starts with the strategic issue of

skills definition. Our results suggest that in this case, the best strategy is to define the skills such that demand rates are balanced. This is in line with the intuition that the benefits of flexibility are highest in settings with high symmetry. Systems which receive the same total mean demand with balanced rates generate more average profit than systems with unbalanced rates, especially when the flexibility costs are high. For example, consider the demand rates $(10, 30, 90)$ and $(40, 40, 40)$, where the former is asymmetric with total demand 130 and the latter is balanced with total demand 120. Now we will compare two systems with these two demand rates in an environment operating with low variance and high capacity costs (configuration 3). When the system receives the asymmetric demand rates, the nested structure gives the best profit performance with an average profit of 3368.74 and an average throughput of 123.05. On the other hand, the same environment with the balanced demand rates leads to the best profit with the overflow structure generating an average profit of 3746.06 and an average throughput of 111.46. Even though the latter system receives less demand on average and so generates less throughput, it has higher average profit (these comparisons are statistically significant). We also observe that it is, in general, best to choose the overflow structure to maximize the profit, once skills have been defined to ensure a balanced ordering for the demand rates.

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 multi-language call centers, where each language has to constitute a different skill. Systems which cannot change their skill definitions, and so the demand rates, need to choose server skill sets given skill definitions. If the demand rates are balanced, then the overflow structure should be adopted to maximize the average profit. When the demand rates are unbalanced, the system should induce the nested structure aligned according to the demand rates. This is true as long as the number of skills is relatively low, and the recommended structure becomes an overflow structure once again if the number of skills is high.

What can we recommend to managers who manage centers where skill sets have already been established? First consider 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. In this setting, managers need to control the demand arrival rates. We know that the nested structure does not perform well, unless the demand rates are ordered as $\mu_A < \mu_B < \ldots$ with large enough differences between them. If the current skill definitions induce this order, all is fine; otherwise there may be a need to re-define the skills and re-build training programs accordingly. As a simple example, consider a bank contact center, where the current types include financial investments as type A, credit cards as type B, and banking as type C, with $\mu_A < \mu_B \approx \mu_C$. In order to induce the above order, we can identify different requests within credit cards and banking, and re-group these requests as type B' and type C', such that, for example, type C' agents need to learn basic information about credits cards in addition to banking, which leads to the order $\mu_A < \mu_{B'} < \mu_{C'}$. Such a change will create a system that operates more efficiently, possibly at the expense of longer initial training.

Next consider a center with an overflow structure, which basically consists of two types of agents: generalists and specialists. If the number of demand types is high, the system can keep the overflow structure since this structure is quite robust in terms of maximizing the profit. If the demand rates are ordered as $\mu_A < \mu_B < \ldots$ with large differences, the nested structure performs better. However, establishing a nested structure in such a system will require a high investment in training. When the number of demand types is low, and skills are defined to result in an unbalanced-Pareto ordering, the 2-chain structure can be implemented. This would require training the specialist agents in another skill. Organizing such a training program may be relatively easy and cheap.

Since the flexibility design problem in call centers concerns resources that are human, there are indirect human resource related costs and benefits associated with different flexibility structures that may drive the choice for a particular structure. As already mentioned in the Introduction, if managers want to pursue an inclusive human resource strategy, the 2-chain or nested structures provide alternatives that are more consistent with this strategy. The overflow structure is consistent with a segmented human resource strategy. While the overflow structure comes out as a robust flexibility structure, as the number of skills

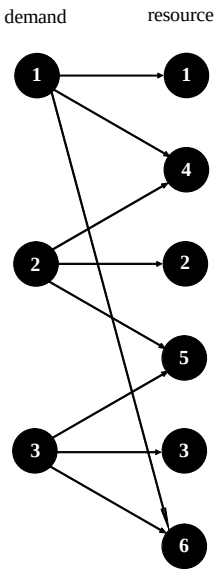

Figure 5: An overflow structure having flexible pools with 2-skill agents

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) demonstrates 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 2-chain structure may be preferred, or the overflow structure needs to be adapted such that there are several partially flexible overflow pools instead of one fully flexible overflow pool (see e.g., Figure 5).

7. Concluding Remarks

We analyzed the capacity decision of multi-resource 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 temporal and stochastic demand variability found in call centers. Suppressing the randomness in capacity simplifies the analysis, and enables us to propose a method that is generally applicable to any flexibility structure or setting.

The numerical comparison of the three flexibility structures, found in practice or recom-

mended by the literature, illustrates the importance of optimizing capacity given a flexibility structure and shows that profit and throughput performance are not necessarily the same. A nested skill-set structure, preferred due to the ease with which it offers a career path to employees, is shown to perform well in settings with matching nested demand rates. On the other hand, a structure with a select group of flexible experts supported by specialists is shown to be essential in settings where additional server skill costs are high. The overflow structure can be implemented as long as the scope of the skills does not exceed one server's learning capabilities. Balanced skill sets will be preferred in settings where costs are not that important, while good customer service levels are essential.

Even though our model does not consider queueing for capacity, comparisons to other models and results from the literature provide support for the use of such modeling in the analysis of flexibility structure. 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. (2008). Analyzing queueing systems in heavy traffic, with their capacity optimized, under the assumption of balanced demand rates, Bassamboo et al. (2008) show the optimality of flexibility structures that are consistent with the overflow structure having flexible pools with 2-skill agents (see Figure 5). 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. (2008) lead to consistent results with our observations lends further support to the robustness of these with respect to modeling specifics.

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

References

Akşin, O. Z., Armony, M. and Mehrotra, V., 2007a. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16(6), 665-688.

Akşin, O. Z., and Karaesmen F., 2002. Designing flexibility: characterizing the value of cross-training practices. *INSEAD, Working Paper*.

Akşin, O. Z., and Karaesmen F., 2007. Characterizing the performance of process flexibility structures. *Operations Research Letters*, 35, 477-484.

Akşin, O. Z., Karaesmen F., and Örmeci L., 2007b. A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective. In *Workforce Cross Training Handbook,* D. Nembhard (ed.), CRC Press LLC, 211-240.

Akşin, O. Z., de Vericourt, F., and Karaesmen, F., 2008. Call center outsourcing contract analysis and choice. *Management Science* 54(2), 354-368.

Bassamboo, A., Randhawa, R. S., and Van Mieghem, J. A., 2008. A little flexibility is all you need: Optimality of tailored chaining and pairing. *Working Paper*, Kellogg School of Management, Northwestern University.

Chevalier, P., Shumsky, R. A., and Tabardon, N., 2004. Routing and staffing in large call centers with specialized and fully flexible servers. *Working Paper*.

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

Çakan, N., 2006. Joint flexibility and capacity design in service and manufacturing systems. *M.S. Thesis*, Koç University.

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

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

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

Graves, S. C., and Tomlin B.T., 2003. Process flexibility in supply chains *Management Science*, 49(7), 907-919.

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 van Mieghem, J. A., 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *European Journal of Operational Research*, 113, 17-29.

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.

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(1), 83-98.

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

Ilgen, D.R. and Hollenbeck, J.R., 1991. The structure of work: job design and roles. In *Handbook of Industrial and Organizational Psychology 2,* M. D, Dunnette and L. M. Hough, editors, 165-207.

Inman, R. R., Jordan W. C., and Blumenfeld, D. E., 2004. Chained cross-training of assembly line workers. *International Journal of Production Research*, 42(10), 1899-1910.

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

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

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., Inman, R.R., and Blumenfeld, D. E., 2004. Chained cross-training of workers for robust performance *IIE Transactions*, 36, 953-967.

Karaesmen, I., and van Ryzin, G., 2004. Overbooking with substitutable inventory classes. *Operations Research*, 52(1), 83-104.

Netessine, S., Dobson, G., and Shumsky, R. A., 2002. Flexible service capacity: optimal investment and the impact of demand correlation. *Operations Research*, 50(2), 375-388.

Robbins, H., and Monroe S., 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 400-407.

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.

Talluri, K., and van Ryzin, G., 1999. A randomized linear programming method for computing network bid prices. *Transportation Science*, 33(2), 207-216.

van Mieghem, J. A., 1998. Investment strategies for flexible resources. *Management Science*, 44(8), 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 & Service Operations Management*, 7(4), 276-294.

Xie, J.L. and Johns, G., 1995. Job scope and stress: can job scope be too high? *Academy of Management Journal*, 18, 1288-1309.

A. Data Sets Used

	$\mu_A < \mu_B < \ldots$						$\mu_A > \mu_B > \ldots$						$\mu_A = \mu_B = \ldots$					
Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	30	10	5	10	10	10	50	90	80	160	70	90	5	10	30	40	80	160
B	40	30	15	40	50	50	40	30	15	40	50	50	5	10	30	40	80	160
C	50	90	80	160	70	90	30	10	5	10	10	10	5	10	30	40	80	160

Table 9: Mean demand rates, μ_A, μ_B, μ_C, in data set I

	$\mu_A < \mu_B < \ldots$	$\mu_A > \mu_B > \ldots$	$\mu_A = \mu_B = \ldots$
A	5	640	80
B	10	320	80
C	20	160	80
...	...	...	...
H	640	5	80

Table 10: Mean demand rates, $\mu_A, \mu_B, \ldots, \mu_H$, in data set II

B. Validity of the GPA Approach

Glasserman 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:

$$E[\nabla_\theta X(\theta)] = \nabla_\theta E[X(\theta)]. \tag{7}$$

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* $\nabla_{\mathbf{c}} E\left[\Phi(\mathbf{c},\mathbf{D})\right]$.

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:

$$|\Phi(\mathbf{c} + \varepsilon e_j, \mathbf{d}) - \Phi(\mathbf{c}, \mathbf{d})| \leq \varepsilon \min\{\bar{p}_j, f_j + t_j\}, \tag{8}$$

where e_j is the j^{th} 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. Until any of the gradients change sign at an iteration t, which means that one of the capacity values passes over the optimal point, a fixed step-size of 1 is

used. After that point, we begin decreasing the step-size according to the rule $b_k = 1/(k-t)$. According to the conditions given below:

$$\sum_{k=1}^{\infty} b_k = \infty \qquad\qquad \sum_{k=1}^{\infty} b_k^2 < +\infty, \tag{9}$$

which were established by Robbins and Monroe Robbins and Monroe (1951) to guarantee the convergence. Our implementation satisfied the two condition and therefore converges.