University of Alberta

Service Quality, Employee Satisfaction, and Cross-Training:
A Broader Look at Workforce Scheduling and Rostering

by

Yong Yue Li ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of *Doctor of Philosophy*

in

Management Science

Faculty of Business

Edmonton, Alberta

Fall 2005

0-494-08676-9

Your file Votre référence
ISBN:
Our file Notre reterence
ISBN:

The fear of the Lord is the beginning of knowledge, …

Proverbs 1:7

To My Parents

Abstract:

Service level, the percentage of customers served within a predetermined time interval, is the most widely used quality indicator in the service industry in personnel scheduling; however, it is doubtful that service level alone represents as broad a measure of service quality as management would desire. The dubious relationship between service quality and service level is what inspired this study. With modern trends of providing more complicated services for more sophisticated systems in a more competitive market, there are many criteria other than service level that need to be considered in order to make good workforce scheduling and rostering decisions. This problem is especially difficult when dealing with systems having time-varying stochastic demand throughout the day, namely $M(t)/G/s(t)$ queueing systems, which are widespread in service industry.

The traditional scheduling and rostering approaches require the application of analytical queueing and integer linear programming methods. Though currently well accepted in the industry, these methods have several limitations. Analytical models involve highly simplified queueing systems and/or nonrealistic assumptions, and might not be able to generate all service quality indicators management is interested in. Obtaining optimized schedules and rosters for a large system is very tedious, and optimality is not always guaranteed. Scheduling and rostering are typically conducted as two independent processes; consequently, employee satisfaction is completely ignored during scheduling and two optimization procedures are needed. In addition, traditional approaches lack the

ability to consider multiple service quality and employee satisfaction criteria simultaneously and cannot handle employees who are cross-trained in multiple tasks.

A novel multi-objective framework which overcomes these many limitations is proposed in this study. Both schedules and rosters are generated in one integrated step; among all efficient ones recognized the best schedule and/or roster is chosen with the direct input of management, in accordance with their particular business strategy. This framework can also potentially be used for cross-trained employees. Cross-training strategies are identified for systems with various demand curve patterns and levels in terms of employee pooling and timing. These strategies will provide guidelines to expand the proposed approach in a multi-tasking setting in the future.

Acknowledgements

I have imagined writing this part of the dissertation many times during the course of my doctorial studies. Without help, I could never have made it to this stage. Recalling the past several years spent at the University of Alberta, I am very grateful for all the encouragement and support I have received.

I thank God for watching over me all the time and keeping me on the right path. He listened to my prayers and blessed me. I thank Him for teaching me and giving me everything.

I would like to express my gratitude to my supervisors Professors, Tarja Joro and Ignacio Castillo, for all the opportunities that they gave me and their patience. They walked me through all the difficulties in completing this project. Their guidance and training provided me with knowledge in research and made me grow into an independent researcher. I especially want to thank Professor Tarja Joro for her tremendous support, care, and advice.

I have been fortunate to have Professors Joseph Paradi and Ming Jian Zuo as my external examiners, not only because their comments and advice improved my thesis but also because their enthusiasm about my work gave me faith in doing the research. I feel privileged to have had them on my examination committee.

It has been a real pleasure to work in the Management Science group. I am thankful to Professor Terry Daniel for chairing my defense, for his great advice, and for his kindness. I am also grateful to Professor Armann Ingolfsson for the summer project that he offered me, which became the subject of my first paper, and his countable comments and suggestions on my work. Professor Erhan Erkut was very accessible for research opinions, excellent teaching tips, and career advice. I have really appreciated his help. Professor Prem Talwar, Professor Payman Jula, Keltie Tolmie, Louise Hebert, and Dan Haight provided me with great help and assistance whenever needed.

The PhD program provided me with a wonderful opportunity to meet future scholars and make friends. First, I would like to thank our Director, Professor David Cooper, as well as Jeanette Gosine, and Kathy Harvey for their excellent work in running this program. In all my years as a PhD student, I have to say that it has always been delightful to see them being there for us. Being able to be around Susan Budge, Edgar Cabral, Kathy He, Jin Li, Annie Niu, Sherry Oh, Yuan Tian, Bo Wu, and Fan Yang have made my life easier because of their care, encouragement, and valuable information. I would like to thank them all and other students in the program for the laughter we shared together and good conversations.

There are also several wonderful people I have encountered during my study: Dave Clyburn, Shauna Crowe, Anne McNaughton, Jarett Prouse, Shaojun Zhao, and Professor Paul Zipkin, all of whom showed their extreme kindness to me. My friends Jarett and Shaojun both inspired me in my research. In particular, Jarett provided me tremendous support and edited most of my thesis.

Finally, I would like to thank my parents, Anliang Li and Shuyi Xu, for their eternal love and faith in me. Their care and understanding have supplied me with the strength and courage to face hardships and future challenges...

Yong Yue Li
July 4, 2005
Edmonton, Alberta, Canada

Table of Contents

List of Tables

List of Figures

1. Introduction

The importance of workforce scheduling and rostering problems as a managerial issue has grown with the exponential expansion of the service industry in our society. To attract and retain customers, organizations in this industry have explored many reforms in their operations such as the extension of business hours. Businesses running 24 hours a day 7 days a week are quite common and many services are available from early in the morning until late at night. Such conveniences have given customers more freedom and choice in service times, resulting in formerly concentrated demand becoming much more spread out, irregular, and easily influenced by customers' whims. This trend had brought new challenges to management – how do you meet this new time-varying demand while still remaining competitive in the market? For most firms in the service industry management of the workforce is one of the most critical tasks they face. It is often a primary source of competitive advantage.

The service industry is typically labour intensive, meaning that labour comprises a large portion of expenditures compared to capital. Staffing costs in the service sector constitute major expenses in total operating costs. Approximately 60 to 70% of operating costs in call centres are spent on employees (Fukunago et al. 2002) and around 30% in hospitality industry (Ernst et al. 2004). Even a small percentage of savings in labour expenditures can make a noticeable difference in a company's profitability. In addition, not only does a company's workforce constitute the core of its service facility, but it also represents a major part of the company's image to the public. Workforce scheduling and rostering therefore significantly impact the performance of an organization and are critical issues in the service industry. Since the 1950's, the highly promising potential of increasing personnel utilization and service quality has increasingly catalyzed research in workforce scheduling and rostering.

In addition, workforce scheduling and rostering decisions influence many people's daily lives. Practically everybody receives some sort of service during the day. There were 11,961,500 people employed in the service-producing sector in May 2004, which comprises around 74.7% of employment in all industries in Canada according to Statistics Canada. An inefficient schedule might generate negative consequences by imposing long wait times on customers or cause service errors resulting from overworked employees. Workforce scheduling and rostering are just as relevant to the satisfaction of employees as to customers. Since human resources are one of an organization's most valuable assets, a successful organization must concern itself with its employees' working conditions, morale, and personal growth in order to retain its strength. Workforce rostering techniques determine employees' working times and the intensity of their workload, and also affect their ability to grow; therefore, they should never be considered trivial.

Workforce scheduling refers to a method that aims to find employee shift arrangements to match service demand while keeping costs under control and satisfying all applicable regulations, rules, and laws (e.g., shift lengths and spacing of breaks). There are three characteristics of the demand for services that make this task particularly difficult. First, random customer arrivals are stochastic and usually vary throughout the day; therefore, it is never certain how many customers are coming in the next moment and staffing requirements can only be forecasted with limited precision. Second, services, unlike physical products, usually cannot be inventoried. The surplus of supply perishes at the time it exists. This presents a unique problem to the management in the service industry and complicates the task of determining an appropriate service capacity. Finally, service quality is difficult for management to evaluate. Unlike goods produced in manufacturing systems, the product of service systems is abstract and intangible. Measurements used for quality control are not at all straightforward to generate.

Workforce rostering is a process of assigning shifts in a predetermined schedule to specific employees in traditional view. The main concern in rostering is how to maximize employee satisfaction without sacrificing service quality or increasing costs. The intrinsic difficulties in rostering lie in the measurement of individual employees' preferences for shifts and how to maximize employee satisfaction as a group. Personal preferences are highly subjective and are therefore not easy to evaluate precisely. When it is impossible to guarantee that all employees will receive their most desired shifts, defining criteria for assessing employee satisfaction as a group becomes problematic.

The workforce scheduling and rostering methods and cross-training strategies introduced in this dissertation are aimed at medium and large scale service facilities with time-varying demand profiles. Such facilities have become very widespread – airline services, retail businesses, financial services, fast food restaurants, police offices, call centres, post offices, and so on. There were for instance 50,600 call centres in the US containing 2.86 million agent positions and 4,500 call centres in Canada containing 212,000 agent positions in 2004; this will increase to 5,300 call centres with 305,500 agent positions in Canada by 2008, according to *McDaniel Executive Recruiter's 2004 North American Call Center Report*. The proposed methods are therefore applicable to numerous service facilities.

Employee scheduling and rostering in a time-varying queueing system are difficult enough to solve in a single-service system. The matter is further complicated due to the fact that a large percentage of these facilities have come to provide multiple services to customers. Instead of having dedicated employees trained in only one task, employees may be cross-trained in secondary tasks so that they can provide several different services to customers. Employee cross-training is an effective strategy to improve employee utilization and substantially improve service quality. It provides further benefits as well, such as facilitating better communications within organizations. However, adding cross-training practices into rostering a time-varying multi-service system creates numerous variations in decisions concerning who should be cross-trained in what tasks, to what degree, and how cross-trained employees should be scheduled. Many intricate complications arise when cross-training optimization problems interact with the original scheduling and rostering problems.

This dissertation aims to contribute comprehensive and practical ideas to these three topics – workforce scheduling, workforce rostering, and cross-training.

The primary objectives of the proposed workforce scheduling approach are to incorporate a broader view of service quality, and to avoid the unrealistic assumptions of currently established methods. This is accomplished by characterizing and quantifying service quality using multiple, complementary criteria, thereby allowing managers to investigate the tradeoffs between labour costs and a variety of service quality indicators. This option is often critical for managers to effectively position their company and decide on a realistic corporate strategy based on these relationships.

The current paradigm in workforce scheduling is to minimize the labour costs subject to meeting a target service level, which is defined as the percentage of customers served within a predetermined time interval. Since the seminal papers by Edie (1954) and Dantzig (1954), this methodology has been applied to workforce scheduling in many organizations – from police departments to laboratories to call centres (see Agnihothri and Taylor 1991, Brusco et al. 1995, Callahan and Khan 1993, Gopalakrishnan, Gopalakrishnan and Miller 1993, Harris, Hoffman and Saunders 1987, Sze 1984, Taylor and Huxley 1989 among others). Service level in these studies is the only criterion employed to measure service quality. However, it is evident that the two are not equivalent; a good service level does not necessarily equal good overall service quality. Multiple service quality indicators are therefore introduced in the proposed approach to remedy this problem.

In addition to this one-dimensional modeling of service quality, another concern in traditional workforce scheduling models is the use of rather limiting assumptions. These include the use of exponentially distributed service times resulting from the *M/M/s* queueing assumption and the steady state assumption arising from the traditional Stationary Independent Period by Period (SIPP) approach. An *M/M/s* queueing system has exponentially distributed inter-arrival and service times and multiple servers. Additionally, the sequential nature of the steps taken to determine server requirements and actual schedules in the traditional method is proven to be problematic in the literature. All these issues are addressed in the proposed scheduling method.

There are three steps involved in the new method. First, many plausible schedules are generated using a heuristic, after which the various service quality indicators considered are calculated for each schedule. A multidimensional efficiency analysis tool is then used to identify non-dominated schedules based on the criteria. Finally, the best schedule is chosen from among all efficient ones using the same analysis tool together with management involvement. Experiments in this work are conducted using industrial data; the results obtained are better than or comparable to those generated by the traditional method in all dimensions except one.

The primary aims of the proposed rostering approach are to consider issues of service quality, labour costs, and employee satisfaction comprehensively and simultaneously; and to build on and incorporate all the advantages of the new scheduling approach. Employee shift preference measurements are studied and various factors which may influence these preferences are identified. Criteria for the evaluation of overall employee satisfaction in a roster are also examined.

Labour costs, service quality, and employee satisfaction cover the essential issues of business operations in the service sector. The objectives of minimizing cost and maximizing service quality and employee satisfaction conflict, and these three aspects are highly interdependent, but all are essential components of a good roster and should therefore be considered simultaneously. Yet little research has been conducted in generating rosters to meet the integrated criteria of these three elements, and the interactions among them are ignored when they are regarded separately. Evaluation of employee preferences and employee satisfaction as a group determines the performance of a roster in terms of employee satisfaction. However, the literature in workforce rostering commonly fails to explore various options for measurements and lacks real-world evidence from use of the measurements that it chooses.

The rostering approach has a three-step framework very similar to that of the scheduling approach. Three heuristics are created to generate many plausible rosters and calculate their employee satisfaction indicators. The next step is to compute all the service quality indicators. The same efficiency analysis tool is then used to identify non-dominated rosters using both employee satisfaction and service quality criteria, and the best roster is chosen with managerial input. The same sets of experiments using industrial data are conducted and the results are compared to those generated by their traditional method. One of the proposed heuristics generates employee satisfaction results that are significantly better than the traditional method in all dimensions. Results from the other two methods are also better than or comparable to the traditional method's results.

To speculate on whether the proposed rostering method, with some adjustments, can be applied in a multi-tasking workforce system, a study of cross-training is conducted in this dissertation, focusing on queueing systems with multiple time-varying demands. This study strives to provide management with information about under what circumstances a cross-training strategy should be applied; and when cross-training is implemented, who should be cross trained and how cross-trained employees should be scheduled.

Service systems have become very sophisticated in modern society as companies endeavor to provide more and better services to their customers. Systems that offer multiple services, have multiple types of customers, or require the use of multiple languages to help customers are pervasive. It is likely that various demands in the system are time-varying and have different characteristics. However, so far, there are few studies in the literature applying cross-training strategies in such systems – this study is one of the first.

Simulation experiments are conducted using various arrival patterns of two time-varying demands to show the impact of cross-training decisions and establish guidelines for cross-training in such systems with two demands. It is found that various combinations of arrival patterns in the two demands have little effect on cross-training decisions but that the average levels of demand do. In a system with non-symmetric demands, cross training employees exclusively in the low demand pool improves service quality performance tremendously. When determining which shifts to assign to cross-trained employees, it is beneficial to position them in periods where the volumes of the two types of demand are similar.

This dissertation is structured as follows. Chapter 2 describes the motivations of, and overviews, the proposed scheduling and rostering approaches. Chapter 3 introduces a plausible schedule generation method and three plausible roster generation methods. The evaluation element for both scheduling and rostering approaches is discussed in detail in Chapter 4. Chapter 5 explains the efficient analysis tool that is used to identify non-dominated schedules or rosters and how it is applied when choosing the best schedule or roster. Cross-training strategies for an $M(t)/G/s(t)$ time-varying queueing system with two demands are illustrated in Chapter 6, followed by experimental results of all approaches introduced in the thesis in Chapter 7. Concluding remarks and directions for further research are provided in Chapter 8.

2. Workforce Scheduling and Rostering: a New Paradigm

2.1 Introduction

Workforce scheduling and rostering are important functions for the daily operations of many businesses, particularly in the service industry. Good schedules and rosters provide customers with timely and accurate services; employees with preferred work times and a moderate workload; and organizations with competitive labour costs and increased quality of the provided service. This dissertation introduces a new paradigm for workforce scheduling and rostering, to consider the needs of customers, employees, and organizations simultaneously, avoid the strong assumptions of traditional models, and enable the generation of schedules and rosters in a single integrated step.

This study focuses particularly on the type of service systems referred to as $M(t)/G/s(t)$ queueing systems, which are widespread in modern society. In such systems, customer demand is random, average demand usually changes throughout the day, the number of servers on duty changes in response to demand, and the probability distribution of customer service times is not restricted to a particular distribution, but any type that fits into the real situation, such as Exponential, Erlang, Triangular, or Gamma distribution. Since the number of needed servers changes with customer demand, it is challenging to generate a cluster of shifts that abide by union regulations, company rules, and labour laws, while also producing aggregated staffing levels that satisfy system labour requirements – this consists of a workforce scheduling problem. Generated shifts then need to be assigned to particular employees, a process which involves a workforce rostering problem.

The most commonly used workforce scheduling methods aim to minimize labour costs while meeting a predetermined target service level, which is defined as the percentage of customers served within a preset threshold time. To accommodate the time-varying system, the span of service operation time is usually partitioned into short and equal length periods e.g. half an hour, which are defined as planning periods. A simple yet powerful analytical model is frequently used to calculate the servers required for each planning period according to the objective function with the assumption that the queueing system follows the *M/M/s* model. Each planning period is considered to be independent from each other. After the number of servers (also referred to as staffing level) is determined for each planning period, a set of shifts is generated using an integer linear programming model.

There are three main flaws with this approach. First is that it imposes strong assumptions about the system, namely that the service time must be exponentially distributed and that operations in one period do not interfere with operations in other periods, neither of which need be true in a particular system. Second, meeting a specified service level is the one and only constraint in the model, and thus is the only indicator of service quality. However, it is very doubtful that a complete measure of service quality can be adequately represented by this single indicator. Finally, the integer linear programming model can become very computationally demanding. Due to the complicated nature of optimization programs, it is difficult to arrive at an optimal solution in a reasonable amount of time when the service system is itself complicated. The proposed approach addresses these three problems in the traditional method by utilizing a very different strategy when solving scheduling problems.

Traditionally, work schedules are determined first, after which rosters are generated by assigning shifts in the schedule to individual employees according to their availability. However, previous literature provides evidence that indicates that this two-step approach can be problematic in that it is very difficult for employees to receive desirable shifts. Moreover, it presents difficulties when trying to balance the conflicting and interdependent needs of customers, who demand quality services; employees, who seek job satisfaction; and organizations, who attempt to minimize the cost of business operations. This study shows that the proposed scheduling and rostering approaches can together not only generate schedules and rosters in a single step, but are also capable of considering these issues of service quality, employee satisfaction, and cost simultaneously.

This chapter introduces the rationale and procedures of the new workforce scheduling and rostering approaches. It explains in detail the limiting assumptions and weaknesses in the traditional scheduling approach and discusses a multiple-objective measure of service quality, in Sections 2.2.1 and 2.2.2 respectively. The advantages of the proposed rostering approach, including a better consideration of employee preferences, are presented in Section 2.3. An overview of the new approach can be found in Section 2.4.

2.2 Motivations for the New Scheduling Method

There are two main objectives of the new scheduling method. First, it aims to solve various problems without imposing the strong assumptions seen in traditional methods such as the SIPP approach. It also seeks to provide a multi-objective view of service quality, instead of adopting service level as the one and only indicator of service quality as is widely applied in the literature.

These two motivations are elaborated in Sections 2.2.1 and 2.2.2.

2.2.1 Limitations of the Traditional Approach

The traditional approach to workforce scheduling consists of four steps (Thompson 1995a): 1) forecast demand, 2) obtain staff requirements based on the forecasted demand, 3) schedule shifts to meet staff requirements, and 4) real-time control. Since steps 1 and 4 fall outside of the actual scheduling process, our emphasis is on steps 2 and 3.

In step 2, a stationary *M/M/s* queueing model is commonly used to estimate the minimum number of servers needed to meet a target service level. A stationary system requires average arrival and service rates to be the same throughout the day. However, in real life, most organizations have time-varying demand. In order to use an $M/M/s$ queueing model under this condition, the service period is partitioned into equal, short planning periods (usually 30 minutes or one hour). It is assumed that the average arrival and service rates are constant in each planning period, and that the system reaches steady state at the beginning of each planning period. With these assumptions, an $M/M/s$ queueing model can then be used to compute the minimum number of customer representatives needed to meet a certain service level in each planning period. This approach of obtaining staffing requirements (the minimum number of representatives needed to meet a target service level in each planning period) is usually called the SIPP approach, which basically uses a series of stationary $M/M/s$ queueing models to calculate the servers needed for a $M(t)/G/s(t)$ service system. Note that each planning period is treated independently in this approach.

It is important to see that blocking (where customers are not able to join the waiting line that reaches its capacity) and reneging (where customers quit remaining in line due to long waits) behaviours are not considered due to the simplified $M/M/s$ assumption, although they are quite common in the service industry. Recall that a $M(t)/G/s(t)$ service system is constantly in transit state and likely to experience blocking and reneging behaviours.

Step 3 transforms daily staffing requirements into a schedule. A shift is defined as a set of intervals during which a customer representative works in a day, and a schedule here refers to a set of shifts that provides the total staffing requirement in a day. For example, a feasible shift may require a customer representative to work from 8:30 to 17:30, with a half-hour lunch break at 12:00 and two 15-minute coffee breaks, one at 10:00 and the other at 15:30. The scheduling problem is an optimization program that minimizes the cost of labour of a schedule while satisfying the staffing requirements in each planning period. There is a considerable amount of literature available that attacks this problem through various approaches with various perspectives (for a brief summary see Gans, Koole, and Mendelbaum, 2003). The standard approach is to use integer linear programming formulations to choose from all shifts to cover the staffing requirements at the minimum cost. However, the number of possible shifts can be very large, not only due to the various possible starting times and shift lengths, but also, and especially, due to the various possible locations of coffee and lunch breaks. Due to the large number of possible shifts, the optimization program is complicated to formulate and computationally expensive to solve: an optimal solution usually cannot be guaranteed.

More importantly, even if an optimal solution is obtained, the schedule does not necessarily provide the target service level established in the first place due to the intrinsic deficiencies in steps 2 and 3. Recall that in step 2 several limiting assumptions were made to facilitate the use of the SIPP approach. Green, Kolesar and Soares (2001) identify the conditions under which the SIPP approach fails and highlight the underlying reasons. One of the limiting assumptions is that the service time follows an exponential distribution; this is rarely the case in real life. In addition, the SIPP approach does not consider blocking and reneging behaviours; behaviours that are well documented in service sector literature. Furthermore, the SIPP approach assumes that each planning period is independent. Thus, customers still waiting in the previous period will not be carried over to the next period, providing understaffed results for those periods, especially during times of heavy traffic. As the planning periods get shorter, the stationary state assumption in the SIPP approach also becomes questionable.

Another apparent limitation lies in the sequential nature of steps 2 and 3. Several studies indicate that workforce scheduling in sequence from step 2 to step 3 can give misleading results (see Easton and Rossin 1996 and Thompson 1999 for details). The main drawback presented in this sequence is that the optimal schedule is generated without considering the employee information; as a result, scheduled employees might not actually be available, or the schedules produced may violate the applicable regulations (e.g., shift lengths and spacing of breaks). Tien and Kamiyama (1982) indicate that further research is needed to simultaneously consider steps 2 to 3. Moreover, recent studies (see for example Ingolfsson, Cabral, and Wu 2002) provide evidence that significant cost savings can be obtained if these steps are considered simultaneously. Thus, an approach that integrates steps 2 and 3 seems to be greatly preferred.

The new scheduling approach proposed in this work succeeds in integrating these two steps. Rather than performing two separate steps, first to determining staffing requirements and then to solve optimization programs to generate schedules, they are instead generated directly from the time-varying demand profile. This method also enables us to avoid the strong assumptions and limitations of the traditional approach, allowing the use of more flexible and realistic models of customer demand.

2.2.2 A Multi-Criteria View of Service Quality

There is a significant amount of literature regarding workforce scheduling in the last fifty years that use the traditional objective function, which is to minimize cost while meeting a certain service level. However, it is the overall service quality that really matters to the management; it is not very convincing that achieving quality service is as straightforward as merely meeting a specified service level. The service level provides some information about the fraction of assumingly satisfied customers, but says little about those who have to wait longer than the threshold time. For example, information about the longest customer waiting time is not available. With the level of competition in modern businesses, management is increasingly under pressure to not only meet a certain service level but also consider other quality indicators to ensure good overall service quality.

While the use of a target service level can currently be considered as a sector-wide standard, one can argue that a single operational measure is not sufficient to capture the performance of service organizations, nor to characterize and quantify service quality. In fact, there is little evidence in service sector research literature that suggests any direct association between service level and service quality.

This observation is not new: indeed, several researchers have previously acknowledged the ambiguous relationship between a target service level and service quality, and have attempted different approaches to resolve or avoid the ambiguity. One of these approaches is to expand the cost component of the objective function to account for the cost of poor service and the cost of waiting (see for example Andrews and Parsons 1993, Grassmann 1988, Koelling and Bailey 1984, and Mabert 1979). The challenge in this approach, however, is the estimation of these costs (see for example Baker 1976, Taha 1981). Such estimation is rarely accurate and apt to be unique for each type of service organization, as it depends on the customers' response to waiting, which is affected by various factors that are very likely to be unique for different organizations (see for example Jackson 2002, Katz, Larson and Larson 1991). The task of estimating the costs thus adds a serious barrier to implementing this approach.

The inadequacy of considering only service level to evaluate service quality can also be verified by looking at the performance of various service systems. It is not difficult to find systems having similar service levels but different values for average wait time, maximum wait time, blocking rate, or reneging rate. In such cases, if the only indicator considered is service level, then the apparent service quality of those systems will be the same, which is obviously erroneous.

Service quality is an abstract concept and is admittedly difficult to evaluate precisely. It is not necessarily clear which operational criteria are relevant to service quality and how the two are related. However, examining various indicators and aspects will definitely give management a better view of the overall quality of service than any single factor.

The proposed multiple-objective measure incorporates cost and several aspects of service quality in workforce scheduling. Service quality is characterized by various indicators such as average and maximum waiting times, average and maximum queue lengths, blocking and reneging rates, service level, and personnel utilization, rather than the usual single measure of service level. In addition, there is no limitation on the criteria which can be incorporated using the proposed approach. Employment of these multifarious criteria provides a broader, more realistic, and ultimately more useful understanding of service quality, from a variety of different angles and perspectives beyond those obtained by the traditional approach.

2.3 The Motivations of the New Rostering Method

Given the success of the proposed workforce scheduling approach, it is natural to further this research by developing a new workforce rostering approach, as well. This approach incorporates the new scheduling method and thus as expected inherits its previously mentioned merits; additionally, it aims to consider employee satisfaction from various perspectives, and do so simultaneously along with concerns of cost and service quality. In Section 2.3.1, the importance of including employee satisfaction is explained. A review of current literature in workforce rostering is provided in Section 2.3.2.

2.3.1 The Missing Dimension – Employee Satisfaction

Human components may well be the most important assets and resources in the service industry, as its success depends equally on both customers and employees. To be profitable, an organization needs to provide quality services to its clients, who are the sources of revenue, and at the same time minimize its operational costs. However, these two objectives usually conflict with each other. Simply cutting costs implies either a decrease in total service hours or a negative impact on employee compensation. The former may affect service quality, which undermines customer loyalty; the latter may influence employee satisfaction, which also may affect customer service, and can result in high turnover, thereby increasing recruiting and training costs. In addition, although perhaps not as obvious as the influence of labour costs on service quality and employee satisfaction, past research has shown a connection between employee satisfaction, and service delivery and resource utilizations.

Bateman and Organ (1983) pointed out that satisfied employees tend to be better corporate citizens; that is, they are more compliant, more altruistic, more dependable, more cooperative, less critical of others, less argumentative, and more punctual. Studies also reveal that dissatisfied employees have a potential negative impact upon service delivery (Schlesinger and Heskett 1991). Moreover, it is documented in the literature that employee satisfaction impacts service quality, employee recruiting, tardiness, absenteeism, and turnover (Elliott 1989, Hung 1992, Mahoney 1978, Ostroff 1992, and Turney and Cohen 1983).

Having noted the influence of employee satisfaction on profitability through its effects on service quality and productivity, it becomes evident that employees' preferences cannot be ignored, even if only the benefits to companies are being considered. Indeed, many factors can affect employee satisfaction towards work, including working environment, involvement in decision-making, training programs, personnel relationships, compensation policies, future opportunities, working time and so forth (Yeung and Berman 1997 and Wright et al. 2001). Among these, working time is one of the few that is relevant to workforce scheduling and rostering. While other factors such as relationships among personnel may also have certain implications, they are not considered in this research because it greatly complicates the problem and it would be extremely difficult to access such data. Silvestro and Silvestro's (2000) research shows that employees' attitudes to work scheduling can influence their delivery of service, productivity, and satisfaction. A roster that fails to take into account the domestic and social needs of staff is likely to give rise to considerable staff dissatisfaction. As the timing of a shift is characterized by work starting time, breaks, and shift lengths, none of these three factors should be ignored in order to satisfy employee preferences concerning work schedules when rostering.

Although the literature has recognized the importance of employee satisfaction since the 1980s, research that incorporates employee satisfaction in workforce scheduling and rostering is sparse. Few papers have addressed employee availability or employee satisfaction with rostering and none have accounted for all three factors mentioned above. Yet, in practice, managers try to maximize employee satisfaction for a given schedule on a regular basis.

Despite its importance, employee satisfaction is often neglected in traditional approaches, partly due to the sequential nature of schedule and roster generation. Since rostering is accomplished only after schedules are found, the satisfaction of employees can only be considered after first dealing with service quality issues. It is thus difficult to ensure that employees receive desirable shifts without compromising the quality of service. The proposed new rostering approach overcomes this limitation by incorporating the scheduling method to enable the simultaneous generation of schedules and rosters. This makes it possible to consider employee satisfaction, cost and service quality issues comprehensively and simultaneously, and thereby strike a better balance between these three factors than was possible under the traditional methodology.

2.3.2 Current Rostering Approach Considering Employee Satisfaction

Traditional rostering for a single day (Moondra 1976, Gaballa and Pearce 1979, Bechtold and Jacobs 1990, Thompson 1995b and Berman, Larson and Pinker 1997 among others) and on a weekly basis (Bailey 1985, Burns and Carter 1985, Jarrah, Bard and deSilva 1994, Easton and Rossin 1996 and Jacobs and Brusco 1996 among others) are problems that have been extensively dealt with. While employee preferences are considered to be important to the rostering problem, the issue has not yet attracted significant attention in the literature. Traditional rostering, even without considering employee preference, is computationally very demanding. As an extremely difficult NP-complete problem, it cannot be solved by general mathematical methods (Bartholdi 1981). Adding the issue of employee preferences clearly complicates the already challenging rostering problem (Ernst et al. 2004).

There are only a few articles in the published literature that incorporate the limited time availability of employees or employee preferences to working time in rostering. The problem these studies try to solve is to roster employees to meet a predetermined staffing requirement while minimizing labour costs, subject to the constraints of employees' limited time availability or preferences.

Both heuristics for integer linear models (Glover, McMillan and Glover 1984, Loucks and Jacobs 1991, Vakharia, Selim and Husted 1992 and Brusco and Jacobs 1998) and linear programming approaches (Love and Hoey 1990, Thompson 1990, Lauer et al. 1994 and Thompson 1996) have been applied to rostering employees with limited time availability or with consideration of employee preferences. Glover, McMillan and Glover (1984) proposed an automatic schedule generation system that considers union rules, management requirements, and restricted employee availability. A heuristic approach that includes seven shift movements was designed to meet the constraints. Loucks and Jacobs (1991) developed a heuristic that assigns employees to shifts explicitly in a weekly time horizon. Their heuristic allows for the scheduling of different tasks to employees; a time period that has the fewest available qualified employees is identified and this period has the privilege of being scheduled to an employee with the specified task skill first. Vakharia, Selim and Husted (1992) developed a two-step heuristic to schedule part time workers. Some restrictions are relaxed in the first step to generate optimal schedules that meet multiple objectives such as employee work time preferences and wages, and these restrictions are reinforced in the second step by modifying the generated schedules through a heuristic by sacrificing labour cost. However, this approach does not consider employees' preferences concerning shift lengths. Brusco and Jacobs (1998) also developed a two-stage heuristic in which several best sets of starting times are obtained first, and a heuristic procedure is run to construct tours, which refers to the times of a day and the day of a week an employee is scheduled.

None of these existing rostering heuristics include breaks; namely, they assume continuous shifts. While omitting breaks can simplify the problem and improve speed, assigning breaks in real time certainly results in understaffing in certain periods, makes employees feel uncertain, and is prone to cause break skipping (Thompson 1996).

In addition to heuristics, linear programming optimization is another approach employed to solve the traditional rostering problem. Two network flow models are used in Love and Hoey's (1990) rostering model. The first model generates a set of shifts that is then assigned to the employees through the second model. This two-step approach can be problematic since the shifts generated in the first step might be infeasible because of limited employee availability (Thompson 1990). Thompson (1990, 1996) developed a series of linear programming models that use set covering approaches to match employees with limited time availability to predetermined shifts. These approaches are the only ones that specifically consider breaks, but they are limited to a daily time frame and only one break is considered for some shifts and none for the rest of them. Lauer et al. (1994) designed a linear programming model that considers part time employees' availability when shifts are assigned. It also involves two phases: optimal shifts are produced in the first phase and assigned to the employees in the second phase. This two-phase approach has the same problem as in Love and Hoey's (1990) formulation. It is difficult to guarantee that the shifts in the schedule generated in the first step can all be assigned without conflicting with employees' time availability.

Most of the current research considers only employee availability when rostering (Glover, McMillan and Glover 1984, Thompson 1990, Loucks and Jacobs 1991, Lauer et al. 1994, Thompson 1996 among others). To our knowledge, Vakharia, Selim and Husted's (1992) study is the only one that incorporates employee preferences in rostering: they use a 3-point scale to denote part-time employees' preferences as unsatisfied, acceptable, or satisfied with respect to various work starting times.

Employees' availability and preferences for working time, though related, are not equivalent to each other. Preferred working times are those that are not only available but also consistent with employees' working habits and inclinations. The preferred working times must be ones that are available to the employees; however, times available might not all be preferred. Therefore, to maximize employee satisfaction, it is the employees' preferences that need to be considered, not merely time availability.

As work starting times, break times, and shift lengths for part timers are the three basic factors that define employees' working time, failure to address any of them tends to not satisfy employees' preferred working time to the fullest extent possible. However, none of the existing algorithms incorporates all three indicators when employee preferences or availability are considered in rostering. This research attempts to solve the rostering problem while technically attending employee preference information in all three factors. The reason that these three factors are considered is discussed in greater detail in Section 5.2.2.

2.4 An Overview of the Proposed Approaches

This section presents the manner of obtaining the best schedule and roster according to multiple criteria under the two new approaches. For the new workforce scheduling method, the process involves the following:

1) Identify all possible shifts;
2) Create demand profiles;
3) Generate plausible schedules to imitate demand profile curves (only combinations of possible shifts are used);
4) Obtain performance criteria for the plausible schedules;

5) Identify the schedules on the efficient frontier and choose the best among them.

Possible shifts are characterized by having breaks and shift lengths that comply with all laws, regulations and union rules. Using only possible shifts ensures that the schedules generated do not violate those rules and regulations. Demand profiles show the trend of the demand and set the minimum number of hours that should be covered by the shifts. The details of demand profiles and how they are generated is explained in Section 3.2.1. A schedule generation method is developed to cover the demand profile with possible shifts using an opportunity positioning heuristic, so that the trend of the demand is reflected in the plausible schedules, and at the same time those plausible schedules are diversified. This method is introduced in Section 3.2.2. Chapter 4 discusses the performance measure evaluation model. An efficiency analysis tool is employed to screen the non-dominant plausible schedules or rosters; this is explained in Chapter 5.

Schedules are generated in one integrated step, which allows the user to avoid the problematic two-step nature of the traditional scheduling approach. By using simulation to evaluate the schedules, the restricting *M/M/s* assumptions of the SIPP approach can be avoided: it is possible to model blocking and reneging behaviours and use any proper service time distribution, as well as to avoid the quick transit state requirement. Furthermore, service quality can be assessed by several criteria, as previously explained. After schedules are generated and evaluated, efficient ones will be identified. An efficient schedule is such that there is no other schedule that performs better than it in all dimensions. Although every efficient schedule is a reasonable choice, which of them is in fact the optimum solution depends on the management's vision of the company. The best solution will thus be chosen through interaction with the management. By investigating the relationship between various service quality indicators and operating costs, managers are able to position their company in an efficient and quantified manner.

The complete process consists of a plausible schedule generator, a simulation model, and an efficiency analysis tool, a structure shared by the rostering process. The key differences between the rostering and scheduling methods are that employee preferences are evaluated in rostering; and that, instead of generating plausible schedules, plausible rosters are created.

Typically, workforce scheduling is concerned with generating work shifts to satisfy customer demand in a cost efficient way, and workforce rostering refers to assigning the predetermined shifts to employees with consideration to various constraints, such as the availability of employees. They are usually solved as separate problems. In this research, however, rosters are generated directly without a specific scheduling step. Our rostering process integrates both scheduling and traditional rostering – that is, shift assignment.

Most of the service facilities generate rosters manually from schedules because of the lack of proper software or the lack of faith in the functions provided by the software. This task, however, is tedious for any facilities having more than 100 employees (Gans, Koole and Mendelbaum 2003). Some service facilities practice "self-rostering" or "shift bidding" to let employees sign up for their own shifts. However, in a big company, getting a desirable shift through this process is akin to winning a lottery for many employees (Gans, Koole and Mendelbaum 2003). Silvestro and Silvestro (2000) also point out that this practice is unmanageable for a facility with over 70 employees. Since it is very common nowadays to have service facilities with more than 100 employees, a method that can generate reliable rosters that match both customer demand and employee preferences with minimal labour costs is in high demand.

The goal of this research is to create a new rostering approach that addresses cost, service quality, and employee satisfaction simultaneously, and which inherits the merits of the proposed scheduling method. All rosters are generated in a single integrated step, given arrival rate, service time, and employee preference information – but not given predetermined staffing levels. The steps to obtaining staffing levels, generating shifts, and assigning them are problematic in terms of cost efficiency and employee satisfaction (Easton and Rossin 1996, Thompson 1999, Tien and Kamiyama 1982, and Ingolfsson, Cabral, and Wu 2003). The proposed integrated approach that incorporates staffing level estimation, shift scheduling, and traditional rostering is the first one that the author is aware of in the literature.

Using a framework similar to the workforce scheduling approach, the proposed rostering method generates a large number of plausible rosters, which are evaluated to obtain performance measurements; the best, most efficient rosters are identified using a data envelope analysis (DEA)-based efficiency analysis tool. The new approach differentiates itself from others because it presents a comprehensive view of not only service quality, but also of employee satisfaction. In addition, limiting queueing assumptions are avoided. The proposed approach provides a view of tradeoff among cost, service quality, and employee satisfaction. With this information, management is able to find efficient rosters that fit into the company's strategic vision, rather than being forced to execute one optimal roster that merely minimizes cost while meeting a target service level and selected employee preferences.

Two rating scales for measuring employee satisfaction are chosen, which are explained in detail in Section 3.3. Three methods are proposed to generate plausible rosters that not only are likely to resemble the time-varying demand profile but also maximize employee satisfaction. Section 3.4 describes the rationale and procedure of the three methods.

3. First Step: Generating Plausible Schedules and Rosters

3.1 Introduction

The overview and rationale of the proposed scheduling and rostering methods were introduced in Chapter 2. In our suggested framework, the best schedule or roster is chosen according to various criteria from many plausible schedules or rosters. Therefore whether the best schedule or best roster is optimal or close to optimal depends on whether the algorithms are able, in a limited quantity and within a limited timeframe, to generate a set of schedules or rosters that contains schedules or rosters resembling the actual frontier, which is benchmarked by the SIPP method in each evaluation dimension. This task is clearly not trivial.

In this chapter, we introduce plausible schedule and roster generation methods. Section 3.2 describes the inputs and formulations of the Three-step Opportunity Positioning method for scheduling. Since maximizing employee satisfaction is an integrated part of the rostering method, we need scientific rating mechanisms to measure employee preferences, which are investigated and presented in Section 3.3. The plausible roster generation approaches are hybrids of the proposed scheduling method and three methods of maximizing employee satisfaction. These three methods are illustrated in Section 3.4.

3.2 Plausible Schedule Generation

3.2.1 Demand Profile

The plausible schedule generation method has two inputs: 1) all possible shifts; and 2) demand profiles. All possible shifts can be identified through consideration of all applicable regulations, rules, laws, and limitations on employee availability. Demand profiles are generated from customer arrival data and average service time through the following steps.

First, the business operation time is divided into smaller time intervals (for example 30 minutes or an hour, depending on the structure of the shifts). A simple measure of the offered load:

$$r = \lambda / \mu \tag{3.1}$$

, where λ is the average arrival rate, and μ is the average service rate per server, gives the average number of servers that are busy. In other words, the offered load r implies the minimum number of servers a system ever needs: the number of servers that would be sufficient to provide service with 100% utilization if inter-arrival times and service times were constant.

Depending on the variability of the arrivals and the expected service quality, the empirical utilization is usually around 65% for a quality driven operation and approaches 100% for an efficiency driven regime (see Gans, Koole and Mendelbaum 2003). Therefore, we assumed that utilization levels vary between (100-ε) % and 100% in the experiments conducted. Note that this range can be targeted to a particular operation regime by varying the parameter ε. The ratio of the offered load r and various utilization levels gives the corresponding server requirements in different levels for each time period:

$$s_t = r_t / U \qquad t = 1, 2, \ldots, N \text{ and } U = [0.5, 1), \tag{3.2}$$

where s_t is the number of servers scheduled in planning period t. $r_t = \lambda_t/\mu$ is the offered load in planning period t, U is the utilization, and N is the total number of planning periods. Several server requirements with different utilization levels for each time period are then created. Together, these server requirements form demand profiles that will act as seeds in the algorithm for generating plausible schedules.

It is intuitively clear that only schedules with staffing levels fluctuating along the proposed demand profiles would potentially perform well. The problem now becomes how to generate plausible schedules that have staffing levels close to the demand profiles, but still differ from each other. Truly random schedule generation has a central tendency: in general, it leads to schedules having a relatively small number of servers in the beginning and the end of the day, and a "bell-shaped" peak in the middle of the day. It is clear that by not biasing the random generation, a very large number of schedules need to be generated in order to guarantee a sufficient number of schedules that match the time-varying customer demand profile. One heuristic approach to accomplishing this task is introduced in the next section. Other possible approaches do exist; see for example Ozden and Ho (2003).

3.2.2 Three-Step Opportunity Positioning Method

Usually, 8-hour shifts are the hardest to assign because they typically have three breaks (two short coffee breaks and one long lunch break) and the working hours between the breaks have upper and lower limits (e.g. no more than 3 hours and no less than 1.5 hours). This characteristic makes many variations of 8-hour shifts possible. It is difficult to determine the starting times of the shifts because of their length, and even more difficult to position the breaks.

In the proposed Three-step Opportunity Positioning (TOP) method, the first step is to determine the starting times of 8-hour shifts. The mechanism in this step is similar to a greedy heuristic. The starting time of a shift is the beginning point of a span of eight hours where the aggregated staffing levels of a demand profile is the maximum; if there are ties among the aggregated staffing levels, the first span of time will be chosen. The formulation is given as follows:

$$\overline{SL}_{\tau_{Start}} = \max_{\tau}\left(\sum_{\tau+8pp-1}^{t=\tau} SL_t \right), \ \tau = 1, 2, \ldots, N - 8pp + 1, \text{ where} \qquad (3.3)$$

SL_t is the required staffing level in the demand profile at time t;

pp is the number of planning periods in an hour;

τ is all the possible starting times for the shifts;

N is the total number of planning periods in the whole operation time of service;

τ_{start} is the starting time of the chosen shift

$\overline{SL}_{\tau_{Start}}$ is the aggregated staffing levels calculated starting from time τ_{start}.

Note that this is not an optimization objective function. $\overline{SL}_{\tau_{start}}$ simply equals the maximum aggregated staffing levels taking over all groups of eight-hour consecutive time slots.

Using this straightforward technique, we can be sure that the chosen shift is used to cover a span of time, where most servers are needed for the eight hour period. However, the positioning of breaks in the shifts is not optimized, but rather a shift is randomly chosen from all possible 8-hour shifts generated in advance. The reasoning here is that the demand profile is not perfect therefore there is no need to match it perfectly and the uncovered time slots due to the breaks can be taken care of in the rest of the steps of the TOP method. This process will continue until at least all the 8-hour shifts required by the company are assigned.

Starting from step 2, only part time shifts are generated. A similar formulation is used in step 2 of the TOP method; however, the shift length is not eight but a random number between the length of the shortest part time shift and the longest part time shift. The aggregated staffing level is calculated for the time span that is the same length of the random shift.

$$\overline{SL}_{\tau_{start},\xi} = \max_{\tau}\left(\sum_{\tau+\xi pp-1}^{t=\tau} SL_t\right),\ \tau = 1, 2, \ldots, N - \xi pp + 1, \text{ where} \tag{3.4}$$

ξ is the number of hours included in the random shift;

$\overline{SL}_{\tau_{start},\xi}$ is the aggregated staffing level of the updated demand profile calculated starting from time τ_{start} for ξpp planning periods.

More importantly, in step two the break(s) in the shift are strategically positioned at the lowest staffing level point in the time range allowed for breaks in that shift. As there are definitely less breaks and usually no lunch break for part time shifts, it is not so computationally demanding to implement the break positioning strategy. This enables the "lumps" in the updated demand profile left over from the 8-hour shift breaks to be gradually removed. A simplified example is given to illustrate this step.

Assume that a shift with a length of 4 hours is randomly generated to cover a demand profile of a 7-hour business operation. The planning period is half hour long, which givens totally 14 time slots for the whole operation time. The staffing level required for each time slot is given in the following table.

Time slot number t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Demand profile	5.2	7.3	6.1	5.8	8.9	9.3	10.1	11.6	11.3	10.8	9	12.3	11.7	6.4

...

Table 3-1. Staffing level required for each time slot

First the aggregated staffing level is calculated according to the randomly generated shift length, in this case 4 hours, as shown in Table 3-2 as follows. The first number for example is the summation of the first eight staffing levels (a time span of 4 hours) listed in Table 3-1.

Aggregated Staffing Level	64.3	70.4	73.9	76.8	83.3	86.1	83.2	---	---	---	---	---	---	---

Table 3-2. The aggregated staffing level values

Note that only the first seven time slots are calculated for the aggregated staffing level as each slot covers the staffing levels of eight planning periods. Starting from slot eight there is less than eight time slots to cover afterwards. The sixth slot has the largest aggregated demand; therefore, it becomes the starting time of the 4-hour shift. As shown in the following table, a 4-hour shift is assigned to cover time slots 6 to 13. Assume that the shift has a half-hour break which is required to be located at least an hour after the beginning of the shift and at most an hour before the end of the shift, the break can then be located at time slots 8 to 11: slot 11 is chosen as it has the lowest staffing level.

Random Shift	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Updated Demand profile	5.2	7.3	6.1	5.8	8.9	**8.3**	**9.1**	**10.6**	**10.3**	**9.8**	9	**11.3**	**10.7**	6.4

Table 3-3. Time slots covered by the shift and updated demand profile

The demand profile is accordingly updated when a shift is assigned and the bold numbers shown in Table 3-3 are the ones that changed by subtracting one from the original staffing levels listed in Table 3-1. This process ends when a negative staffing level appears in any time slot of updated demand profile.

In the third step of the TOP method, the aggregate strategy used in step one and two is no longer suitable, as nonpositive staffing levels start to appear in some time slots in the updated demand profile. These time slots should be either covered by a break in a shift or not covered at all. Note however that it is our intention that at some random time slots the supply profile exceeds or goes below the demand profile. The principle of this scheduling approach is to expect that the randomness emerging in the schedule generation process would bring better performance results than simply matching the demand profile with least labour costs.

This last step involves locating the first time slot having positive staffing level, from where a condition check will be applied to determine if it should be the starting point of a shift. No matter how short a planning period is, employees' working times are usually counted by hours; therefore, the average hourly labour demand is used as a condition to determine if an employee should work for that hour. The primary condition is that the average staffing level of one-hour period is greater than one. The secondary condition is that the average staffing levels of two adjacent one-hour periods are larger than 0.5. If either condition is satisfied for the first hour or first two hours, a shift will be assigned starting from the time slot having the first positive staffing level. The shift length depends on the primary condition of future time slots. For example, if the average staffing level of the fourth hour satisfies the primary condition, then the shift is at least four hours long depending on the average staffing level of the fifth hour. When the condition is not satisfied for the first hour, a new search will start to find the second time slot that has positive staffing level. The same condition check will apply and the same method will be used to determine the shift length whenever a shift is assigned. This tactic guarantees that a shift will not be assigned for those isolated low value staffing levels in the demand profiles. The breaks of the shifts are again strategically positioned to level the staffing levels of the demand profile.

Although full timers' schedules are more or less randomly generated using the TOP method, they do not violate the union constraints, company rules, and labour law, since schedules are combinations of the shifts, which are generated with the consideration of applicable regulations.

This method can be easily modified to fit into to the situations where no breaks are included in the shifts and is also used as the shift generating component for the plausible roster generation, which is introduced in Section 3.4.

We now discuss potential ways to measure employee preferences, which is a vital part of maximizing employee satisfaction in roster generation.

3.3 Employee Preference Measurement

3.3.1 Various Factors Considered for Employee Preferences

In the rostering methods we seek to maximize employee satisfaction. Without measuring employee preferences, there is no metric by which to evaluate satisfaction. It is critical then to consider various aspects that affect employee preferences and look at possible ways to provide measurement methods that are feasible and appropriate.

There are many factors that could affect employee satisfaction, e.g. working environment, involvement in decision making, training programs, personnel relationships, compensation policies, future opportunities, and working times

(Yeung and Berman 1997, Wright et al. 2001); however, due to the scope of this research, only working time is considered as it is one of the most relevant to rostering. Work starting time, times of breaks, and shift lengths for part time employees are the factors that characterize employees' working time.

Work starting time plays a significant role in describing employees' preferences. Both time availability and working habits factor in when employees determine their daily preferred working times. Instead of assuming homogeneous preferences for every working day, various employee preferences on different days can be accommodated in the proposed approach.

Breaks are important to workers in the service industry. Service facilities are required by law to schedule one or more breaks when shift length exceeds certain hours. It is possible that employees will have preferences for their break times as well. When shifts are designed, times for breaks should at least abide by the regulations and union rules for the breaks and preferably satisfy employees' preferences at the same time. Although it might be difficult in practice to deal with the preference information for breaks, especially when rostering a large group of employees, theoretically this approach is able to consider constraints on breaks involving both regulations and preferences.

For part-time employees, shift length is a sensitive and crucial factor: some part-time employees can work only limited hours per day because of their other commitments, and some have to work at least a certain number of hours to maintain a minimum income. The travel time employees invest in going to and from work also impacts employees' expectations with respect to the number of hours they can work. It is therefore a key factor as well when considering part time employees' preferences.

When rostering employees on a weekly basis, one needs to pay special attention to employees' preferences for days off. In most cases, employees should have two days off in a week; most service facilities are open six or seven days a week and therefore not all the employees can take days off on weekends. It is then necessary to find out whether employees prefer to work on weekends and, if they do not mind working on weekends, their preferred days off.

Work starting times, break times, and lengths of shifts have a significant impact on employees' satisfaction with their schedules on a daily basis, and their days off are an additional crucial factor to consider when rostering on a weekly basis (Ernst et al. 2004). After identifying all the factors, the next step is to find accurate and efficient methods to evaluate employees' preferences with respect to these factors.

3.3.2 Rating Employee Preferences

A particular shift involves information on starting times, break times, shift length, and the day of the week that a shift occurs in; therefore, when evaluating employees' preferences for shifts, employees' views are obtained on the shift's blended effect, as contributed to simultaneously by all the above-mentioned factors. However, employees' preference as an attribute of their perception is not directly observable or quantifiable, and thus it is not obvious how it could be measured.

One alternative to modeling employees' preferences may be to use a utility function, which can potentially be estimated by collecting and analyzing a great amount of employee preference data. However, especially in a labour-intensive situation, where automatic workforce scheduling and rostering are really important, it is tedious and computationally demanding to determine the weights of the function for each employee. In addition, the service industry usually has a high turnover rate, which implies that this task has to be carried out continuously. Therefore, we are inclined to use measurement, an easier method than a utility function to characterize employee preferences. For organizations that have had a utility function developed, it can certainly be accommodated in the proposed approach.

Measurement is a tool that assigns evaluating numbers to an attribute to represent differing degrees of the attribute (Netemeyer, Bearden and Sharma 2003, DeVellis 1991, Haynes, Nelson, Blaine 1999, Nunnally and Bernstein 1994). Developing measures of subjective aspects such as employees' preferences for working time has long been a research area in social science and psychology. There are a few scaling techniques for measurement that have been developed. They can be classified into comparative and noncomparative methods.

3.3.2.1 Comparative Scales

Comparative scales provide a means to make a direct comparison of objects, in this case shifts, and their data must be able to be ranked in order (Malhotra 2004). Since employees' preferences for different shifts are comparable, which means that they have the rank-ordering property, it is possible to use a comparative scale technique to measure them. There are several of such techniques developed, such as paired comparison scaling, rank-order scaling, and constant sum scaling. See Malhotra (2004) for details.

Paired comparison scaling compares objects in pairs, with each object being compared with others in a group one by one. Binary numbers are used to record the preference of one attribute of an object over the other; the preferred one is denoted by one, and the not preferred one by zero. When data are analyzed, the object with the most ones is recognized as the most preferred. In the case of a large number of choices, the requirements of this scaling technique are large.

Rank-order scaling requests the respondents to rank all the items in order. Since it has the characteristic of forcing the respondents to discriminate among the items, and ties are not allowed, this scaling method can be demanding for the respondents when a large number of objects are presented or when it is not easy for respondents to observe the differences among some objects.

Constant sum scaling, also known as point allocation, is a process of distributing a budget of units such as certain points among a set of objects. It can discriminate objects not only from the ranking of the points but also from the gaps in between, therefore giving more information on respondents' preferences for the objects. Objects that are indifferent to the respondents can have the same points. In general, this technique requires less work for the respondents than the two previous ones.

3.3.2.2 Noncomparative Scales

In noncomparative scales, respondents are required to evaluate objects one at a time, and they do not have to discriminate among the objects. Noncomparative techniques include continuous rating scales and various direct rating scales.

In a continuous rating scale, also known as a graphic rating scale, respondents express their perceptions by making a mark on a line, with the two ends of the line representing opposite extreme situations. Using this method, respondents are generally not confined or confused by the provided categories, usually numbers, and are able to give visual and nonbiased information. However, the analysis can be tedious.

Direct rating scales provide respondents with different categories in ordered positions. There are different kinds of direct rating scales depending on the number and the range of chosen categories and their presentation. The basic idea is that respondents choose a category to represent their opinions on an object. To the respondents, doing so can be considered to be the easiest scaling method to understand and conduct, which implies that it might provide high quality measurement data.

3.3.3 Potential Measurement Scales of Employee Preferences

Among the scaling techniques that are discussed, the ones that can potentially be used in the proposed framework are the constant sum scale and the direct rating scale.

3.3.3.1 The Constant Sum Scale

When using constant sum scale, respondents are given a fixed number of points (a budget) to allocate to different categories. Points assigned to categories can range anywhere between zero and a specified fixed number as long as the sum of the points distributed to all the categories equals to the budget. Objects that receive zero points represent those that are undesired by the respondents, and objects that receive the same points are those to which the respondents are indifferent. These qualities are suitable to characterize employees' preferences for shifts as it is likely that, given all the possible shifts with different starting times, break times, and shift lengths for part time employees, some of them might not be desired and some might be equally preferred. As the data collecting process can easily be computerized, the potential problem of calculating the sums incorrectly can be avoided and the employees will not have the extra burden of keeping the budgets balanced.

When a constant sum scaling mechanism is adopted, employees who allocate more points to a particular shift than other employees will have a higher chance of obtaining this shift. If an employee does not get any preferred shifts, an arbitrary shift will be assigned. The risk of getting an arbitrary shift encourages employees to make true and strategic choices. If employee preference data are collected periodically, the fear of getting undesired shifts will motivate them to learn how to distribute their budget intelligently in order to "win" their preferred shifts. It then can be safely assumed that the employees will have incentives to quickly learn about the points allocation game, will adjust their strategies if they fail to get their preferred working shifts in the previous period, and might have different budget distribution in different weeks.

It is very important for some service facilities to be able to prioritize their senior employees' preferences when assigning work shifts. A constant sum scaling system can easily accommodate such a requirement by giving senior employees more points than lower ranked employees. With more points to be distributed to their preferred categories, senior employees have a higher possibility of receiving their preferred shifts.

A variation of constant sum scaling was applied in the airline industry several years ago, using a technique that was referred to as preferential bidding system (Gamache et al. 1998, Byrne 1988, Moore, Evans, Ngo 1978). In this case, employees used their budgets to bid on a limited number of predetermined shifts, but not all possible available shifts. Under this system, the employee who had the highest bid for a shift was certain to receive that shift unless there was a draw.

3.3.3.2 The Direct Rating Scale

A direct rating scale requires respondents to rate each object individually by giving it a numeric score, usually an integer in a given range. The ease of understanding and conducting the rating for employees and the efficiency of data administration for organizations make it another possible choice to obtain employee preference information (Malhotra 2004). Shifts that receive the lowest rating points are interpreted as not preferred, while shifts that receive the same points are considered equally preferred. Unlike the constant sum scale, there is no restriction on the sum of the scores assigned to all the shifts.

When designing a direct rating scaling system, one has to determine the range of the points respondents can give to the objects. There is a large amount of literature published on this subject but, despite its being a comparatively mature subject, finding the optimal number of scales for ratings is still an unsolved problem (Preston and Colman 2000). In attitude and opinion measures, five or seven response categories are the most used rating scales (Bearden, Netemeyer and Mobley 1993, Peter 1979, Shaw and Wright 1967). Vakharia, Selim and Husted (1992) used a three-point direct rating scale to measure part time workers' work starting time preferences in their research. Some studies have shown that five-point scales have a high reliability (Jenkins and Taber 1977, Lissitz and Green 1975, McKelvie 1978), and five-point or higher scales have similar and sufficient validity of measurement (Preston and Colman 2000). Therefore, a five-point scaling might be a good choice for measuring employees' preferences for work shifts.

3.3.4 Comparison of Different Measurement Scales

Both the constant sum scale and the direct rating scale have their own advantages and disadvantages when measuring employee preferences. There is some indication that the result from these two most commonly used measurements can be different (Doyle, Green, Bottomley 1997). Research has shown empirically and theoretically that a direct rating scale might be better than a constant sum scale (Doyle 1999, Bottomly, Doyle, Green 2000). While a constant sum scale technique imposes two tasks on employees, to make their judgments for their preferences for the shifts and to balance their budgets, a direct rating scale requires only the former. It also seems that a constant sum scale might be elastic because of the tendency of respondents to assign more scores when more scores are left and fewer when fewer scores are not used (Doyle 1999). A direct rating scaling in general might be easy to use; however, constant sum scaling has the advantage of not tying the points to given response categories and, as a result, points aggregation can be easier.

A direct rating scaling procedure measures employee preferences directly and provides a means to evaluate. However, employees might be able to deceive the evaluation system by giving false information and, because of employees' different habits and perceptions of the scores, it is possible to generate an illogical preference result such that all the shifts were assigned the same preference score.

It is comparatively easy to implement senior employees' priorities in a constant sum scaling system. In a direct rating system, seniority is difficult to represent by points and thus it has to be implemented through other arrangements such as senior employees' priority to be scheduled before others.

It is almost impossible to evaluate any subjective view such as preferences in a very accurate manner. Any method available can only facilitate an understanding of the subject. Either a constant sum scaling or a five-point scale direct rating method can potentially provide sufficient information on employee preferences for work shifts and consequently either can be used in this study.

The involvement of employees in making rostering decisions by providing preference information, and the transparency of the process through sharing the preference information among employees and making public the final rosters, potentially reduces employees' dissatisfaction with the arrangement of their schedules. Intuitively, employees who do not get their preferred shifts also tend to blame themselves, not just the scheduler, for not assigning the points strategically.

We now turn our attention to the generation of plausible rosters.

3.4 Plausible Roster Generation

In this section, an approach that automatically generates rosters that minimize labour costs and maximize service quality and employee satisfaction is introduced. The framework of this approach is as follows:

1) To generate plausible rosters;
2) To evaluate generated plausible rosters;
3) To identify the non-dominated rosters;
4) To find the roster that best matches management's needs.

Each step is elaborated in the following subsections and also in Chapters 4 and 5.

Plausible rosters are generated given employee preference information, average service time, and average arrival rate in each planning period – the equally divided quarter-hour or half-hour time intervals segmented from the total working hours.

In this process, it is intend to generate a great number of plausible rosters that are biased to both the demands of customers and employee preferences. As it is difficult to optimize a roster to meet the objectives in all three general dimensions – labour costs, service quality, and employee satisfaction – heuristics are appropriately employed to generate various rosters, each of which can potentially be the best one. An optimization algorithm can generate only one roster; however, in the case of multiple conflicting objectives, an optimum solution is extremely difficult to achieve. In addition, when a variety of potential best solutions are presented using heuristics, the tradeoffs among the objectives can be investigated. Several scholars have agreed that management is more inclined to heuristic solutions for large problems because of the flexibility exhibited in the results (Glover, McMillan and Glover 1984, Glover and McMillan 1986, Holloran and Byrn 1986, Taylor and Huxley 1989 and Bechtold and Jacobs 1990).

The roster generating process is not absolutely random since there are two goals that the plausible rosters need to achieve – to meet demand and to satisfy employees' preferences. Instead of assigning employees to a set of predetermined shifts (a schedule), a rostering heuristic consisting of two parts is applied to generate a shift and assign it to an employee simultaneously one by one so that the roster is able to meet both criteria. The first part of the heuristic bounds the generation process in such a way that the supply profile provided by the roster resembles the customer demand profile, as it is intuitively clear that the roster that meets customer demand with cost constraints usually have this characteristic. The second part of the heuristic is responsible for satisfying employees' preferences. A shift generated using the first part of the heuristic will be assigned to an employee immediately via the second part; next, another shift will be generated and then assigned to another employee. This is repeated until the number of total hours required by the demand profile is exhausted. Using this approach we do not need to produce a schedule first, instead a roster is generated directly.

The TOP method is employed in the first part of the heuristic. As was explained in Section 3.2, the inputs for the method are all possible shifts and demand profiles. All possible shifts refer to all types of shifts that are characterized by shift lengths and breaks that abide by legal regulations and, possibly, employee preferences. Demand profile indicates the number of servers needed according to average service requests in each planning period. It is generated by using a simple measure of offered load: $r = \lambda / \mu$, where λ is the average arrival rate in each planning period and μ the average service rate per server. The number of servers needed in each planning period is not definite because of the variations in the average arrival rate in the planning periods and the requirement of service quality. To ensure that the service quality desired can be offered by at least some of the plausible rosters generated, various levels of demand profiles are produced by using the ratio of offered load r and utilization U, the latter of which is in the range of 100-ε % to 100%. Plausible rosters are generated according to each demand profile. The TOP method generates shifts that provide aggregated staffing levels that vary along the demand profile. It is likely that schedules having such a characteristic potentially provide good customer service performance.

Three methods are proposed to implement the heuristic to satisfy employee preferences, which will be integrated with the TOP method. All three methods are conducted following every execution of generating a shift in the TOP heuristic except the one introduced in Section 3.4.2, which does not consider the demand requirement until a certain stage. As the rosters generated are plausible and great in number, all three preference-oriented methods, which are presented next, aim to maximize the average preferences.

3.4.1 Naïve Match

In this approach, a shift generated by the TOP method will be assigned instantly (before the second shift is generated) to an employee if he or she is available on a particular day, has not been assigned a shift yet, and is among those who give the highest preference score for this shift. Whenever there is more than one available employee with the same highest preference score for a particular shift, one of them will be chosen randomly, and the chosen employee will be marked to differentiate him or her from those who were not assigned a shift. The working hours contributed by the employee will be subtracted from the demand profile. Afterwards, another shift that needs to be filled will be generated the TOP method based on the updated demand profile and then assigned to another employee whose preference for the shift is among the highest available. The same process repeats until the updated total hours needed are equal to or less than zero.

Let s_n be the shift generated in the first step on iteration n. Let p_{ni} be the preference score employee i gives to s_n. Let a_i, a binary variable, denote the availability of employee i, where 0 stands for available and 1 stands for not available. Let P be the chosen employee's preference score. Thus

$$P = \max_i p_{ni} \tag{3.5}$$

where

$$a_i = 0.$$

Since the employee who fills in the shift generated in the first step is randomly chosen, and such rosters will be generated in bulk, it is be expected that the variety given by the randomness and mass quantity will produce a few rosters that are good in at least one dimension and possibly in all three.

3.4.2 Feed and Fill

The main objective of the feed and fill approach is to have as many employees as possible working for their favorite shifts. It starts with randomly picking employees and assigning them shifts ranked with their highest preference scores. When necessary, this method can be arranged to guarantee the best shifts for certain employees. If an employee allocates his or her highest score to several shifts, one of them will be selected randomly. Every time a shift is assigned to an employee, the employee will be marked as unavailable and the demand profile will be updated by diminishing the hours provided by the shift in each time interval. This process continues until the hours demanded in one or several time intervals in the demand profile are equal to or less than zero.

Afterward, the second part of the approach is the same as the one described in the naïve match method. A shift is generated according to the updated demand profile and then assigned to one of the remaining employees whose preference score for the shift is the highest. Both the employee availability and demand information is updated accordingly; the process stops when the hours needed reach zero or a negative number.

3.4.3 Dynamic Match

The dynamic match method, which maximizes the average preference score while introducing randomization into the process, resembles a dynamic programming algorithm. The objective function is given as:

$$Max \qquad \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}A_{ij} \tag{3.6}$$

s.t.

$$\sum_{j=1}^{m} x_{ij} \leq 1, \text{ for } \forall\, i$$

x_{ij} is binary

where

x_{ij} = 1, if jth shift is assigned to the ith employee; otherwise, $x_{ij} = 0$

A_{ij} = the points the ith employee pre-assigns to jth shift

n = total number of employees

m = total number of shifts

The constraints ensure that one person can at most work one shift on a particular day.

Although the problem looks like an ordinary optimization one, it cannot be solved by a standard method such as a set-covering approach that requires all shifts to be generated first and then assigned to employees, since the shifts produced by the TOP method need to be assigned dynamically to employees one by one while they are generated. A quasi-deterministic dynamic programming approach is thus adopted to find the optimal result.

When the first random shift is generated the TOP heuristic, each possible employee who can potentially be part of the optimal solution is listed as the starting point of candidate solutions. As one of the objectives is to maximize the minimum satisfaction, employees who have a zero preference score for the shift using a constant sum scale and a preference score equal to one using a direct rating scale are not considered as candidates. There are two reasons for this. First, if these employees are considered, the smallest satisfaction score in the roster is already the minimum, an outcome which is unlikely in the optimum solution and not desired. In addition, if they are included as candidates, then all the employees become candidates and, consequently, the subsequent calculation will become very cumbersome and computationally demanding. For the remaining iterations for each candidate, the available employees with the highest preference scores for the randomly generated shifts will be assigned to them. When a tie occurs for a shift, both or all employees will be listed as candidates until one or some of them are required in the later iterations and the remaining one will obtain the shift. If there are still residual ties at the end of the iterations, either an employee can be randomly selected or other management considerations can intervene for the final decision. The solution for each iteration from each possible starting point is not difficult to find from the results obtained from the preceding ones. Comparatively, ties are more difficult to deal with. When all the candidate rosters are generated, the one with the highest total preference score will be chosen. If there is a tie, the one with the maximum minimum preference score or minimum number of the minimum preference scores will be selected.

The formulation of the approach described is as follows. Let s stand for the list of starting candidates. Let x_{ni} be the preference score of employee i who has the highest preference score for iteration n. Let $f_n(s)$ denote the total preference score up to iteration n. Let a_i continue to represent the availability of employee i, and p_{nj} the preference score of employee j for the shift n generated in iteration n. Thus

$$f_1(s) = x_{1i} \qquad (3.7)$$

s.t.

$x_{1i} \neq 0$ (constant sum scale) or

$x_{1i} \neq 1$ (direct rating scale)

$s = 1, 2, \ldots$

and

$$f_n(s) = f_{n-1}(s) + x_{ni}$$

s.t.

$a_i = 0$

$x_{ni} = \max_j p_{nj}$

$n = 2, 3, \ldots$

$j = 1, 2, \ldots,$ total number of employees

An example is used to explain this approach. Suppose that there are four possible shifts and three employees. Each employee has submitted his or her allocation of preference score for each shift, as shown in the following table.

	Shift 1	Shift 2	Shift 3	Shift 4
Employee A	30	20	20	30
Employee B	40	0	60	0
Employee C	50	30	10	20

Table 3-4. Employee Preferences

This approach is applicable for both constant sum scaling and direct rating scaling methods. The former is used to illustrate it. In the example, each employee is given one hundred points to allocate, except Employee C, who is assumed to be a senior employee and who is given 110 points.

Shifts are generated by the target plausible schedule generator. Suppose Shift 2 is the first one generated. Since all three employees are available, there are three options for Shift 2 with corresponding preference points in step one. If Shift 1 is the second one generated, one needs to find the best solution for both shifts, given all three options in Step 1. Table 3-5 shows the results of Step 2. The preference score for each employee's assigned shift is shown in the parentheses.

If Shift 2 is assigned to	The optimal result for Shift 1 is	Total preference scores
Employee A (20)	Employee C (50)	20 + 50 = 70
Employee B (0)	Employee C (50)	0 + 50 = 50
Employee C (30)	Employee B (40)	30 + 40 = 70

Table 3-5. Step two computation

Suppose that in the third step when Shift 3 is generated, whoever is left needs to work for this shift, it is easy to locate the optimal solution among the results of the three starting conditions. The results are presented in Table 3-6. Scores shown in the first column are from the "Total preference scores" in Table 3-5.

If Shift 2, 1 assigned to	The only available employee for Shift 3 is	Total preference scores
Employee A and C (70)	Employee B (60)	70 + 60 = 130
Employee B and C (50)	Employee A (20)	50 + 20 = 70
Employee C and B (70)	Employee A (20)	70 + 20 = 90

Table 3-6. Step three computation

Using the dynamic match approach, the optimal result can be guaranteed with much less computational effort than that required to fully exhaust all possible combinations. The sequential approach perfectly suits the requirement of biasing to both demand and employee preferences in a plausible roster generation method.

One might argue that it is not necessary to use this algorithm. Whenever a shift is generated, it simply needs to be assigned to the employee who allocates the highest point to this shift. However, doing so might result in a solution that is far from optimal. In the example given above, if this approach is used, Shift 2 will be assigned to Employee C, since his or her preference point for Shift 2 – 30 is the highest among all the employees. When Shift 1 is generated, Employee C is no longer available and the shift will be assigned to Employee B. The resulting total preference score is only 90 using this naïve approach, compared to the optimal solution of 130.

The above approaches and discussions are also applicable using a direct rating scaling method. The only difference is that the range of preference points assigned to the shifts is between 1 and 5 instead of a usually wider range as in the constant sum scaling method. The pseudo-codes of these three heuristics are presented in Appendix 1.

4. Second Step: Evaluating Plausible Schedules and Rosters

4.1 Introduction

In the previous chapter, the plausible schedule and roster generation methods were presented. This chapter introduces a method for obtaining the performance measures of generated schedules and rosters. These measures estimate the values of all service quality indicators and thus are used to determine whether a given schedule or roster will be chosen. In addition, Section 4.2 discusses the strengths and weaknesses of an analytical versus a simulation model and the reasons why a simulation model has been adopted in this work. Section 4.5 describes the specifications of the developed simulation model.

4.2 Analytical Model versus Simulation Model

The purpose in the second step of the proposed scheduling or rostering approach is to find out the performance measures of the generated plausible schedules or rosters. There are two mainstream methods that are both suitable to solve the problem. It is not essential to the proposed approach which method to use as long as the performance measures that are considered can be measured correctly.

Solving business problems requires an extensive understanding and analysis of management and operation systems. To do so, mathematical models are usually developed to represent these systems in order to facilitate comprehension of the interrelationship among system components or to estimate system performance under different scenarios and therefore enable the users to solve problems and make high-quality decisions. There are two types of mathematical models, analytical and simulation, both of which are widely applied.

Analytical models use mathematical methods to seek exact solutions to problems. They can numerically calculate some performance measures such as average service level and average wait time of several particular stochastic systems. It is usually quick to run once developed and provides accurate results for compatible systems. A simulation model can simulate various systems and estimate most performance measures within certain confidence intervals. The computation time is usually longer compared to an analytical model especially if many replications are needed to validate the results.

Analytical models were used to be restrictive with unrealistic assumptions such as reaching stationary state in ignorable time, omitting reneging behaviours, and deeming service times as exponential etc. However, some progress has been made in recent years to flex some assumptions in analytical models. Ingolfsson et al. (2005) compared several approximation methods that calculate transient state probability for an $M(t)/M/s(t)$ system. Garnett, Mandelbaum, and Reiman (2002) analyzed an $M/M/N+M$ (N servers and M queue capacity) reneging model where customers' tolerance of wait time is assumed to be exponentially distributed. Whitt (2004) proposed a diffusion approximation method to estimate steady state probability to wait for a $G/GI/n/m$ queue (a system that has a general arrival process, identical and independent general service distribution, n servers, and m queue capacity). A state-of-the-art analytical model summary can be found in Mandelbaum et al. (2002), which includes algorithms with a reduced set of the assumption issues addressed. However the reasons holding us back from choosing analytical models are their limitations of relaxing only one or two assumptions and calculating only a few performance measures.

In workforce scheduling and rostering, there are many uncertainties involved in the operation systems, such as time-varying arrival rates throughout the day or the fact that customers might leave the system without being served if waiting too long. We want to demonstrate in this work that our proposed framework can be applied to general and complex systems, where transient state operations, reneging behaviours, and general service time distribution can all appear at the same time. Many businesses have a time-varying arrival rate, which requires a corresponding change of service capacity. Changing capacities requires longer time for a system to reach steady state and thus its time of being in transient state cannot be ignored. Reneging behaviour is widely documented in the literature and service times in general do not necessarily follow an exponential distribution. If systems have a combination of such characteristics exist, which we believe is realistic, it is important to show that our approaches are applicable. In addition, due to the multi-objective nature of our framework, we expect to obtain various performance measurements. However, it is non-trivial to obtain all the criteria listed in Section 5.2.1 from an analytical model. Analytical models are therefore not adopted in this research as we are not aware any of those that are capable of undertaking complex systems and generating various performance measures in current literature. Though analytical models usually require less processing time, we believe that it is beneficial to sacrifice the speed to obtain more realistic and accurate performance measures and avoid apparent nonrealistic assumptions.

Computer simulation is defined as a process of building a mathematical or logical model about a real system and experiment with the model to gain insight into the system's behaviour (Pristker 1986). A discrete event simulation model is therefore adopted to mimic the daily operations. The most important purpose of employing a simulation model is to avoid the apparent nonrealistic assumptions discussed above, which are frequently seen in analytical models. Nevertheless, the subjects of the study itself – proposed scheduling and rostering methods – do not depend on the simulation model and can easily be adapted to an analytical model once an adequately sophisticated one is developed in the future.

4.3 Computer Simulation Technique

Simulation modeling plays an important role in presenting business applications for two reasons – the increasingly complex systems emerging in the real world and the dramatic progress in computer technology. The first reason creates the need and the second provides powerful and affordable tools for conducting simulations. In the 1950's, FORTRAN or Assembly languages were used to write simulation models. Later, special simulation languages such as GPSS, SIMSCRIPT, SLAM and SIMAN were used for simulation modeling. Recently, many software packages have become available that provide drag-and-drop graphical user interfaces. The progress made in the simulation languages and software makes the learning curve of a simulation modeler moderate and the simulation model development cycle short. Advances in computer hardware have improved the speed of simulation processes tremendously in recent years. Simulation modeling therefore has proliferated into various aspects of the business world and has become a standard tool. It has been applied in industries such as health care, call centres, financial services, and manufacturing. Chung (2004) provides a detailed discussion and references on simulation languages, software, and applications.

Simulation models are categorized into three types (Law and Kelton 2000). A simulation model can be either static or dynamic. A static model represents a system at one point or one period of time such as the Monte Carlo simulation, whereas a dynamic model describes a system over the passage of time. Most simulation models are stochastic, which involves the use of random inputs; deterministic simulation models in which all the inputs and outputs are certain are rarely built. There are two types of dynamic models: continuous and discrete. In a continuous model, the state of entities changes continuously with time. In contrast, the state of entities in a discrete model changes instantly at a point in time.

In service facility systems, customers' states change with the passage of time at any given moment. For example, events such as a customer arriving, joining a queue, starting to be served, finishing being served and leaving, occur in a sequence at a particular point in time. To simulate such a system, the simulation model should be dynamic, stochastic, and discrete. Discrete event simulation (DES) imitates systems that evolve with time and have moving entities that encounter various events at distinct instants (Kelton, Sadowski and Sadowski 2001). Therefore, in this study a DES model is developed to represent a service facility in general.

4.4 Operation of Service Facility

Establishing a framework to generate workforce scheduling and rostering for typical service facilities is the purpose of this research. The service facilities are representative of services provided in various industries such as healthcare, service centres, banking, restaurants, airport check-in counters, etc. The entities involved in the systems can be tangible like people, or intangible like phone calls or requests. Despite the various kinds of services provided and the different layouts and characters of the facilities, the simulation model extracts the common operations and processes that typical service requesters go through.

The list of events involved in service facilities are the following:

1. Entities arrive in the system
2. Entities enter the system if the system capacity allows them to do so
3. Entities leave the system when its facility capacity is full (blocking)
4. Entities join the waiting line
5. Entities receive services
6. Entities might quit waiting before receiving services if the wait time is beyond their tolerance (reneging)
7. Entities leave the system after the service is finished

The arrivals of entities are assumed to have time-varying rates throughout the day. The average number of arrivals per half hour or quarter hour is used to present such variation. Entities arrive in a random manner following a Poisson process. The time-varying nature of the arrival process is seen in many businesses: fast food restaurants usually have three peaks, and hospitals one or two peaks per day. It is also one of the main reason that makes the scheduling and rostering problems challenging.

Most facilities have capacity constraints, such as the number of trunk lines in service centres and number of seats in a restaurant. Entities are able to enter the system only if the system does not reach its full capacity; otherwise, they have to be rejected, a behaviour known as blocking.

After entities enter the system, they either wait in line or receive a service immediately. If they wait in line it is possible that they may leave the system without getting served when their waiting time is longer than their tolerance threshold. In the simulation model, each entity has its own distinct patience level, which is given by a probability distribution. When entities successfully reach a server the service provider, it is assumed that the server serves only one entity at a time and that all the servers have the same capability of providing the service. Thus, the average service time is the same for all servers, although, the service times vary for different customers following a probability distribution. Customers leave the system once they are served.

4.5 Description of the Simulation Model

The simulation software Arena was used to build the simulation model. It was chosen because of its capability of handling sophisticated situations and its ability to consider a wide variety of alternative schedules. Since hundreds of schedules need to be evaluated by the simulation system, a VBA script was written to automate the schedule reading and result recording process so that they could be evaluated efficiently.

Each schedule was replicated 100 times to generate various performance measures. The rule of thumb is that the half widths range within 10% of the average of performance measurements over 90% of time for all the measurements. The accuracy of the results can be improved by increasing the number of replications, yet efficiency is as important as accuracy, and the current setting provides a run time of around 25 seconds for testing each schedule. Using 100 replications for all the experiments was determined to be sufficient for both the validity of the results and the time requirements for conducting the experiments.

A common random number technique was employed for each schedule, which implies that the same set of random numbers is generated when simulating the performance of the service system for all the schedules in an experiment. By doing so, one can be certain that the differences exhibited in performance are not by accident as a result of random number generation but as a result of the variations in the schedules.

One simulation model was developed in this study using data from a major North American utility company. The average arrival rate in each half-hour period of the total operation time of the utility company's service call centre is calculated using four-day customer arrival data collected by the company. Distribution of service times are determined by the best fit of the collected weekly data. The capacity of the service centre is limited by the total number of trunk lines the company has, and therefore any calls received when all the trunk lines are taken will be discarded and counted as blocked entities. The total number of trunk lines is 150, which is the sum of number of servers and all waiting slots (this setting is especially common in call centre industry). Reneging behaviour distribution is roughly estimated due to the lack of data. Each arrival is assigned a random patience threshold time according to the distribution, and thus it cannot receive service if its waiting time is longer than the patience threshold time. When this occurs, the arrival is discarded and counts towards the reneging rate.

The detailed information of the simulation models is embedded in the experiments conducted and is elaborated later in Section 7.2, where the empirical scheduling experiments and their results are presented.

5. Third Step: Selecting the Most Preferred Schedule or Roster

5.1 Introduction

Plausible schedules and rosters are generated using the methods introduced in Chapter 3. Chapter 4 describes how their performance measurements are obtained. This chapter presents a multi-objective analysis tool to evaluate each schedule or roster in accordance with its performance measurements. In Section 5.2, indicators related to service quality and employee satisfaction are described. The concepts of Data Envelope Analysis and Free Disposable Hull are introduced in Section 5.3. Section 5.4 shows an interactive approach to find the schedule or roster that best suits management's interests or an organization's strategy.

5.2 Factors Considered

There are more factors to be considered when choosing from generated plausible rosters than schedules, because a schedule is a set of shifts, which have not been assigned to employees yet, and thus does not involve employee satisfaction. Rosters on the other hand need to be evaluated according to not only the service quality indicators as schedules, but also employee satisfaction indicators. Both sets of indicators are introduced next. Note that this is not a comprehensive list. The paradigm presented in this work can accommodate any number or indicators, and the list of indicators can be tailored based on the particular application.

5.2.1 Service Quality Indicators

We consider eight service quality indicators to reflect various aspects of service quality.

Average wait time

Average wait time is the most natural indicator for service quality. It is the average of the wait times experienced by all served customers in simulation. A wait time average per customer is calculated for each replication and the average wait time that is used for schedule or roster evaluation is the average of the customer averages of all replications in one simulation. However, it is important to notice that average wait does not account well the outliers: in this case the customers with extremely long waits.

Average service level

Service level is the most popular quality indicator used in call centres, emergency departments, and many other service industries. One reason for its popularity is that it accounts the outliers and the tail behaviour in general better than average wait. Another reason is that it can be tailored to the needs of the particular application: It specifies the percentage of customers who are served within a wait time limit, which is set by the management. Especially in a time sensitive environment it approximates the percentage of satisfied customers (e.g. call centre) or customers who receive adequate care (e.g. emergency department).

In this research, a service level is calculated as the percentage of customers who receive service within the threshold time among the total processed customers during the entire operating hours for each replication. The average service level indicator is the average of the service levels in all the replications, and one replication in this case represents a day of operation.

Maximum wait time

Maximum wait time shows the worst scenario of wait time. In each replication, a maximum wait time among all the served customers is recorded, and the average of all the maximum wait times produced in all replications is the one used for evaluation purpose. Instead of the maximum, a high percentile such as 95^{th} or 99^{th} percentile can be used.

Average queue length

In reality, there is always limited waiting space and the existing of a long queue can visually discourage customers to join the line for services (e.g. fast food restaurants). This indicator is therefore useful for some service facilities. In this work, it refers to the average of all the average queue length calculated for each replication over the entire operation time.

Maximum queue length

Maximum queue length should never be greater than queue capacity. When this indicator is approaching the capacity of waiting space, it shows customer accessibility problems. A maximum queue length is calculated for each plausible schedule or roster using simulation, which is the average of the maximum queue lengths appeared during each entire operation time of all replications. Again, instead of the maximum, a high percentile such as 95^{th} or 99^{th} percentile can be used.

Average blocking rate

Block rate refers to the percentage of customers who are not able to join the waiting line as the queue is at its capacity. It is sometimes also called balking rate. Blocked call center customers would receive busy signals when they call, whereas blocked car washing customers would not have spaces to park their cars. A high blocking rate means either inadequate queue capacity or inadequate labour capacity. In each simulation replication, the total number of blocked customers is counted and the blocking rate of the replication is calculated. The average blocking rate of a schedule or a roster is the average of the blocking rates of all replications.

Average reneging rate

The percentage of customers who abandon the queue due to excessive long wait is defined as reneging rate, which is also referred as abandonment rate. This is the only indicator that reflects customers' tolerance to wait time. An isolated wait time does not necessarily suggest if a customer is satisfied as the same person could have various expectations to the wait time depending on the situation. Thus reneging rate is a more realistic indicator for unsatisfied customers. Similar to average blocking rate, a reneging rate is generated for each replication, and the average reneging rate of a schedule or roster is calculated by obtaining the average of the reneging rates of all of the replications.

Average employee utilization

Employee utilization shows service quality from a different angle. On one hand, high utilization suggests low labour costs; on the other hand, when it is approaching 100% in a stochastic system, it implies inadequate capacity – very busy servers and a long queue for a surge of arrivals. Should the levels of service be the same, a system with lower utilization would be preferred. In each replication, an average utilization per employee is obtained, and the average employee utilization of a schedule or roster is the average of the average utilizations of all replications.

The eight indicators introduced above are considered important service measurements that could be included in service quality evaluation. As mentioned, depending on the nature of the operations of a particular service facility, there might be more or less indicators that management is interested in. The objective is to introduce an approach that enables multiple, complimentary criteria to be processed. It is expected that this approach is able to present an extensive view of service quality and tradeoff among the indicators and labour cost to facilitate management decision making process.

5.2.2 Employee Satisfaction Indicators

Employee satisfaction indicators are developed basing on employee preference measurements – the constant sum scaling and direct rating scaling, which were introduced in Section 3.3.3. Using either of them, we understand to some extent how each employee feels about each possible work shift. When a roster is created, the idea in the proposed approach is to generate a great number of plausible rosters that meet customer demand and employee preferences as much as possible. Although maximizing all employees' preferences is one of the objectives, it is virtually impossible to satisfy every single employee's preference to the highest level with a group of employees that have various perceptions and needs while at the same time considering other constraints such as meeting customers' demand. We therefore need to find a way to measure the merit of a roster according to the satisfaction of the whole employee population through employees' individual preferences for shifts. All employees are treated equally when evaluating a group of employees' level of satisfaction. Seniority is singled out over other considerations, as mentioned in Section 3.3.

Two obvious and intuitive measurements of goodness are maximizing the overall employee satisfaction on average and maximizing the satisfaction of the least satisfied employees.

When maximizing the average overall employee satisfaction only, we have a general and neutral view of the overall satisfaction of all employees as a group; however, very dissatisfied individual employees are ignored. A roster with a substantial percentage of employees assigned the lowest rated shifts that might achieve the maximum average satisfaction may not be as good as one with a much lower portion of least satisfied employees and slightly less average satisfaction.

If the measurement of goodness is maximizing the preferences of the least satisfied employees, in order to ensure that the least number of employees are to receive undesirable shifts, the overall satisfaction may have to be sacrificed. For example, a roster with a few persons assigned to his or her lowest rated shift but with a very high average satisfaction level may be replaced by a roster with no employees in the lowest rated shift but significantly less average satisfaction. When such an objective is employed, the result may not be in the managements' interest.

It is extremely difficult to estimate how a very dissatisfied employee's performance will affect customer satisfaction compared to the performance of a mildly dissatisfied employee. If a general correlation between employee satisfaction and customer satisfaction can be obtained, we may be able to consider one measurement to be superior to the other; however, the lack of a quantified evaluation of the relationships between employee satisfaction and customer satisfaction makes this comparison impossible.

Each measurement has its own limitations. If only average satisfaction is maximized, those who receive the least preferred shifts are neglected. When we only maximize the preferences of the least satisfied employees, the overall goodness of satisfaction may not be considered. To cover both ends, the two measurements should be somehow combined. One approach is to maximize the average satisfaction while constraining the number of employees having the lowest rated shifts within a certain percentage of total employees. Another approach is to meet both objectives simultaneously. The latter is not difficult to implement in the proposed approach because of its multiple objective mechanism: both measurements and various percentiles of employees working lower rated shifts can be included as one of the objectives and used as indicators for measuring a roster's ability to maximize the goodness of total employee satisfaction.

In all the experiments conducted for this research, average preference score of the entire scheduled staff is used as the indicator for average satisfaction; whereas the lowest actual preference score in a roster and the percentage of the employees receiving shifts with the smallest possible preference score in the rating system are the two indicators applied for the least satisfied employees. The result from experiments conducted later supported our argument of the limitation of having only one goodness measurement. The two figures below show that there is little correlation between average satisfaction indicators and least satisfied employee indicators, which means that it is not sufficient to consider only one and it is possible to find such rosters that will perform well in both directions.

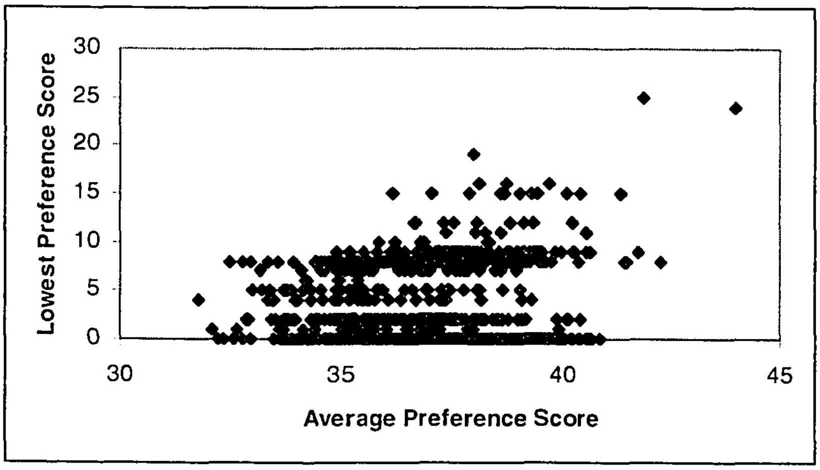


Figure 5-1. Correlation between average preference score and lowest preference score

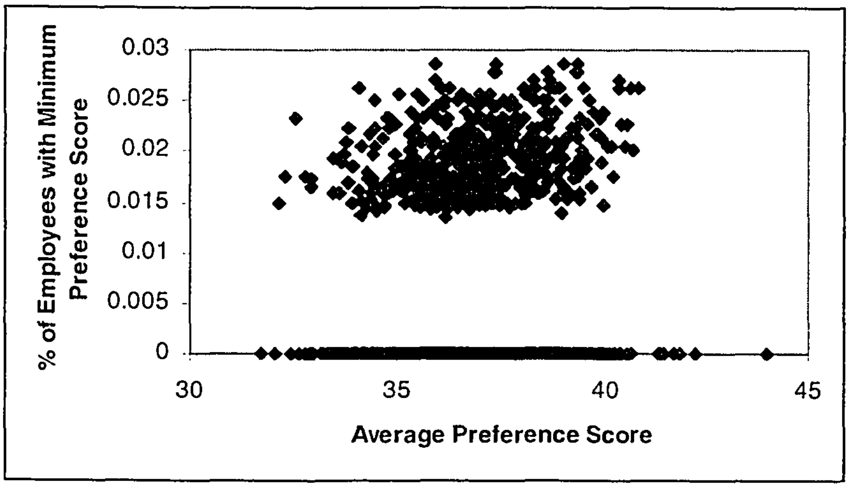


Figure 5-2. Correlation between average preference score and percent of staff receiving shifts with minimum preference score

A method that considers all the indicators of employee satisfaction and service quality simultaneously is described in the next section.

5.3 Data Envelope Analysis (DEA) and Free Disposable Hull (FDH)

Every schedule is evaluated from the perspectives of cost and service quality; every roster is examined from three aspects: cost, service quality, and employee satisfaction. Eight indicators are incorporated to represent the service quality and three are used for employee satisfaction. A method that is able to simultaneously evaluate schedules in all nine dimensions and rosters in all twelve dimensions is therefore needed.

DEA (Charnes, Cooper and Rhodes 1978) and its non-convex counterpart FDH (Easton and Rossin 1991) are efficient analysis tools that determine if a Decision Making Units (DMU) is efficient. They measure relative efficiency of comparable units using multiple inputs to produce multiple outputs. A DMU can be a company, a hospital, or a school – an entity that converts multiple inputs (such as capital and human resource) into multiple outputs (such as various products). The process of finding those efficient DMUs is to determine which ones can produce the most outputs with the least inputs. DEA and FDH's unique ability to express the relative inefficiencies as a single figure – an efficiency score – makes them appealing choices. The efficiency score can be used to identify the efficient units and measure the inefficiency for the inefficient units.

DEA and FDH models are defined here in combined orientation (see for example Joro, Korhonen and Wallenius 1998). A variable return to scale DEA model is used with the assumption that convex combinations of the units are allowed (for various DEA models see for example Cooper, Seiford and Tone 2000). Following are the detailed mathematical formulations.

Consider n DMUs having m inputs and p outputs. Let X and Y be matrices containing inputs and outputs measures for each DMU, where $X \in \Re_+^{m \times n}$ and $Y \in \Re_+^{p \times n}$, which are nonnegative elements. Assume all units in the data set are unique. We denote x_j (the jth column of X) the vector of inputs consumed by DMU$_j$ (namely, x_{ij} is the input i used by DMU$_j$); y_j (the jth column of Y) the vector of outputs produced by DMU$_j$. The problem can be formulated as a linear (in FDH case mixed integer) programming problem.

$$\max \quad Z_0 = \theta + \varepsilon(es^+ + es^-) \tag{5.1}$$

s.t.

$$\begin{aligned} & Y\lambda - \theta y_0 - s^+ = 0 \\ & X\lambda + \theta x_0 + s^- = 0 \\ & \lambda \in \Lambda \\ & \lambda, s^-, s^+ \geq 0 \\ & \varepsilon > 0 \end{aligned}$$

$$\Lambda = \begin{cases} \{\lambda \mid \lambda \in \{0,1\}\} & \text{FDH} \\ \{\lambda \mid \lambda \in \Re^n_+, e\lambda = 1\} & \text{DEA with variable returns to scale} \end{cases}$$

Where $e = [1, \ldots, 1]^T$. Z_0 is the above-mentioned efficiency score that measures how far away a DMU is from the efficient frontier. When Z_0 and all slack variables s^-, s^+ are zero, a DMU is considered efficient.

In this work, a schedule or roster is considered as a DMU; however it does not have well defined inputs and outputs. Since inputs are the ones that should be minimized and outputs are the ones that need to be maximized, those indicators that need to be minimized (such as cost and average waiting time) are treated as the inputs and those that should be maximized (such as service level and average preference score) are defined as outputs. This way the original use of DEA is altered and it is employed as a tool to identify schedules that have low cost and high service quality, and rosters with additional consideration of high employee satisfaction compare to others.

An efficient schedule or roster is the one that is not worse than any other ones in all dimensions. Here the concept of *efficiency* and *weak efficiency* are formally defined. Using the notation defined above, we consider the following set:

$$\mathrm{T} = \{ (y, x) \mid y = \mathbf{Y}\lambda, x = \mathbf{X}\lambda, \lambda \in \Lambda \} \tag{5.2}$$

Definition 1. DMU j, j = 1,2, ..., n, is *efficient* iff there does not exist another (y, x) $\in$ **T** such that $y \geq y_j$, $x \leq x_j$, and $(y, x) \neq (y_j, x_j)$.

Definition 2. DMU j, j = 1,2, ..., n, is *weakly efficient* iff there does not exist another (y, x) $\in$ **T** such that $y > y_j$, $x < x_j$.

These definitions apply to both DEA and FDH, and the difference between is the convexity assumption, which is illustrated in Figure 5-3.

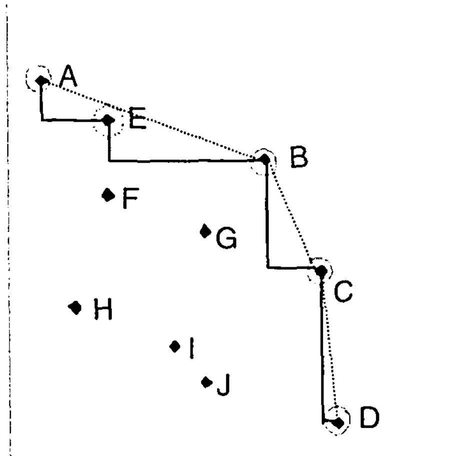


Figure 5-3. DEA and FDH Efficiency

Suppose that there are two outputs (in our case, criteria) to maximize. The gray dashed line illustrates the frontier of a DEA model and the solid grid lines show the FDH frontier. Units A to D in the graph are efficient in both DEA and FDH and units F to J are inefficient. Unit E is on the frontier of FDH as no other units dominating it. However it is dominated by a combination of unit A and B due to the convex combination assumption in DEA.

A DEA model requires a convexity assumption whereas schedules and rosters can only contain integer numbers of employees. Furthermore, the convexity assumption implies that a convex combination of two schedules or rosters will generate indicators that are the convex combination of the indicators of the two schedules or rosters. Because of the nonlinear nature of stochastic queueing systems, the convexity assumption is questionable in this case, and FDH is a more suitable model to determine the efficient schedules or rosters. In FDH the efficient frontier is discontinuous and composed by the existing schedules or rosters.

Even though the FDH model itself is computationally more demanding than DEA one due to the binary constraint, there is a benefit of using FDH. Tulkens (1993) presents a maximin algorithm that can be used to solve the FDH problem:

$$\theta = \max_{j}\left(\min_{h,i}\left(\frac{y_{hj} - y_{0j}}{y_{0j}}, \frac{x_{0j} - x_{ij}}{x_{0j}} \forall h = 1,\dots,p; i = 1,\dots,m \right) \forall j = 1,\dots,n \right). \tag{5.3}$$

Given that the problems are relatively large scale (in this work, 1050 schedules and 1050 rosters), and that we have to solve the problem for each schedule or roster, the maximin algorithm offers better computational speed than the linear programming (LP) model for solving DEA.

Finally, it is important to notice that several performance indicators can contain zero data (such as reneging rate, minimum preference score). This can be problematic in the DEA / FDH framework. However, the efficiency classification in the models discussed in this work (DEA with variable scale and FDH) is *translation invariant* (see for example Ali and Seiford, 1990 or Cooper, Seiford and Tone, 2000). This means that we can perform an affine displacement of the data (i.e. add a positive constant to values of any input or output), and efficient units remain efficient while inefficient ones remain inefficient. Note that efficiency scores of the inefficient units may change, but that is not problematic in this work since we are only seeking to identify the efficient schedules or rosters.

5.4 Using FDH to Find the Best Schedule or Roster

As an efficient DMU is one that is not worse than any other DMUs in all dimensions, it is possible to have efficient schedules or rosters that are good at only one specific aspect. To find the efficient schedule or roster that fits organization's strategy and managements' requirements, more information is needed from the organization.

In order to locate the best schedule or roster among the efficient ones, management's preferences with respect to cost, service quality, and employee preferences need to be incorporated into the analysis. In this work, the method used to incorporate these preferences is to determine *an ideal schedule or roster*. This presented approach is closely related to superefficiency models in DEA (see Anderson and Petersen, 1993) and reference point based multicriteria optimization models (see for example Figueira, Greco and Ehrgott 2005 for details).

The management is asked to express their preferences in a form of an ideal schedule or roster. The schedule or roster is expressed in terms of its performance indicators: the cost and the values of the various service quality and employee satisfaction indicators. After this, an efficient schedule or roster closest to the ideal schedule or roster is identified and presented to the management.

Figure 5-4 below illustrates in a two-dimensional space how an ideal DMU can be projected to the frontier and the best available one for management can be located.

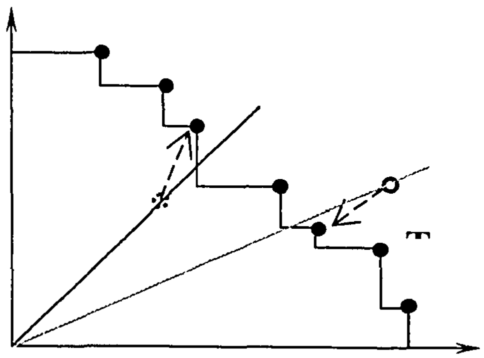


Figure 5-4. Project the ideal DMU to the efficient frontier.

A possible counter-argument for this approach is that the management is likely to present an unrealistic "pie in the sky" ideal schedule or roster. However since the management is well aware of the general tradeoffs between the labour cost, service quality and employee satisfaction, they should be able to position the ideal schedule realistically. Furthermore, it is in their best interest to be able to locate as good schedule or roster as possible, and this high-level ownership of the problem should be an adequate incentive to setting a realistic ideal schedule or roster.

The process of choosing the best schedule or roster is interactive, which means that the management can adjust an ideal schedule or roster until the most preferred one among the efficient ones is found. The additional benefit management will gain during the search process is a comprehensive understanding of the interactions and trade-offs among the indicators considered.

The interactive procedure of choosing the best schedule or roster has two advantages. First, it happens after the efficient frontier of schedules or rosters are identified. When an adjustment of ideal schedule or roster is made, management does not need to wait for the lengthy schedule or roster generation procedure rather, a new unit can be located quickly among all the generated ones. Second, there is no need to estimate utility or objective function. Though a utility function seems straightforward and has its benefits, it is complicated to generate and the interpretations of the relationship among weights are not clear. This is especially crucial when dealing with multiple correlated objectives.

Mathematically, the projection can be carried out with the same FDH formulation that was used earlier to evaluate the efficiency of the generated schedules and rosters. However now x_0 and y_0 are the criteria values of the ideal schedule and roster, and they are not included in X and Y (see Anderson and Petersen, 1993). The DMU that has non-zero λ value is the efficient schedule or roster closest to the ideal one. Since FDH's efficient hull is a combination of the free disposable hulls of the individual DMUs, the projection in FDH is a single schedule or roster unlike in DEA, where the projection is usually a combination of several DMUs.

Note that this approach enables us to project both dominated or infeasible ideal schedules or rosters.

Again, it is possible to avoid solving the FDH mixed integer LP. It is relatively straightforward to set up an interface to perform a line search between the ideal schedule or roster and origin. Once we have located from the line such a point that dominates only one of the efficient schedules or rosters, we know that the dominated schedule or roster is the ideal one's projection in the efficient frontier.

The projection seems trivial when considering a two-dimensional case as shown in Figure 5-4. However, similar graphs cannot be produced when more than three dimensions are included, in which case the proposed interactive search procedure will be the only effective approach to sort through the efficient schedules. Here the general principle of exploring the efficient frontier is to alternate the ideal point (schedule or roster) that is projected to the frontier, but use always the same projection direction (a radial projection towards to origin). An alternative method for exploring the efficient frontier is to fix a point, and project it to the efficient frontier with various weight combinations that reflect the preferences (here management's preferences). The benefits of both approaches are discussed extensively in the Multicriteria decision making literature. In general, the approach presented has some computational advantages, and there are convincing arguments that it is easier to express preferences as targets than as weights. For example Figueira, Greco and Ehrgott (2005) provided a great introduction to the current state of Multicriteria decision making literature.

6. Developing Cross-Training Strategies

6.1 Introduction

Service systems having multiple functions are common in the business world. In financial industry for instance, several examples can be found – telephone banking systems that provide customer account information, update clients' addresses, process reporting of lost or stolen cards, or handle requests to increase the credit limit; some branches in big cities in North America offer personal financial consulting in different combinations of English, French, Spanish, Mandarin, or Cantonese. Similar cases are popular in industries such as health care, fast food, police department, and call center.

While the scheduling and rostering approaches developed in this dissertation are presented in a single task system, the demand of each service provided by these systems is likely to have its unique time-varying stochastic arrival pattern and possibly unique stochastic processing time. To extend the existing approaches to solve a multi-tasking employee rostering problem, we conduct an investigation in this chapter to find out factors affecting cross-training decisions and multi-tasking employee scheduling guidelines.

Employee cross-training has been identified as an effective strategy to reduce cost and improve service quality and employee moral for systems having multiple demands for workers (Hopp, Tekin, and Van Oyen 2004, McCune 1994). To implement this strategy however presents challenges to multi-service organizations. The decision-making in cross training is critical as adjustments are required in management, administration, operation processes, and employees. This work presents managerial insights and policies of workforce rostering for systems having multiple time-varying stochastic demands using simulation. It is the first one that is aware in the literature that tackles multi-tasking problems with time-varying arrivals.

It is intuitively easy to understand the benefits of cross training from the following example. Suppose a service facility provides service I and service II, which are served by employee groups A and B exclusively. The setting is referred to as the dedicated configuration in Figure 6-1 as groups A and B are dedicated to services I and II respectively. When the arrivals of demand for services I and II are random, situations such as all workers in group A are occupied and customers of service I are waiting in line whereas one or more employees in group B are idle happen. If employees in group B are also trained for service I, they then can serve the customers in the queue I, which would shorten customer waiting time and improve employee utilization. Training employees in group A for service II or employees in group B for service I is called cross-training. A system has cross-trained employees is referred as a flexible configuration as shown in Figure 6-1, because all or some employees in either groups have the flexibility to serve both types of cusomters. Those employees are called flexible employees.

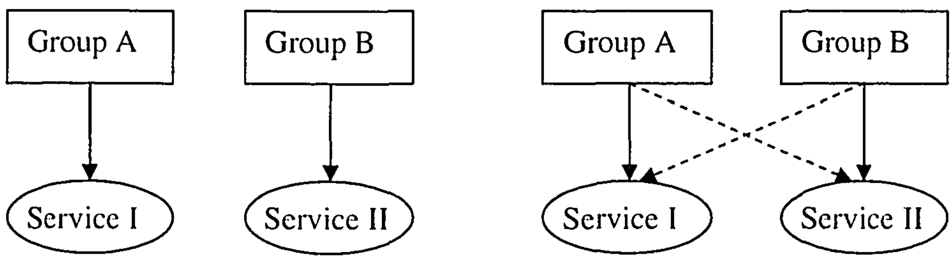


Figure 6-1. Dedicated configuration verse flexible configuration

Two benefits can easily be identified for this practice – improved service quality and increased personnel utilization. It allows the two individual services to share employee capacity; namely instead of always having integer numbers of servers for each demand, a fraction of either server capacity can be contributed to either demand practicing multi-tasking. Hopp and Van Oyen (2004) summarized that in addition to increasing productivity and responsiveness, cross training improves service quality through more knowledgeable employees and fewer customer handoffs and indirectly facilitates employees' learning, communications among co-workers, problem solving skills, retention, and ergonomic effects (less repetitive tasks, etc.). Foegen (1993) pointed out positive psychological impact to employees through cross training. Reports from nurse administration in Highland Park Hospital in Illinois (Lyons 1992), Chaparral Steel Co., 800-Flowers, CAN Insurance Companies, and Pillsbury of General Mills, Inc. (McCune 1994) all supported this view. In addition, Pfeffer (1995) identified cross utilization and cross training as one of the strategies to produce sustainable competitive advantage.

Best of all, it seems that no direct additional cost is incurred and indeed no significant extra labour or equipment investment is required implementing multi-tasking. However, training itself can be costly and time consuming, industries usually pay higher salaries to cross-trained employees to encourage participation and boost retention, and the operations of multi-tasking in a system might be threatened by efficiency (e.g. setup time of switching tasks) and quality (e.g. employee learning curve) issues. The training time of new employees for call centres averages 4.5 weeks and it takes the rookies six months to be fully productive (Stuller 1999). The daily operations of systems having cross-trained employees can be very complicated. Centring on the trade-offs between the potential benefits and additional costs and the implementation of cross-training and cross-training related scheduling and rostering problems, a lot of research has been intrigued.

Due to the complexity of the cross-training and multi-task scheduling problem, almost all the past research focused on systems with stationary stochastic arrivals and most of the results were drawn from manufacturing systems. The most investigated systems have a steady state and a uniform working time among all employees. Several important questions were explored for these systems under various assumptions.

Under what circumstances would cross training be effective?

Hottenstein and Bowman (1998) summarized from sixteen simulation studies of dual-resource constrained systems that cross-trained workers are always preferred in job shop type system. Agnihothri, Mishra, and Simmons (2003) concluded that cross training is effective when the premium rate of employee (the extra salary paid to employees cross-trained per time unit) is lower than 40% of the unit wait time cost and personnel utilization is higher than 70%. Campbell (1999) found that among seven factors considered, demand variation is the factor that makes cross-trained employees the most valuable, which indicates that cross training strategies should be applied when high demand variation presents. Similarly, high variety of service time also calls for cross-trained employees (Agnihothri, Mishra and Simmons 2003). Employee utilization is another indictor to consider as Agnihothri and Mishra (2004) found that it is not beneficial to apply cross training if employee utilization is less than 60%.

Should all employees be cross-trained or only certain percentage of them?

It was found that the contribution of multi-tasking staff diminishes while the portion of it increases; therefore, considering cost and benefit only, organizations would always be better off cross-training part of their staff crew (Fryer 1974, Park 1991, Jordan and Graves 1995, Felan and Fry 2001, Agnihothri and Mishra 2004 among others). In addition, the higher the variability of service time and the lower the training and premium cost the higher percentage of employees should be cross-trained (Agnihothri, Mishra and Simmons 2003). Complete training of all the crew is only desired when the routing of service to employees with the right skill is prone to mistakes or long travel time is expected (Agnihothri and Mishra 2004). Compbell (1999) pointed out that the percentage of employee cross-trained depends strongly on the variability of demand. Pinker and Shumsky (2000) induced employee learning curve and its impact to service quality into a stochastic service system model and found that the optimal mix of specialized or flexible workers depends on the size of the system and learning rate of the secondary skill.

To what extent should the secondary skills be trained?

Agnihothri and Mishra (2004) concluded that it is recommended in most cases to cross-train employees extensively rather than briefly. This is especially true when employee and job mismatch is prone to happen and travel time during job exchanging is not negligible. However, at even low levels of cross-training the contribution could be significant, and the rate of diminish of the marginal return of cross-training depends on the demand variation – the more the variation the larger of the contribution of the high level training (Campbell 1999). This is also consistent to the conclusion made in Hottenstein and Bowman (1998)'s review that systems can benefit from workers whose secondary skills are not as good as their major skills. The potential problem is that they are more likely to be assigned to jobs that require the most efficient skills, which results in a forgetfulness of the cross-training and potentially losing of secondary skills.

How many skills should employees impart when there are more than two types of demand?

When three job types are presented, training more employees in one secondary skill is more desirable than training less staff with two secondary skills. For same number of cross-trained employees, it is more effective to train them to the full extent in one secondary skill than at lower level in two secondary skills (Agnihothri and Mishra 2004). Hottenstein and Bowman (1998) also revealed in their review that there is little benefit from cross-training more than three skills per worker. The most recent finding is from Wallace and Whitt (2004). They suggested, after investigating various cross-training combinations of representatives in a call centre with six types of demand, that most resource-sharing benefit occurs when servers possess two skills and occasionally three skills. All three studies drew similar conclusion.

Having acknowledged the benefits of including cross-trained employees, scholars have developed analytical queueing models with multi-tasking servers. Green (1985) modeled a static stochastic queueing system with two kinds of servers to estimate the average delay time and other measures. Mandelbaum and Reiman (1998) modeled pooling of queues and servers analytically, and compared the affect of having specialized servers only, flexible servers only and mix types of servers. Zhang and Tian (2004) used the matrix analytic method to calculate the distributions of queue length and wait time of a Markov queueing system with multi-task servers.

Optimization methods and heuristics are developed to find the optimum mix of cross-trained and non-cross-trained employees. Chakravarthy and Agnihothri (2005) provided guidelines for managers to determine if cross-training is necessary. They also proposed an analytical tool to calculate the optimal mix of employees with or without multiple skills assuming exponential service time and inter-arrival time.

Even if the right mix of employees is determined, it is a complex problem to schedule a mix of dedicated and flexible employees. Eaves and Rothblum (1988) defined a flexible worker scheduling problem in a manufacturing system as a discrete dynamic program and developed a sequence of linear programs to solve it.

Previous studies provided insights on cross-training policies, models to accommodate cross-trained employees and algorithms to find the optimal mix of specialized and flexible workers. However, there is a lack of systematic investigation to the systems with time-varying arrivals. Though, presumably some findings from the past contributions might be applied to these systems, there are still many remaining questions not answered. This research is aimed to investigate the following issues in a two-demand system.

1. How do various combinations of time-varying arrival patterns impact the cross-training decision?
2. When two arrival patterns are similar, will the intensity of demand affect cross-training decision?

3. How do cross-training and multi-task scheduling decisions interact?
4. What are the guidelines of scheduling multi-tasking employee in a time-varying stochastic service system?

These questions are unique for time-varying stochastic queueing systems and have rarely been examined in the literature. The guidelines of cross-training decision-making drawn from static state queueing system might provide insights and explanations to some phenomena in the time-varying demand system but the results are not necessarily the same. This study intends to discover situations that are suitable for cross training and principles of pooling and timing for multi-tasked employee scheduling.

Previous findings in steady state systems (Agnihothri and Mishra 2004, Hottenstein and Bowman 1998, Wallace and Whitt 2004) have shown that it presents little value to have employees trained for more than two skills in most cases. Intuitively, it is also understandable that it is difficult for employees to maintain several skills at operational level simultaneously. Therefore, this research focuses on cross-training one secondary skill only, and it is believed that the principles lie in a two-job system would help understanding of multi-job systems. Even though skill based routing, a mechanism that dispatches jobs to servers based on their skills, is an important and relevant issue in a system including multi-tasking employees, it is not included in this discussion. In future research, tying skill based routing with multi-task employee scheduling is a promising trend; however, the focus of this study is on scheduling.

In the next section, system specifications and assumptions are described. Section 6.3 presents how various combinations of arrival patterns impact cross-training decisions. The policies of multi-task rostering are introduced in Section 6.4.

6.2 Problem Statement and Assumptions

The type of the queueing systems that are studied in this chapter provide two kinds of services, each of which has its own unique time-varying demand pattern, required service time, and a group of employees scheduled to meet its demand. The purpose is to improve the quality of both services by reaching higher service level or shorter wait time without increasing the total working hours by practicing employee cross-training. Two independent $M(t)/G/s(t)$ queueing systems that share servers partially or entirely with various configurations are studied to accomplish this goal.

In order to find out the impact of various combinations of different arrival patterns to multi-task scheduling decisions and determine the scheduling policy, sinusoidal artificial arrival data is used to imitate various arrival patterns in terms of state of variation in a day, and level of average arrival rate. It is assumed that service times follow Gamma probability distribution, as it is commonly used to simulate service times in service systems (Seila, Ceric and Tadikamalla 2003).

As this work is an initial attempt to unveil the affect of cross-training to time-varying multi-service systems, the intention is to keep the first study manageable and straightforward. Blocking and reneging behaviours are not included in the current model and can be considered in future research. The multiple-objective framework can also be applied in this study; however, combining the quality indicators for two queues can be a tedious practice. Since average wait time gives a direct indication of costing of waiting, it is adopted as the main service quality indicator to compare the performance of various scheduling and cross-training strategies. Average personnel utilization is also recorded to keep track of the assumption that the improvement of service is the result of reducing employee idle times.

A simulation model is in favor over an analytical model in this case for its ability to accommodate various combinations of two $M(t)/G/s(t)$ queues and schedules composed by shifts with a mixed combinations of skills. Simulation models are flexible to adapt to various scenarios and much easier to implement. The advancement in computer technology makes simulation an even more powerful tool as its major accusation, run time, is no longer a hinderer. An experiment with 100 replications in this study takes no more than 2 minutes.

Several assumptions are included to further simplify the simulation process. Due to the usually adopted data collection procedure, which is to record the service time for each customer, the difference among employees is blended and factored into the data when probability distribution is estimated. It is therefore safe to assume a homogeneous service capacity among all the employees in the simulation model.

In the service industry, it is common to observe various services that are provided not in a strict order; for example, banking services, call centre services, retail service for various merchandises. This research is concentrated on systems providing parallel services, namely there is no requirement that the two services need to be provided in a sequential order.

It is also assumed that the match between customers' requests and servers' skills is always correct, which implies a perfect dispatching system. The simulation model does not consider the setup time of switching tasks, which can be realistic in many cases such as call centre services and retail services in which the setup times are negligible.

Campbell (1999) concluded that extensive training of a secondary skill is especially appreciated in a system with high demand variation. In a system with time-varying demand, it seems more preferable to train employees to full efficiency. Therefore it is assumed that all cross-trained employees are equally good at both tasks.

6.3 The Impact of Various Combinations of Arrival Patterns

In a system that provides two kinds of services, the arrival patterns of the two demands might be different. The high volumes of both demands might appear at the same times of a day or not; the averages of the two types of arrivals in a day might be similar or very different. In this section, the impact of various combinations of arrival patterns to multi-task scheduling decision is investigated. A simulation model is developed to compare the behaviour of time-varying stochastic service systems with two types of customers and have employees that are all dedicated, partially dedicated (namely, partially flexible), or all flexible.

6.3.1 Shape

It is interesting and important to know if multi-tasking is more favorable when the peaks of arrivals of two kinds of demand happen at the same times of a day than appearing at different times. Both situations can happen in the real world.

For instance, in retail businesses the arrivals of customers coming to buy a refrigerator might be very similar to the ones coming to buy a computer. Namely, on a regular day, the variation of demand of refrigerator sales representatives should likely have similar waves of variation as computer sales staffs. As both customers would visit the store in similar time periods, it seems like they are competing for services, and thus this combination of demand patterns is called a competing pattern.

An opposite example could be a call centre that provides services to North America and Europe. Because of the time difference between the two continents, during the peak hours in North America the demand from European customers would be low and vise versa. Such a combination of demand patterns is called a complementary pattern as the aggregated demand sets off the ups and downs of each demand and becomes stable throughout the day.

Simulation experiments using artificial data are conducted to compare the behaviours of systems having competing pattern and complementary pattern. Systems with various percentages of cross-trained employees are tested. The purpose is to find out if the benefits of having cross-trained employees would be the same in a system with competing pattern arrivals compared to a system with complementary pattern arrivals.

Sinusoidal functions are employed to generate competing and complementary patterns of arrivals. In a system with competing pattern, both arrival rates λ_t follow function $\lambda_t = \lambda_{avg} A \sin(2\pi P/18 - \pi/2)$, in which λ_{avg}, the average arrival rate, is 30/hour; A, the relative amplitude, is 0.8; P, the number of peaks, is 1; and t is the time index. Figure 6-2 illustrates two time varying average arrival rates of a competing pattern. The mixed colours of black and grey on the arch are to demonstrate the two overlapping demand curves. In the complementary pattern, one arrival rate follows the same sinusoidal function as the competing pattern; the other is simply its flip, which follows function $\lambda_t = \lambda_{avg} A \sin(2\pi P/18 + \pi/2)$. The parameters of both functions in complementary pattern have the same values as the ones given in competing pattern example. The illustration of a complementary pattern is shown in Figure 6-3 as follow. Service rate is assumed to follow Gamma distribution with average 12 per hour.

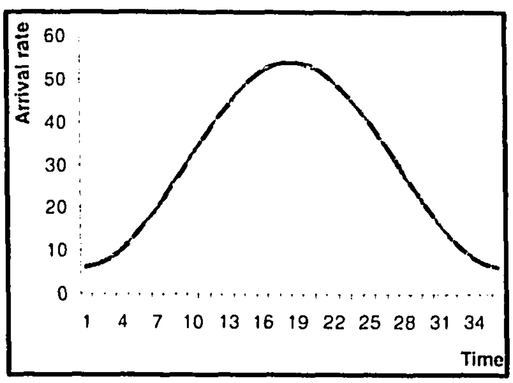


Figure 6-2. Competing pattern

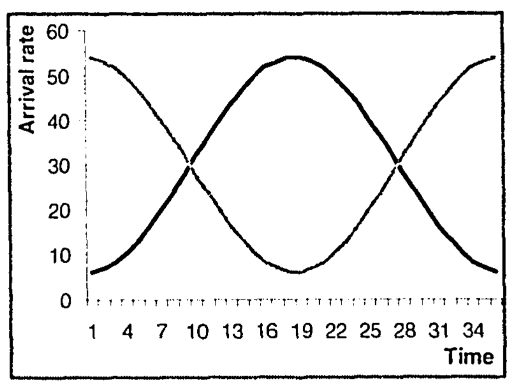


Figure 6-3. Complementary pattern

Five experiments were conducted for each arrival pattern with 0%, 25%, 50%, 75%, and 100% cross-trained employees. Each set of experiments starts with two groups of dedicated employees scheduled for each service, then approximately 25% of employees in each group were randomly picked and assigned to cross-trained group, and more employees are randomly chosen to be cross-trained until all the employees in both originally dedicated groups are trained for both services.

When employees are chosen to be cross-trained in a completely random manner, the simulation model does not necessarily give reasonable results. The half-width of average wait time in that situation does not converge even with a large number of replications. In that case, adjustments are made to the employees randomly chosen to be cross-trained to ensure that in each group (two dedicated groups for each demand and one flexible group for both demands) there are almost always employees available during the whole operating time. After such changes are made, the half-widths of average wait times were within 5% for all experiments, each of which is run 100 replications, and the results are presented in Figure 6-4. During the process of the experiments, it is found that a system with three mixed groups of employees might not be stable if employees cross-trained are not positioned strategically. Management should be aware that cross-training does not necessarily bring benefits automatically.

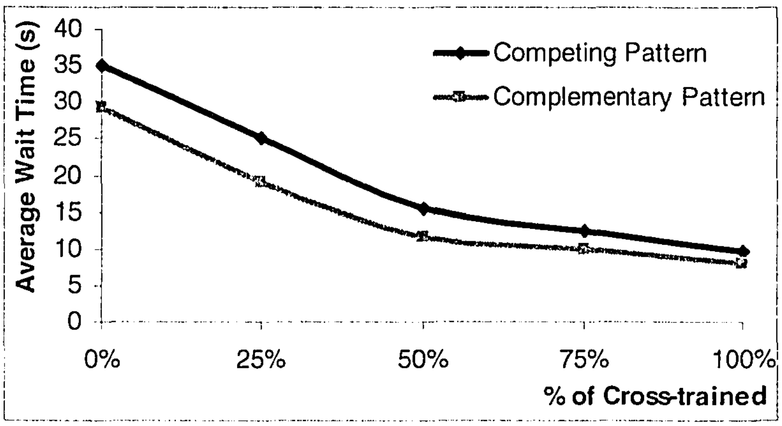


Figure 6-4. Average wait times vs. percent of cross-trained employees in systems with competing or complementary patterns

Contrary to intuition, the figure above shows that there are no significant differences between the competing pattern and the complementary pattern. Even though people would have assumed that the complementary pattern performs better with cross-training strategy, the benefit of having more than 50% of employees cross-trained diminishes even a little quicker than the competing pattern in this set of experiments. It was also more difficult to generate schedules that have stable performance in a system with a complementary demand pattern. The average personnel utilizations of both patterns are around 55%, and increasing while more employees are cross-trained. Management therefore should consider using multi-tasking strategy as long as the utilization of employees is not very high and realize that the shapes of the time-varying arrivals have little impact on the multi-tasking decision making process. However, the complementary demand pattern requires more caution when generate schedules for a partially cross-trained group. Some guidelines for scheduling employees with mixed skills are provided in Section 6.4.

6.3.2 Level

In the previous set of experiments it was found that there were no significant differences when two types of time-varying demand fluctuated in the same or the opposite directions, yet there is one more characteristics missing but worth capturing - the average arrival rates. Some facilities experience low volumes of customer arrivals of both demand (see Figure 6-5) while others have high volumes (see Figure 6-6). Experiments were conducted in this section to find out if the volumes of demand would affect the decision making in cross-training; e.g. does heavy traffic makes partial cross-training more favorable and easier to be implemented?

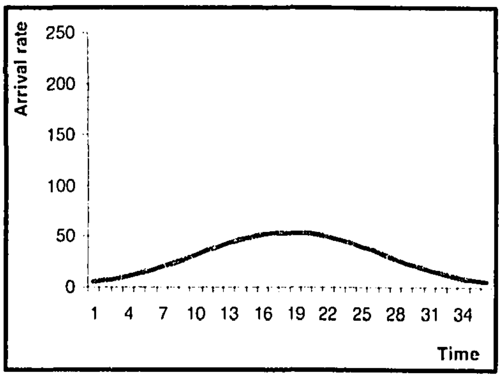


Figure 6-5. Low level demand

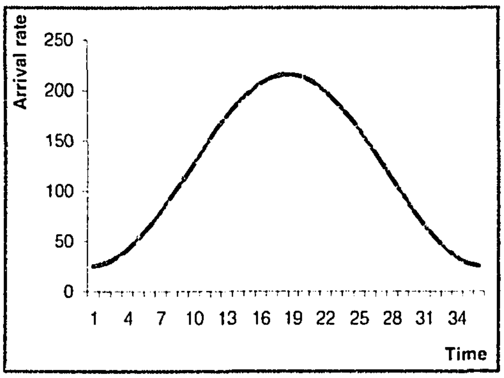


Figure 6-6. High level demand

Five experiments were carried out for each case at cross-training level 0%, 25%, 50%, 75%, and 100% respectively. It is assumed that both demands in each case had exactly the same time-varying average arrivals. The average arrival rate for the low volume case is 30 customers per hour and 120 customers per hour for the high volume one. The arrivals follow the same sinusoidal function as is shown in the previous section. Service rate for both sets of experiments are 12 customers per hour and follow a Gamma distribution. Each experiment was run for 100 replications. The results are shown in Figure 6-7.

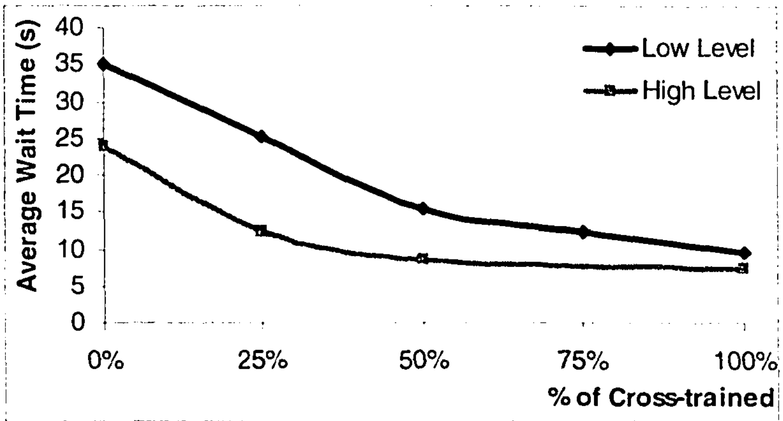


Figure 6-7. Average wait time vs. percent of cross-trained employees in systems with low or high demand

It shows that when traffic is heavy, little improvement can be achieved to cross-train more than 25% of employees. In addition, the experiment results of cross-training 75% of the employees for the high volume case is not listed as no meaningful average wait time could be generated in this experiment. Many efforts have been made to find a schedule that generates average wait time that converges to a reasonable half width but were not successful. It was also a very difficult process to randomly generate a schedule in the heavy traffic case that has a converged average wait time for the case of having 50% of cross-trained employees.

This phenomenon is not a total surprise as Mandelbaum and Reiman (1998) pointed out that cross-training has to be dealt with care in a heavy traffic system and a partially cross-trained group sometimes can make a stable system unstable. After examining an analytical model with non-time-varying stochastic arrivals, they concluded that cross-training efficiency is affected by arrival variability and cross-training is beneficial when variability is low. In a heavy traffic system, it is likely to have more variable arrivals which would cause unbalanced utilization of employees with different skill sets; consequently, the system becomes unstable.

Management should be aware of this case in that cross-training does not necessarily bring benefits. A system with a high volume of arrivals is not suitable for cross training the majority of the employees and the benefits from cross-training diminishes more quickly than in a light traffic system. Additional attention is needed in a heavy traffic system when scheduling a group of employees with mixed skills.

6.4 Policies for Rostering Cross-Trained Employees

Results from the previous section reveal that cross-training does not necessarily shorten wait time automatically. Under some conditions it is simply not appropriate to cross train employees. However in most situations cross training is desirable with careful planning done in advance. In this section, more experiments are conducted to establish the guidelines for cross-training employees and scheduling a mix of dedicated and flexible staff.

A system providing two types of services that have different demand levels but similar amplitude of change in demand will be investigated. In such a system with two different sizes of dedicated employee groups, it is difficult to determine which one should be cross-trained, these from the pool dedicated to demand with high arrival volume or vise versa. When this decision is made, the next step is also a complex process, which is to determine how to schedule the cross-trained employees so that their skills can be utilized more efficiently. Investigations about the type of employees to be cross-trained and the timing of scheduling flexible employees in a mixed congregation are presented as follow.

6.4.1 Pooling

In reality it is possible that two services in one facility have different changing volumes of customers but similar changing direction of arrivals. For example, the number of reports of lost and stolen cards might have lower volume than the requests of ordinary banking services but they both can experience the same peak time. When management tries to apply the cross-training strategy, they need to determine who should be cross-trained – the employees help report lost and stolen cards or those provide services to ordinary banking requests.

Various pooling combinations are tested in a system having two types of arrivals following sinusoidal functions; one has 30 average arrivals per hour and the other has more demand with 120 average arrivals per hour and both have 0.8 relative amplitude. Each of the two sets of experiments includes 0%, 25%, 50%, 75%, and 100% cross-trained employees among total employees from both groups whereas one starts cross training employees solely or mostly in the group with high demand and the other in group with low demand.

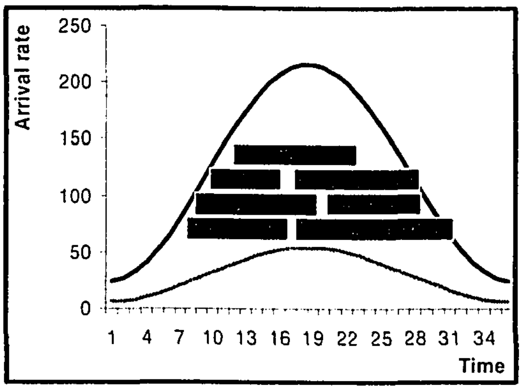


Figure 6-8. Cross train in pool with high level demand

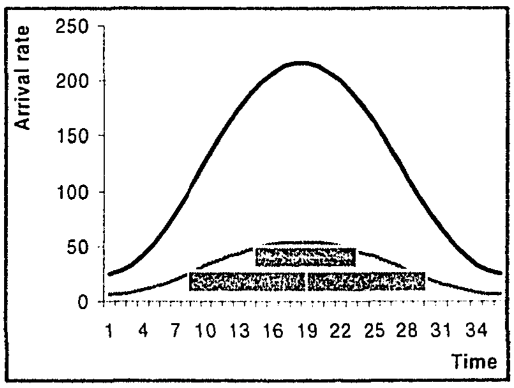


Figure 6-9. Cross train in pool with low level demand

Figures 6-8 and 6-9 show the two time-varying demand curves. The flat one has fewer arrivals than the steep one. The bars underneath the curves illustrate shifts assigned to employees. The black bars in Figure 6-8 symbolize the shifts assigned to employees who are originally dedicated to provide the service with high demand and later cross-trained. The grey bars in Figure 6-9 are hours contributed by employees used to serve the low volume demand.

By conducting the two sets of experiments, we aim to find out if there is any impact to decision making to train employees in different pools. If there is, training employees in which pool is more likely to provide better benefits? Among all the eight experiments conducted (the 0% and 100% cross-trained scenarios are the same for both sets), each was run 100 replications. When the total number of employees cross-trained exceeds the total number of employees in one particular pool for a certain experiment, they are randomly assigned to the employees in the other pool. Figure 6-10 shows the results.

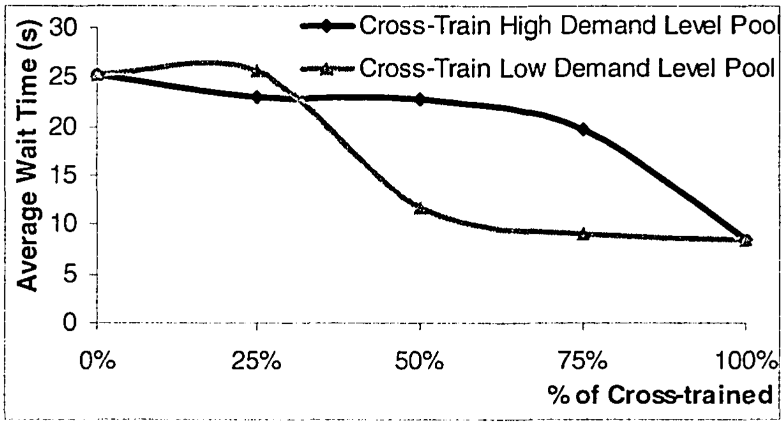


Figure 6-10. Impact of training employees in different pools

As shown in the figure above, it is obvious that training employees in different pools has dramatic impact to system performance. With 25% cross-trained employees, average wait time drops slightly if they are from the high demand pool (big pool) and increases slightly if they are from the low demand pool (small pool).

When 50% were cross-trained the average wait time plunges in the case that most cross-trained employees are from the small pool, and almost remains the same when employees in big pool were exclusively cross-trained. Due to the difference in the demand, low demand pool is significantly smaller than the high demand pool. To have 50% employee trained, cross-training all the employees in the small pool is still not enough, several are randomly chosen from the big pool to be cross-trained.

Similar magnitude of decrease in average wait time is observed when 75% of employees are cross-trained in both cases. It is intuitive that cross-training all employees from small pool would bring the most benefit. Partially cross-training employees in either pool seems to bring the least improvement to the wait time.

6.4.2 Timing

After who should be cross-trained is decided, it is time to consider when they should be scheduled. There are periods when two types of demand having similar average arrivals (see Figure 6-11) and periods where there is a huge gap between average arrivals of the two types of demand (as shown in Figure 6-12). Does the timing of scheduling have impact to the system performance? At which periods should cross-trained employees be scheduled to maximize their utilization?

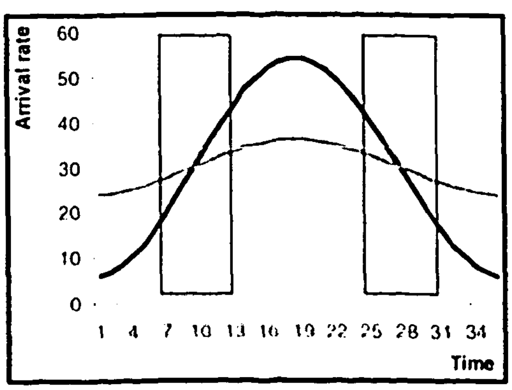


Figure 6-11. Schedule cross-trained workers at times with similar demands

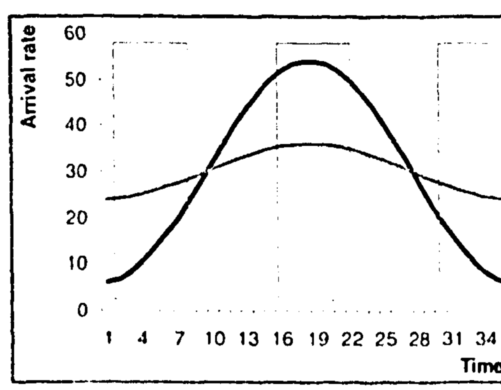


Figure 6-12. Schedule cross-trained workers at times with diverse demands

In one case, cross-trained employees are mainly scheduled to work during the times that the two demands have similar volumes, which are shown visually as the shaded areas in Figure 6-11. Whereas in the other case, cross-trained employees would cover mostly the three shaded areas in Figure 6-12 where there is a big gap between the arrival volumes of two demands. Eight experiments are conducted since the performance results are the same for both cases when 0% and 100%

employees are cross-trained. Shifts that cover at least the shaded areas in both cases are randomly chosen to be assigned to cross-trained employees. Each experiment was run 100 replications and the 95% confidence intervals of average wait times are always within ±10% of the average. The impact of cross-training employees in the cases of different timing is shown below in Figure 6-13.

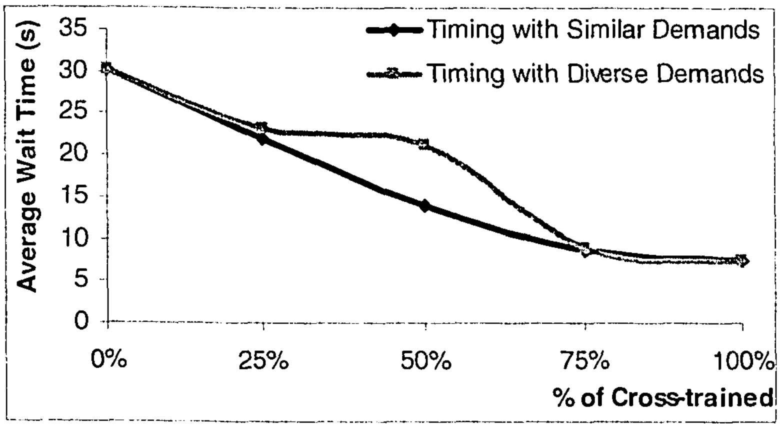


Figure 6-13. Impact of scheduling cross-trained employee at different periods

The results for both cases are very similar with the only discrepancy at 50% cross-trained point. It is easy to understand the reason that the results are almost the same at 75% cross-trained point as it is inevitable to have multi-skilled employees working for both periods. All 25% cross-trained employees in the diverse demand case were assigned to work during the shaded area in the middle in Figure 6-12. This arrangement brings average wait time down significantly but not as much as assign them to the two shaded areas in Figure 6-11. When the same numbers of cross-trained employees are added to the system to the two shaded areas on both sides in Figure 6-12, there is almost no improvement (the half width in this case is 4.9% of average). We can conclude from the experiment results that management should consider scheduling cross-trained employees at periods where both types of demand are similar.

6.5 Conclusion and Future Research

This chapter examines how different shapes and levels of time-varying arrivals would impact cross-training decision making. It is discovered that various combinations of shapes have little impact yet different demand levels have significant impact to cross-training strategies. Cross-training is a good tactic to be implemented in a light traffic system. In a heavy traffic system, it is not recommended to cross-train the majority employees, as this might easily trigger an unstable system. Section 6.3.2 provides detailed explanation.

Further investigations are conducted to establish guidelines for cross-training and scheduling in terms of pooling and timing. It is found that in a system with non-symmetric demand, cross training employees in big pool exclusively has limited benefits. It is always better to train at least all employees in the small pool and some employees in the high demand pool. Timing is a very important factor to consider when scheduling cross-trained employees in a system with time-varying arrivals. When cross-trained employees are scheduled at periods where the average volumes of the two types of demand are similar, the systems generate better performance results.

This research brings insights to cross-training strategies in a time-varying service system with two types of demand, which is an important issue that has not been clearly addressed in the literature. It characterized the systems into systematic categories and unveiled the factors that do and do not affect the cross-training process. It also provides guidelines for cross-training scheduling, with which our proposed rostering approach can be altered to accommodate multi-tasking systems, further descriptions can be found in Section 8.1.

Additional studies can be conducted to focus on the savings of labour hours while maintaining the same level of service. Skill based routing can be incorporated into the research as a condition dependent factor and strategy.

7. Applications

7.1 Introduction

To find out the performance of the proposed workforce scheduling and rostering approaches, we experimented with real problems from industry in this chapter. Section 7.2 applies the scheduling approach to generate schedules for a call centre of a major North American utility company. In Section 7.3, the flexibility of the employee preference biasing heuristics in the rostering method are tested by 16 $M(t)/G/s(t)$ queueing problems, and one of the heuristics is employed in Section 7.4 to solve a rostering problem using empirical data from the same company as in Section 7.2.

7.2 A Scheduling Experiment Using Empirical Data

In this section, the proposed approach is applied to empirical data and the results are compared to a schedule adopted by the company and to optimized SIPP results. Schedules generated from the proposed approach and the SIPP method are both tested by the same simulation model with common random numbers to ensure a fair comparison of the results. The data is from the call centre of a major North American utility company and consists of call arrivals, service times, reneging times, and regulations on legitimate shifts. In this call centre, calls coming in are answered directly by representatives or, if all the representatives are busy, are put into a wait queue. Any representative is able to answer all the questions and the unit labour cost is the same for both full- and part-time employees. Figure 7-1 shows actual arrival rates on a typical day. Call volumes are low at the beginning and the end of the day, and the peak appears at around 11:00 a.m. The arrival rates used in experiments are the average of one week's data. Service times are combinations of talk time and after-talk time in seconds, which is found to follow an Erlang distribution with parameters $m = 5$, $B = 54.4$ and shift = 10 (mean = 282.5 and standard deviation = 121.9). Reneging behaviour due to customer impatience is a rough estimation from the wait times (in seconds) of reneged customers before they quit the call (see Table 7-1). There are a total of 3671 possible shifts generated for both full-time and part-time employees according to the company's labour regulations.

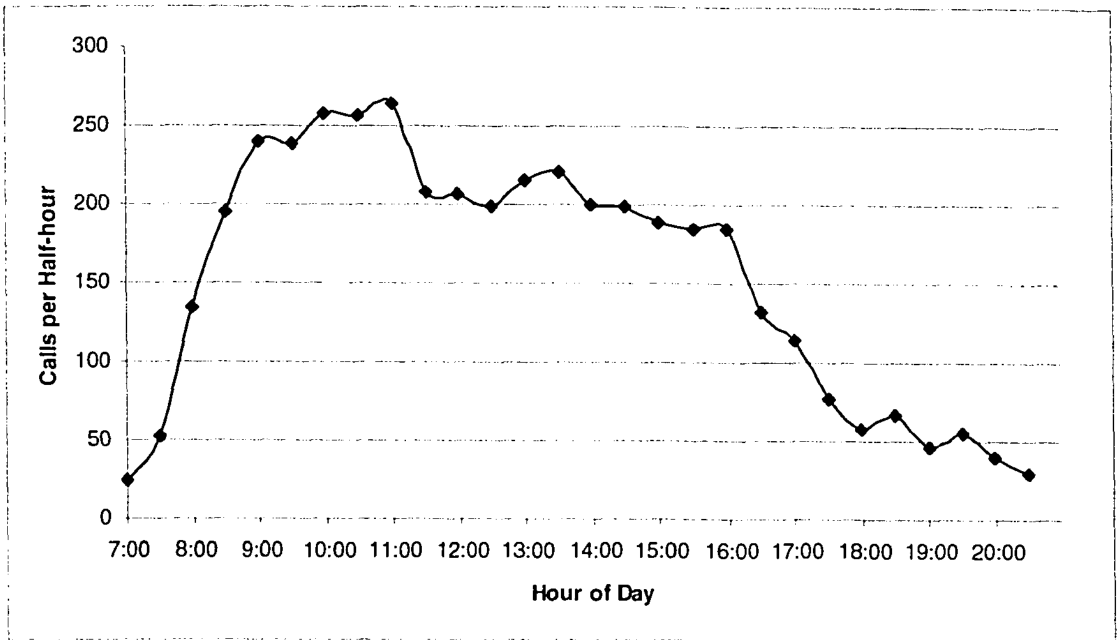


Figure 7-1. Call arrival rate on a typical day

Wait time uniformly ranges		Probability to Leave
From (s)	**To (s)**	**without Being Served**
0	10	1.5%
10	20	1.0%
20	30	1.0%
30	40	1.2%
40	50	1.0%
50	60	1.0%
60	120	3.0%
120	200	10.0%
200	500	80.3%

Table 7-1. Reneging probability distribution

According to the proposed approach, the offered load r is first calculated for each time slot given corresponding average arrival rates and average service rates. We estimate, by varying the parameter ε, a wide range of utilization from 38% to 96%

in steps of 2%, therefore 30 demand profiles in total are generated. Using the TOP algorithm, 35 plausible schedules are generated for each demand profile, which yields 1050 plausible schedules in total.

Experiments are conducted to determine the number of plausible schedules needed in order to include most of the possible outcomes. Figure 7-2 shows the difference between generating 35 rosters per demand profile (in total 1050 plausible schedules) versus generating 100 rosters per demand profile (in total 3000 plausible schedules) visually in two dimensional spaces (Hours paid vs. Average service level). The scatter points from 1050 plausible schedules are displayed in dark colour, which are plotted on top of the data from 3000 plausible schedules in gray colour. It is clear that the data from both clusters provide almost the same amount of information, whereas the time required for running the simulation experiments for 1050 schedules is only one third of the time needed for 3000 ones. Thus, we determine to generate 35 plausible schedules per demand profile.

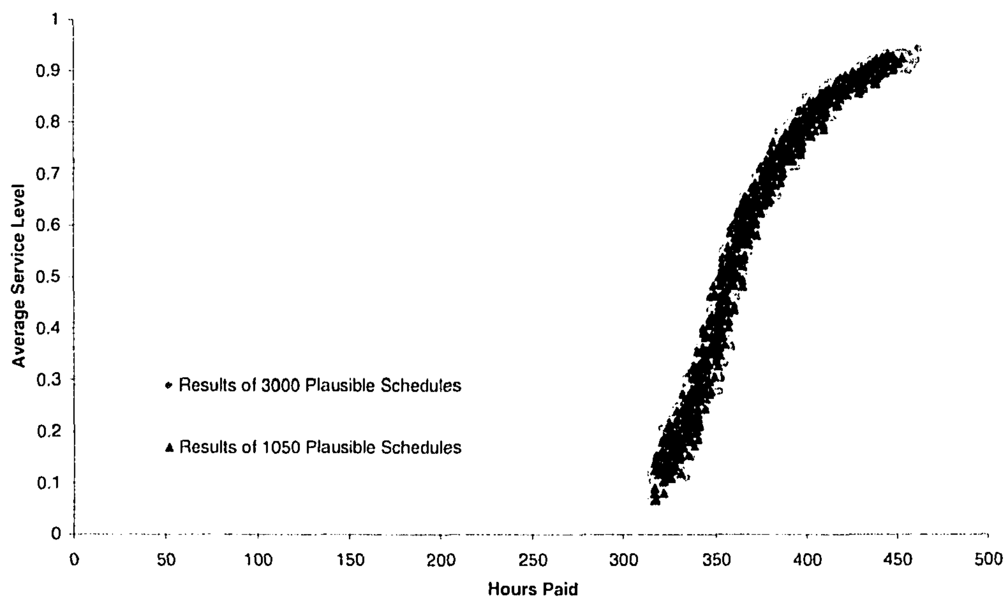


Figure 7-2. Justification for number of plausible rosters needed per demand profile

When plausible schedules are generated, shifts are first assigned to at least 25, at most 40 full time employees according to the demand profile. After this, shifts are assigned to part time employees until the remaining staffing levels in the demand profile are mostly nonpositive (for detail, refer to Section 3.2). Once all 1050 of plausible schedules are generated, a simulation model is used to obtain their performance measures. Each schedule is run for 100 replications, the run-time for one set of 100 replications being less than 25 seconds. The plausible schedules and their staffing levels s_t were generated in Excel by VBA code. A simulation model built in Arena (Rockwell Software Inc., 2000) was called by Excel to simulate each problem while all the parameters were read into Arena from Excel. Reasonable confidence intervals are achieved with 100 replications and the performance criteria generated by the simulation model were written and stored in Excel.

The outputs of the simulation model collected as indicators of service quality are:

- Average waiting time (average time customers spend in queue for service)
- Maximum waiting time (average of the longest time a customer spends in queue for service for each replication)
- Average queue length (average number of customers in queue for service)
- Maximum queue length (average of the largest number of customers in queue for service for each replication)
- Reneging rate (percentage of customers who abandon queue prior to service)
- Blocking rate (percentage of customers who attempt and are not able to join queue)
- Service level (percentage of customers served within a predetermined time interval – threshold time)
- Personnel utilization (percentage of customer representatives' busy time over total work time).

We note that instead of maximums, upper tail measures such as the 95th percentiles could be used as well for the worst case analysis.

We realize that some indicators are highly correlated for example average wait time and average queue length, see Figure 7-3, yet some are not correlated such as maximum wait time and maximum queue length as shown in Figure 7-4.

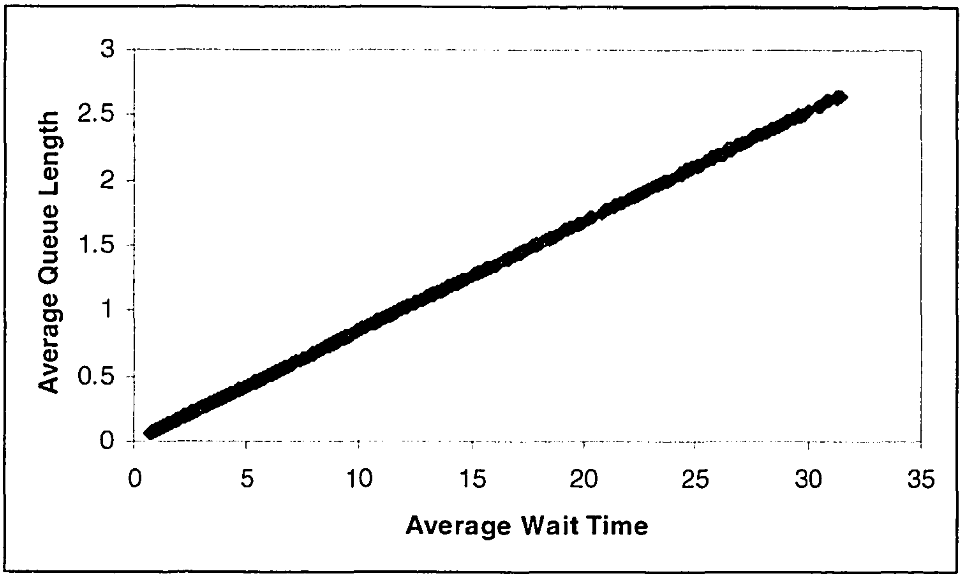


Figure 7-3. Correlation between average wait time and average queue length

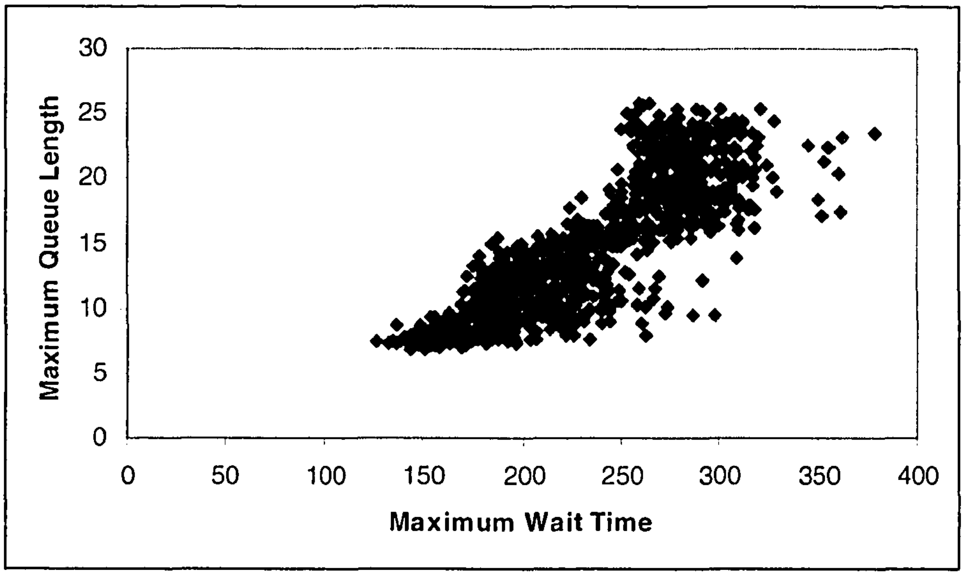


Figure 7-4. Correlation between maximum wait time and maximum queue length

In general, the goal was to control the confidence intervals (CIs) so that all confidence intervals are within +/- 10%. This target was met by almost all the schedules for all quality indicators except average reneging rate, with which 90% of schedules met. The number of schedules that has CIs not within +/- 10% and the largest CIs among all the plausible schedules for each indicator are presented in Table 7-2.

Indicators	Number of Schedules NOT within Target CI	Largest CI Encounters among All Schedules
Average waiting time	4	11.62%
Maximum waiting time	0	8.04%
Average queue length	7	12.50%
Maximum queue length	0	7.39%
Average Reneging rate	104	14.21%
Average Blocking rate	0	0%
Average service level	0	6.38%
Average personnel utilization	0	0.45%

Table 7-2. CI information of all the quality indicators evaluated for plausible schedules

Note that although there is more variation in Reneging rate than desired, the actual CIs are short in absolute terms: the Reneging rate with longest CI is 0.08% +/- 0.01137%.

The efficiency analysis yields altogether 813 efficient schedules. Again, in order to compare the results to the optimized SIPP ones, the 1050 schedules from biased random sampling are compared to the optimized SIPP results generated with various target service levels as well as to the schedule adopted by the company in all service quality indicators. To facilitate the comparison, all indicators are determined with the simulation model. This is illustrated in Figures 7-5 to 7-11.

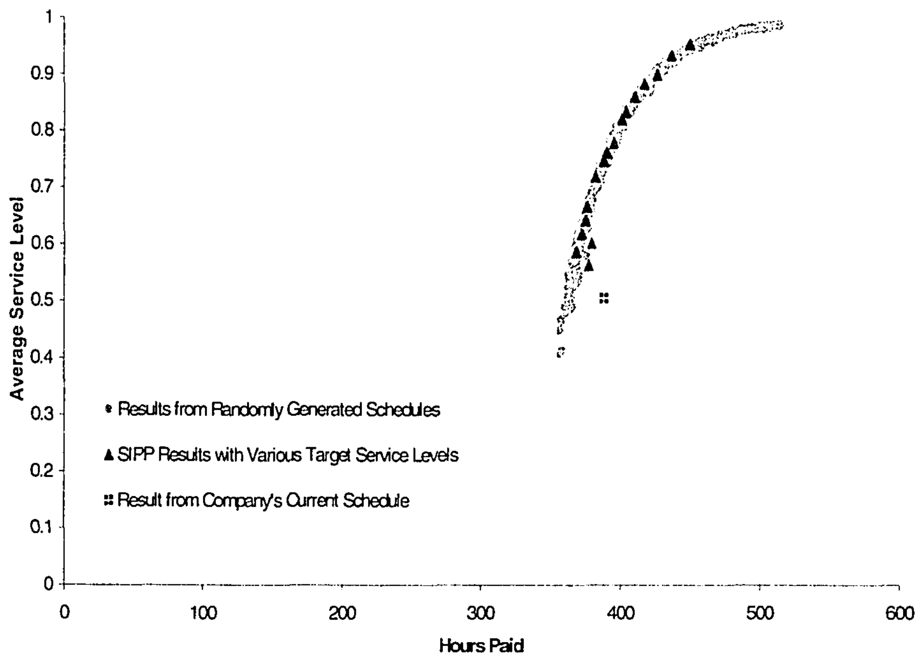


Figure 7-5. Average Service Level vs. Cost

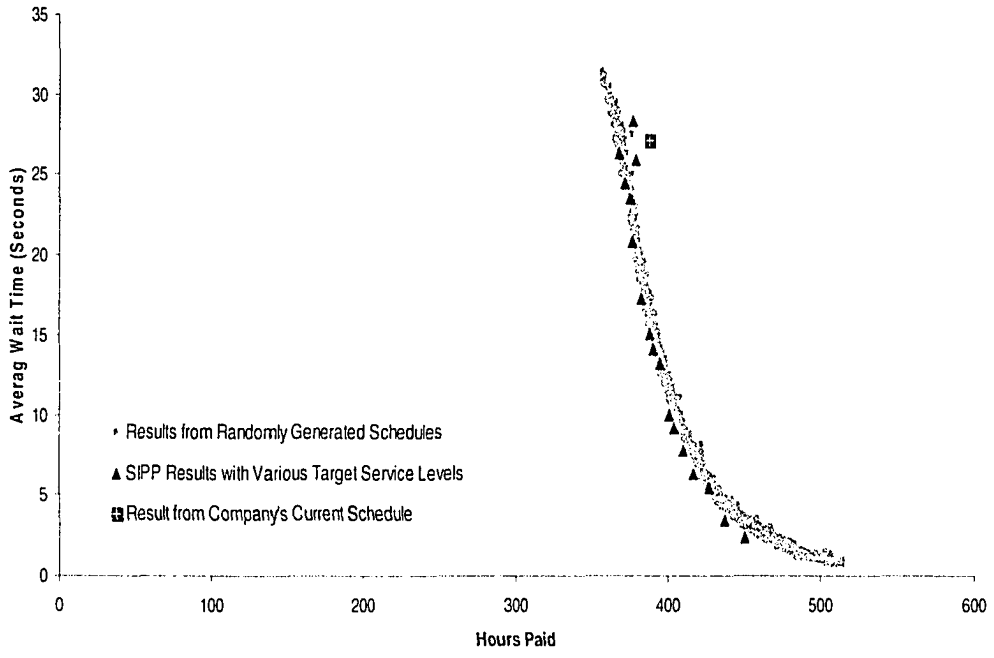


Figure 7-6. Average Wait Time vs. Cost

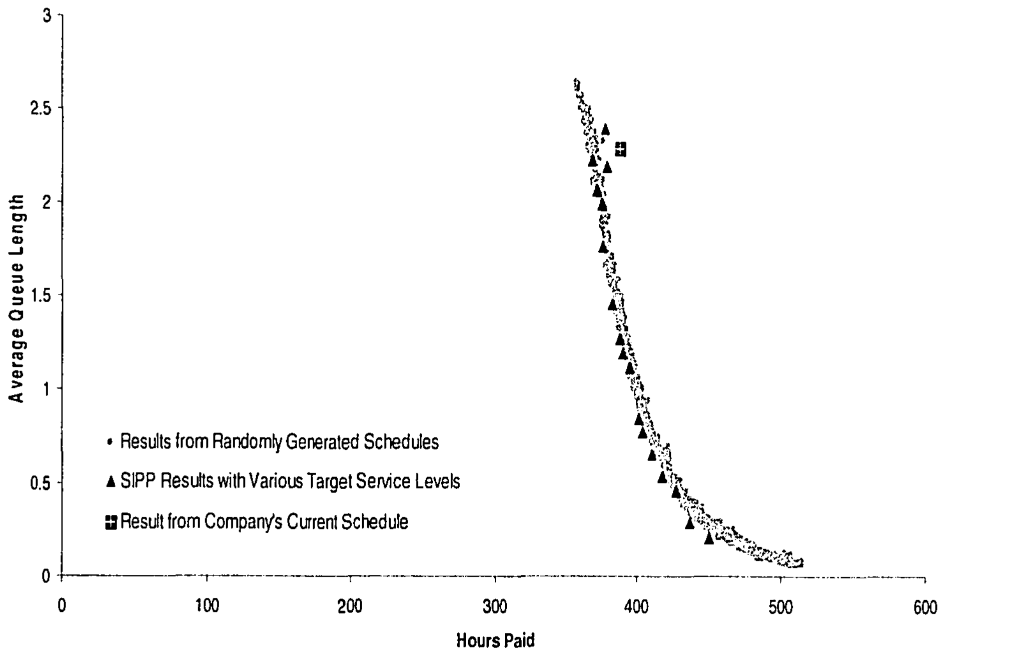


Figure 7-7. Average Queue Length vs. Cost

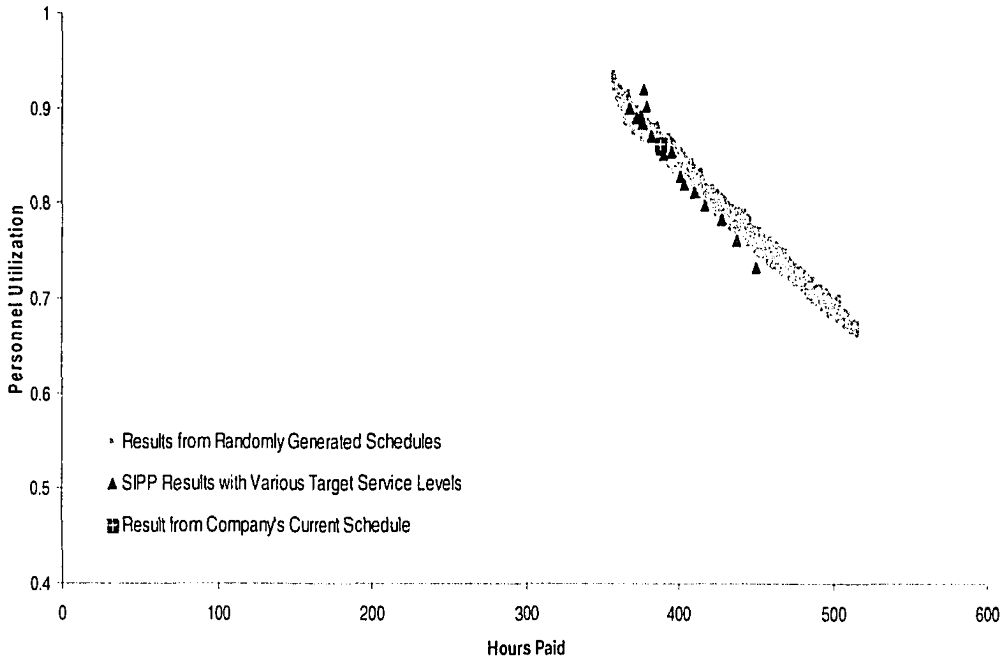


Figure 7-8. Average Utilization vs. Cost

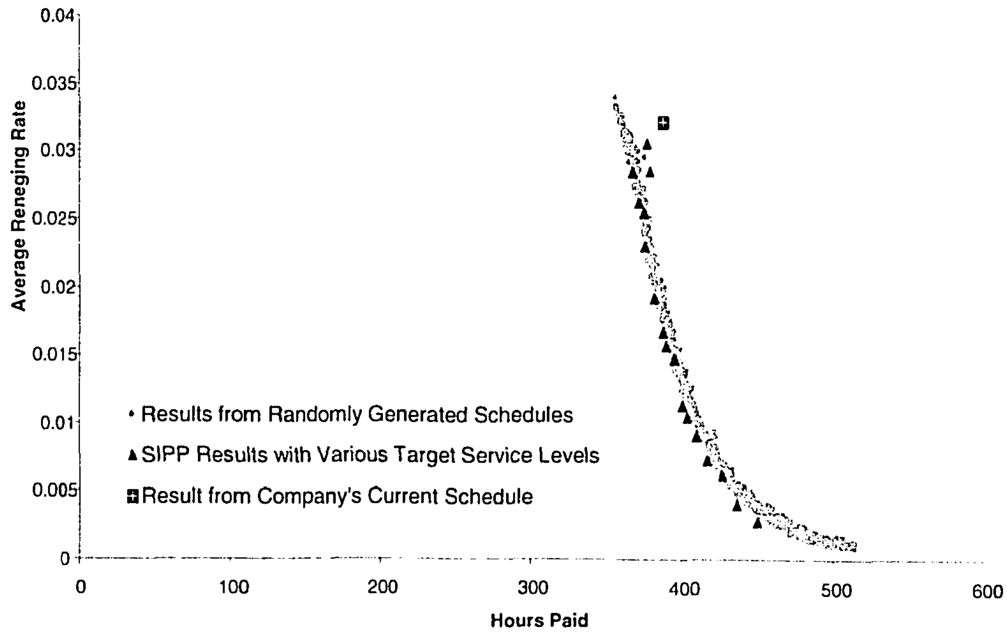


Figure 7-9. Average Reneging Rate vs. Cost

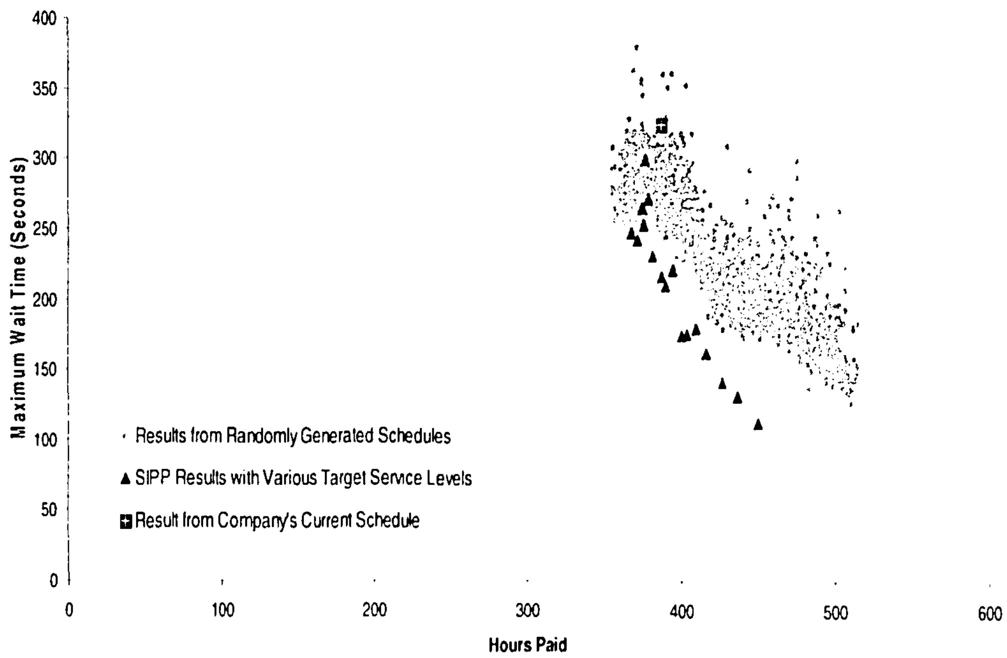


Figure 7-10. Maximum Wait Time vs. Cost

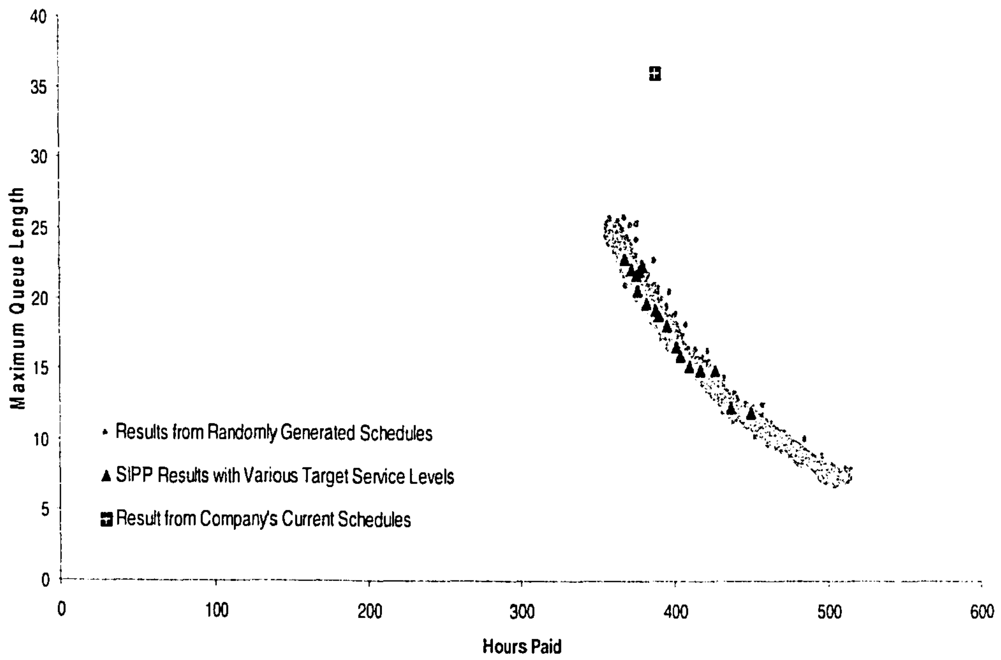


Figure 7-11. Maximum Queue Length vs. Cost

Note that there is highly nonlinear relationship between the cost of labour (number of hours) and average service level. Although in the two-dimensional case it would be possible to estimate the frontier by solving several SIPP problems and varying the cost level, it is critical to remember that once we introduce additional dimensions this approach loses its appeal as the number of optimization problems required to generate a reasonable representation of the frontier will grow exponentially.

Although there is correlation between the criteria (say, for example, service level, average wait, and queue length), including multiple criteria provides additional insight. Table 7-3 lists pairs of schedules that are almost identical in service level and cost (number of hours). In all pairs, there is difference in average wait time and, for example, in pairs 771/776 and 922/933 the service level and average wait give slightly contradictory information on which schedule would be the best.

Schedule #	Average Wait (seconds)	Service Level	Number of Hours Paid
190	21.429	68.61%	378
201	19.714	68.76%	378
407	12.19	81.86%	401
357	10.933	81.88%	401
771	2.871	95.73%	459
776	2.877	95.79%	459
836	2.645	96.14%	466
827	2.463	96.32%	466
851	1.955	97.22%	476
854	1.654	97.28%	476
922	1.164	98.09%	492
933	1.285	98.11%	492
1012	1.296	98.2%	495
975	1.075	98.3%	495

Table 7-3. Average wait time, service level, and number of hours paid

After all the efficient schedules are identified, the next step is to select the best one among them. As the graphs indicate, the efficient schedules differ from each other in their cost and service quality characteristics.

In order to select the final schedule among the efficient ones, interaction with the management is needed. Table 7-4 shows some possible target schedules and the corresponding closest efficient schedules. By trying various ideal schedules, the decision maker is able to locate various schedules that match his or her needs.

Schedule	Avg. Renege %	Avg. Wait	Max Wait	Avg. Q. length	Max Q. length	Service Level %	Util.	Hours
Target #1	0.10%	1.00	140.00	0.08	7.5	98.50%	0.68	500
Sched. #995	0.10%	0.921	138.75	0.078	7.41	98.60%	0.687	500
Target #2	1.32%	12	420	1.5	21	65.00%	0.75	450
Sched. #404	1.18%	10.182	271.75	0.86	15.99	83.71%	0.84	403
Target #3	1.75%	13	470	1.75	27	75.00%	0.9	500
Sched. #368	1.36%	11.934	273.09	1.007	16.88	81.14%	0.857	398

Table 7-4. Targets and corresponding schedules

This way, the approach proposed in this thesis gives the management a broader understanding of the service quality with several criteria as well as additional insight on the interaction between the cost and that quality.

7.3 Rostering Experiments Using Sinusoidal Arrival Data

To find out the performance of the rostering heuristics introduced in Chapter 3, they are tested with artificial arrival, service time, system capacity, customer reneging behaviour, and employee preference data. Sixteen scenarios combining a high and low value of service rate and threshold time of service level and various shapes of arrival processes were created to test the flexibility of the proposed approaches to maximize employee satisfaction. The heuristics that are unique in the rostering method yet are irrelevant to the scheduling results are applied and the outcomes concerning employee satisfaction are presented in this section.

Arrivals are time varying and follow the Poison process. The average arrival rate changes according to a sinusoidal function: $\lambda_t = \lambda_{avg}(1 + A\sin(2\pi P/18 - \pi/2)$.

Each element is explained as follow.

λ_{avg} – the average arrival rate of the day;

A – the relative amplitude. A large number means that arrival rate changes dramatically in a day;

P – the number of peaks. It is assumed that arrivals are lightest at the beginning of the day, increase steadily to a peak, and then decrease. Number of peaks indicates the times that arrivals increase and decrease.

By changing those parameters, the proposed employee preference biasing heuristics could be tested with arrivals changing by level, variation, and frequency of variation. It was assumed that there are 18 working hours per day. This setting implies that the queue system begins with an empty system every day. The "$-\pi/2$" expression manipulates the sinusoidal function to generate light arrivals at the beginning and end of the day and peaks in the middle of the day, which is in accordance with the common sense of real problems.

The other model parameters of the test problems are average service rate μ and threshold time of service level τ. Average service time is assumed to be homogeneous among different servers and throughout the day. Service time is set to follow a Gamma distribution with a low average service rate μ 3/hour verse the high 12/hour. The threshold time of service level is another parameter that might have impact to the proposed approach. It is determined to be either 20 seconds or 300 seconds. Table 7-5 lists the high and low values of the five factors that are considered having the most impact on the performance of an $M(t)/M/s(t)$ queueing system.

Parameters	Low Level	High Level
Arrival process		
λ_{avg} = Average arrival rate	30 / hour	120 / hour
P = number of peaks per day	1	3
A = Relative amplitude	0.2	0.8
Service process		
μ = Average service rate	3 / hour	12 / hour
Service level parameter		
τ = Threshold time	20 seconds	300 seconds

Table 7-5. The high and low levels of parameters used for sinusoidal experiments

Other parameters such as planning period and operation time of business were not considered as changing variables in this experiment. We understand that the changes in those parameters might lead to more sophisticated conclusions, at this stage of research we choose to keep the experiments concise. Following a fractional factorial 2^{5-1} experimental design, 16 test problems are included in total and their specifications are given in Table 7-6.

Problems	λ_{avg} (/hour)	P	A	μ_{avg} (/hour)	τ(seconds)
1	30	1	0.2	3	20
2	30	3	0.2	3	20
3	30	1	0.8	3	300
4	30	3	0.8	3	300
5	120	1	0.2	3	300
6	120	3	0.2	3	300
7	120	1	0.8	3	20
8	120	3	0.8	3	20
9	30	1	0.2	12	300
10	30	3	0.2	12	300
11	30	1	0.8	12	20
12	30	3	0.8	12	20
13	120	1	0.2	12	20
14	120	3	0.2	12	20
15	120	1	0.8	12	300
16	120	3	0.8	12	300

Table 7-6: List of the 16 testing problems

In addition, the planning period and business operation time are set to be 30 minutes and 18 hours per day respectively. The average arrival rate and number of servers available is constant in each planning period. The constant average arrival rate is calculated as the average value of the sinusoidal λ_t function during the corresponding period. Using a constant average arrival rate in each planning period on one hand simplifies the implementation of the simulation model; on the other hand, is consistent with the format of real life data. The system capacity is set to be 50 trunk lines and the reneging time is uniformly distributed between 60 and 600 seconds among all the test scenarios.

For each testing problem, 1100 plausible rosters were generated with the procedures described in Chapter 3.

When generating rosters, shifts need to be assigned to virtual employees with synthetic preference scores to all shifts. Those data were computer bred with certain rules and assumptions. Section 7.3.1 introduces how they were generated and the reason beyond it.

7.3.1 Generating Artificial employee Preference Data

Several assumptions were made when employee preference data were generated for each shift to make the artificial data more realistic using a constant sum scaling measurement. First all possible shifts, from which to choose, are listed with shift lengths ranging from three hours to eight hours and covering the eighteen service operation hours.

Full time employees work 8-hour shifts only, yielding 21 possible shifts, and there are a total of 30 points to be allocated among these shifts. Each shift can have a preference score ranging from 0 to 30 and the summation of the preference scores assigned to all 21 shifts has to be 30. It is assumed that the preferred shifts would cluster according to starting time in a neighborhood varying from one to seven planning periods and that the preference scores for each shift in the cluster are similar.

Part time employees can work 3- to 8-hour shifts, there are 156 various shifts they can possibly work, and each employee has 200 points to allocate to their preferred shifts. The preference scores for each shift were made to range from 0 to 50, and the sum of the scores from all possible shifts 200. Employees presumably would like short, median, or long shifts but not a mix of these, and that their preferences for the starting time would be consistent with shifts having different lengths. Another assumption is that it is likely that a cluster of shifts with adjacent starting times would also receive similar preference scores.

Employees do not have to assign scores to every shift; the shifts that they do not prefer at all will receive a zero preference score using a constant sum scaling method. Because of the various combinations of arrival rates, service rates, and threshold times of service level, the number employees need to be scheduled changes dramatically among the sixteen test cases. It will be very unrealistic that the employees provided for the rosters of each scenario are from the same pool, as the number of employees required ranges from less than 20 to more than 300. Small, median, and large pools of employees were therefore generated with their personal preference data to fit various staffing requirements in the sixteen scenarios. When the number of staff required is similar, the rosters generated for different scenarios are from the same pool of employees. Below is a table showing the types of pools from which to generate rosters in the sixteen scenarios and the number of employees included in each pool.

Employee Pool	# Part Timers	# Full Timers	Applies to Test Problem
Small	10	30	9, 10, 11, 12
Median	20	190	1, 2, 3, 4, 13, 14, 15, 16
Large	40	400	5, 6, 7, 8

Table 7-7. Composition of employee pools and corresponding test problems

Plausible rosters produced using this artificial data were simulated using Arena 5.0 (Rockwell Software Inc, 2000) and their performances were evaluated according to two criteria – cost, and employee satisfaction.

7.3.2 Employee Satisfaction Results from Plausible Rosters

The performance of the three employee preference biasing heuristics – naïve match, feed and fill, and dynamic match – are compared with the traditional SIPP and set covering approach (an integer programming model that maximizes total preference) basing on the three employee satisfaction indicators discussed in Section 5.2.2.

Because of the limited space and the difficulties with seeing all the graphs for all the sixteen scenarios and three methods at once, a summary table is created to give an overall view of the results of the entire 48 experiments. Three symbols "X", "-", and "☺" are used to represent the situations where the benchmark result is worse than, similar to, or superior to the frontier of each indicator versus cost (labeled by the indicators only) given by the plausible rosters; please refer to the sample graphs given below for a visual explanation. Figure 7-12 to 7-14 are the average preference score versus costs of 3rd, 4th, and 9rd scenario respectively of the naïve method. The results generated from method naïve match, feed and fill, and dynamic match methods are given in sequence by the three symbols in each cell of Table 7-8.

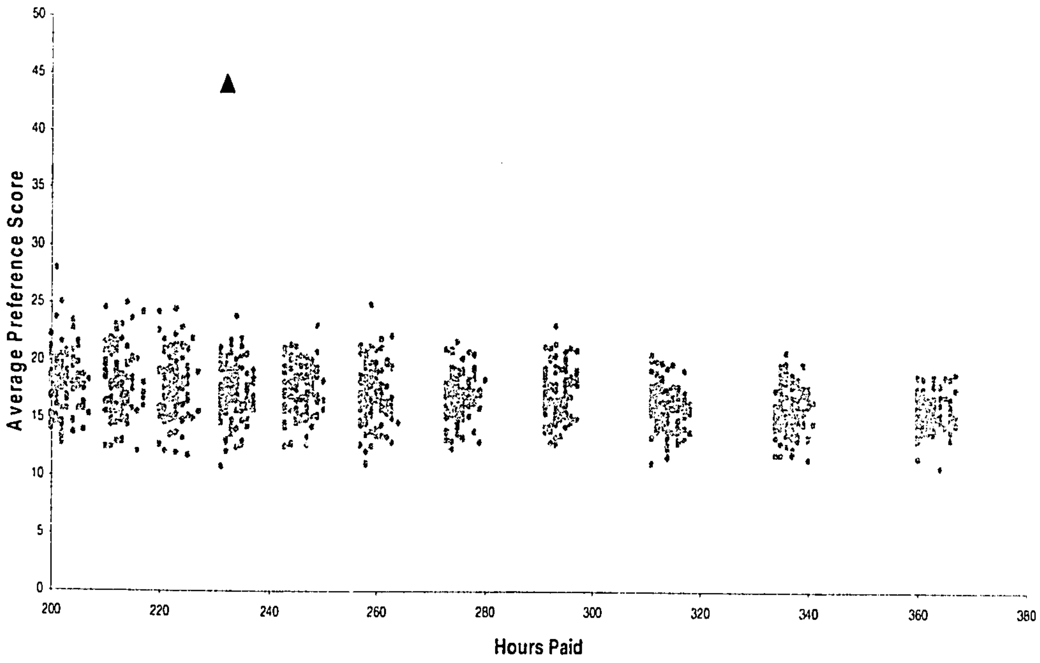


Figure 7-12. An example of "X".

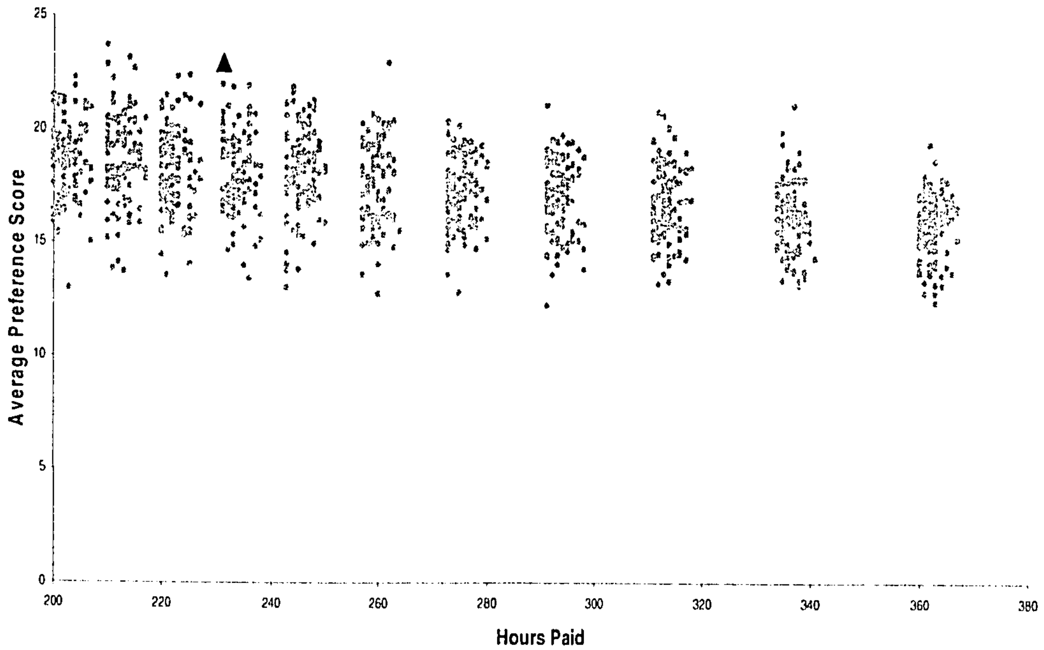


Figure 7-13. An example of "-".

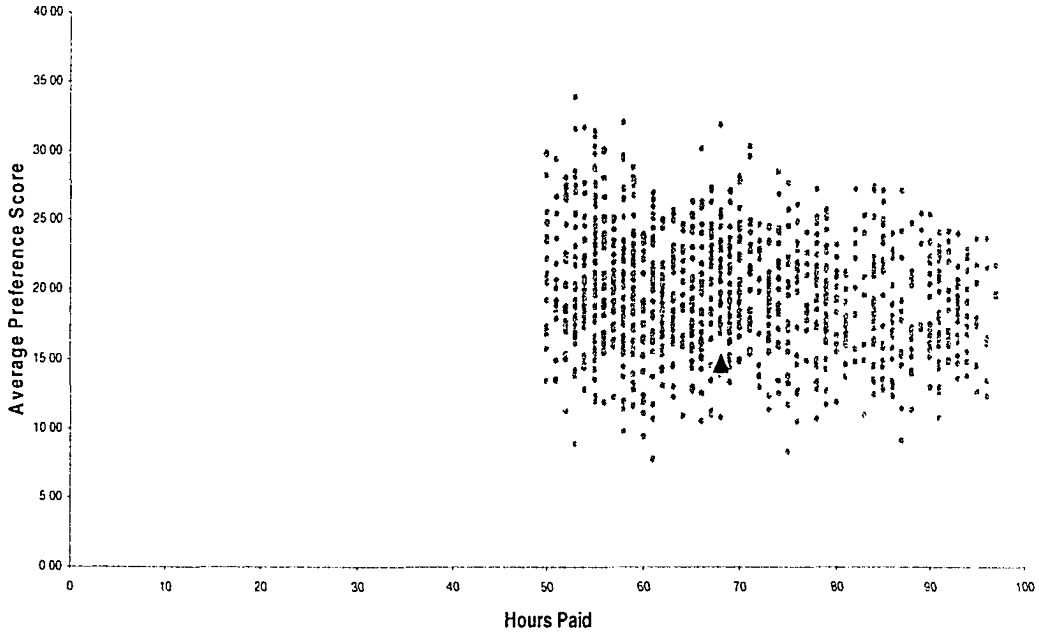


Figure 7-14. An example of "☺".

Scenarios	Avg. Preference Score	Lowest Preference Score	% of Min.
1	X☺☺	☺☺☺	☺☺☺
2	X☺☺	☺☺☺	☺☺☺
3	XX☺	☺☺☺	☺☺☺
4	-☺☺	☺☺☺	☺☺☺
5	-☺☺	☺☺☺	☺☺☺
6	XX☺	--☺	☺☺☺
7	XX-	☺☺☺	☺☺☺
8	XX☺	☺☺☺	☺☺☺
9	☺☺☺	☺☺☺	☺☺☺
10	☺☺☺	☺☺☺	☺☺☺
11	☺☺☺	☺☺☺	☺☺☺
12	☺☺☺	☺☺☺	☺☺☺
13	X☺☺	☺☺☺	☺☺☺
14	X☺☺	☺☺☺	☺☺☺
15	XX☺	☺☺☺	☺☺☺
16	☺☺☺	☺☺☺	☺☺☺

Table 7-8. Summary of employee satisfaction results for 16 scenarios

From Table 7-8, all three methods work well in most scenarios. Particularly, the SIPP and set covering combination never generate better solutions in all three employee satisfaction indicators compared to the results of dynamic match method. Naïve match and feed & fill methods provide very good results to reduce the number of least satisfied employees but in some scenarios have low average preference score. One problem of set covering method is that it can only have one constraint – maximize average preference score; therefore, though in some cases its average preference scores outperform those of the naïve match and the feed & fill methods, it has a much higher percent of employees assigned their least desired shift. Another problem is that all shifts are generated then assigned to employees when SIPP and set covering combination is used, thus employee satisfaction is totally out of the picture when shifts are generated. When an ill match presents between the schedule and employee satisfaction, little can be done.

7.3.3 Comparison of the Three Employee Preference Biasing Heuristics

Three methods are employed to generate rosters that consider maximizing total employee satisfaction. Total satisfaction is described by three variables – average preference score, the lowest preference score of a shift assigned to an employee in a roster, and the percentage of employees that received the minimum preference score in a rating system. As shown in Tables 7-9 to 7-11 the dynamic match method consistently outperforms the other two methods in all three dimensions and among all sixteen scenarios. Naïve match and feed & fill methods in some scenarios generate low average preference score, while feed & fill method is slightly better than naïve match results. The following table of graphs gives a visual comparison of the three methods for the three indicators. Scenario one is used as an example. Though having the best performance, dynamic match method is significantly slower compared to the other two methods. In the worst case, the speed can be ten times slower. Feed & fill method is the fastest one among them. To generate 1100 plausible rosters, it takes at most one hour on a Pentium IV GPU computer.

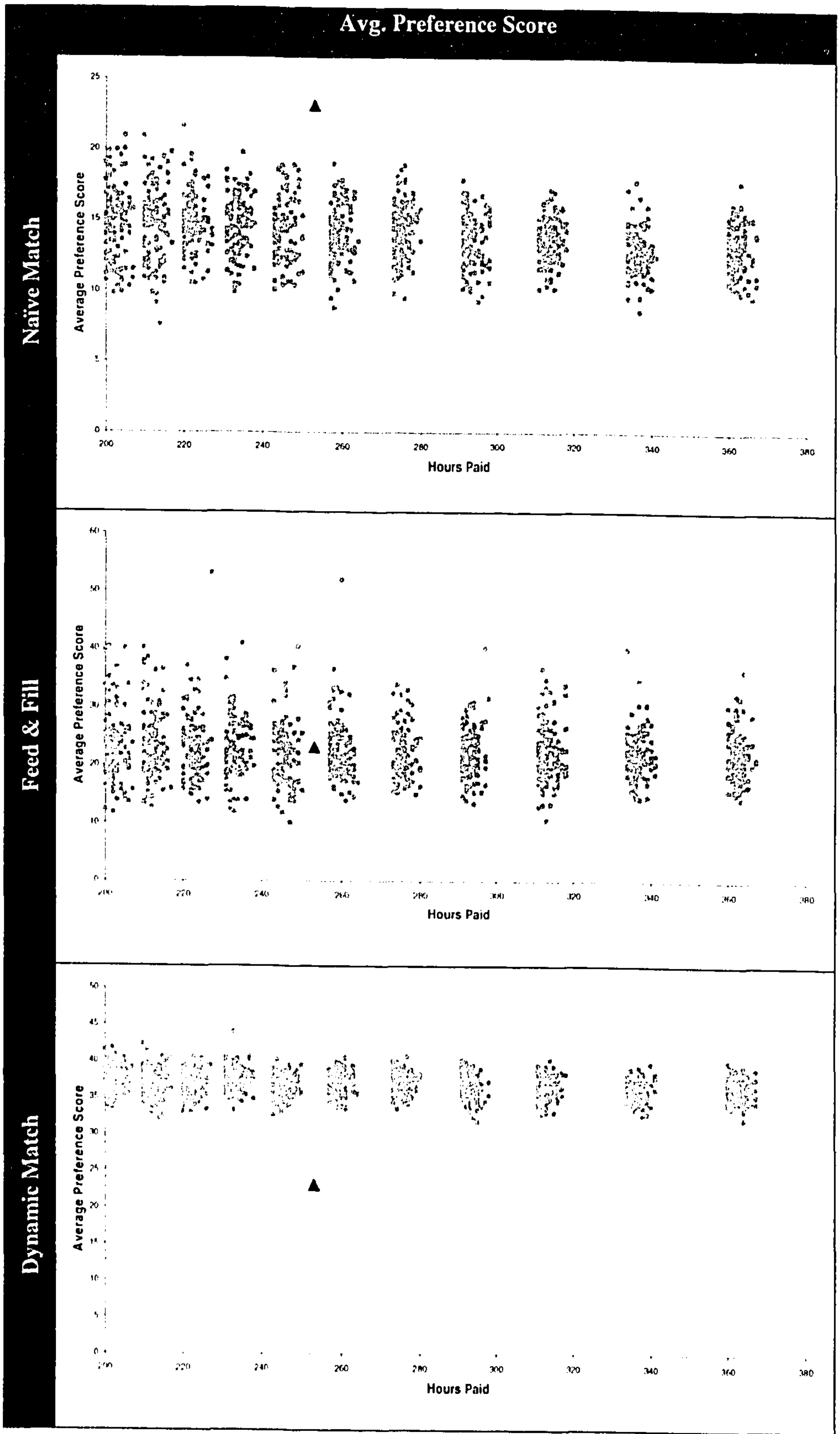


Table 7-9. Comparing the three employee-preference-biasing methods (average preference score vs. cost, scenario 1)

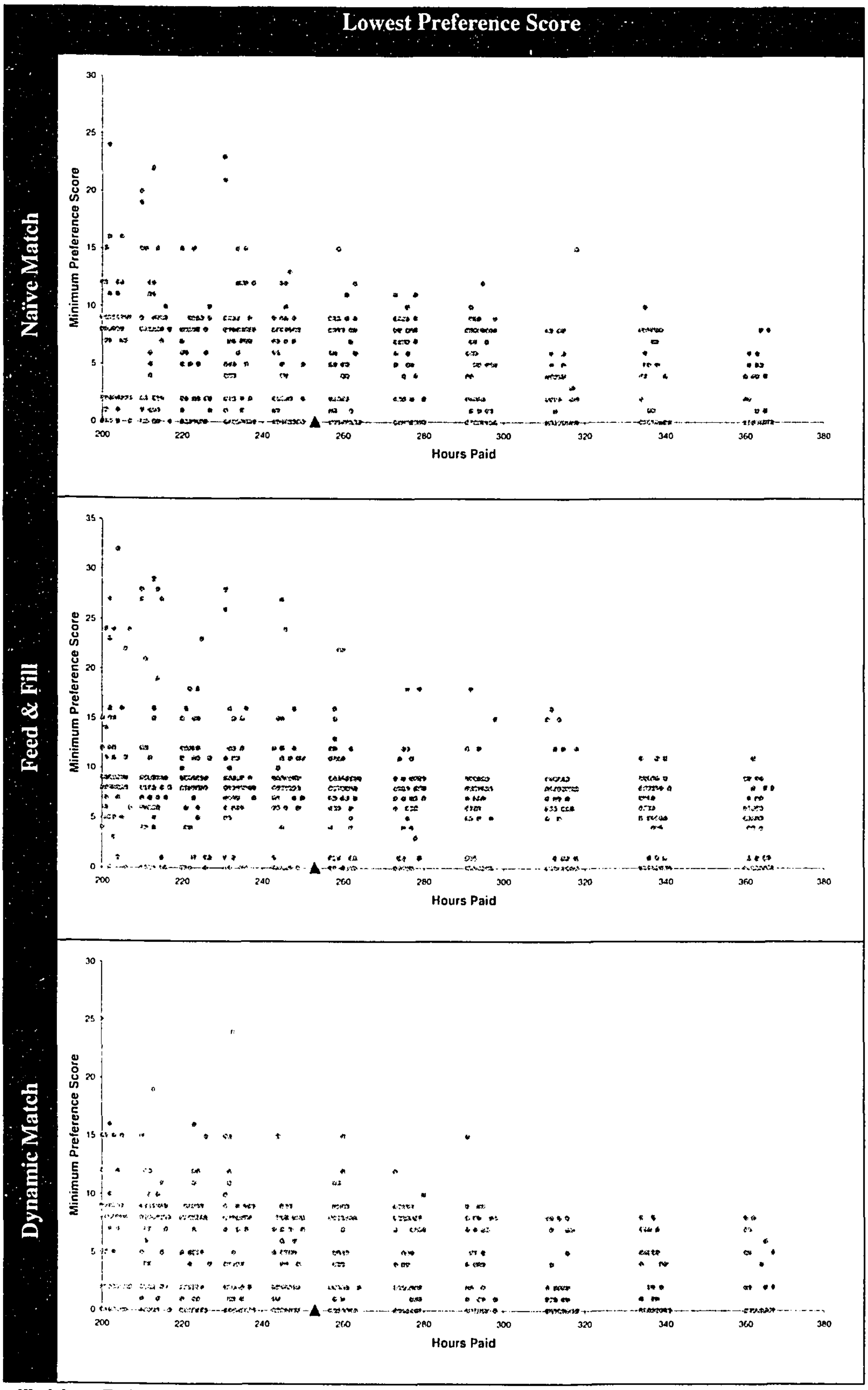


Table 7-10. Comparing the three employee-preference-biasing methods (lowest preference score vs. cost, scenario 1)

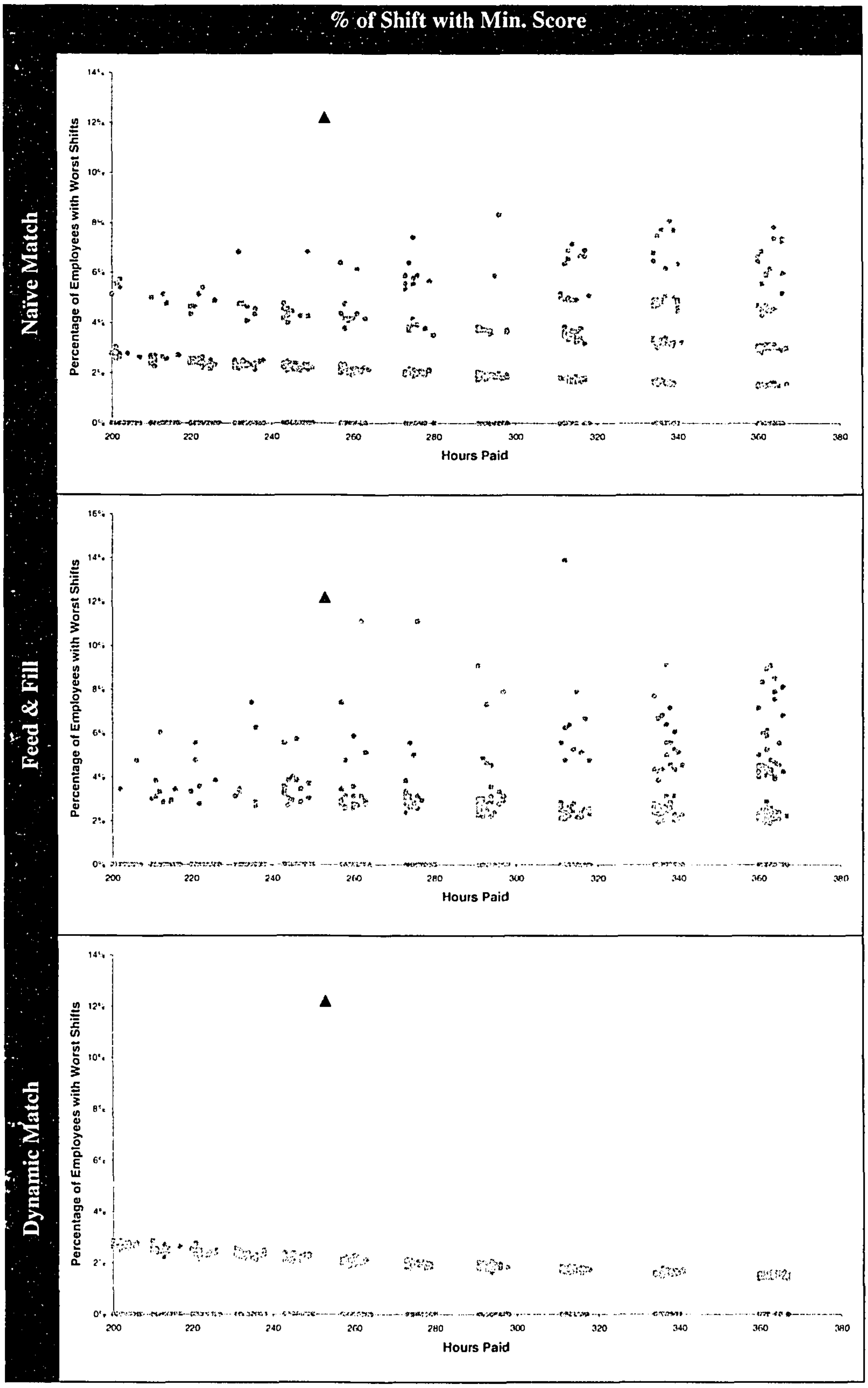


Table 7-11. Comparing the three employee-preference-biasing methods (percent of shifts with minimum score vs. cost, scenario 1)

7.4 A Rostering Experiment Using Empirical Data

The proposed automatic roster generator using a direct rating scaling is applied in this section to empirical data provided by the same utility company call centre. The call arrival curve is therefore the same as the one displayed in Figure 7-1, which has a strong and a weak peak before and after noon. Call volume starts low in the morning at around 30 calls per half hour, rises steadily to the high peak of about 260 calls per half hour, then drops a little and waves around 200 calls per hour until 4 pm, and drops constantly to around 30 calls per half hour at the end of service operation time.

In this particular call centre, incoming calls are processed directly by representatives and customers will be put on hold if all the representatives are busy. The schedule planning period is set to be 15 minutes, which implies that the average arrival rate is constant in this period. Arrivals follow a Poison process, and service times are found following the Erlang distribution with shape = 5, scale = 54.4, and shift =10. The trunk line capacity of 150 might lead to blocking behaviour if all the lines are busy at the same time. The reneging behaviour is estimated from the weekly record of each abandoned phone call to follow a discrete probability distribution based upon customer wait time. The distribution detail is given in table 7-1 same as the one used for testing the scheduling method.

The company collected employee preference data using a survey designed by us (see Appendix 3). Direct rating scale is used to find out employees' preference of working start time and at the mean time the range of preferred shift length is requested. Scale ranges from 1 to 5 indicate the least to the most that an employee would desire to start working from a specific time. There are a total of 113 employees including 22 full timers who only work 8-hour shifts. The company collected 49 completed surveys, which according to the schedule manager is the most sophisticated preference data the company ever collected. Since we were informed that those who do not care to provide preference data are willing to work on any shifts, we assumed that they are satisfied with all the possible shifts provided.

Besides the arrival rate pattern and different scale rating system for employee preferences, the empirical experiment also included breaks in the shifts. Depending on the shift lengths, at least one 15-minute break and at most three breaks including a half-hour lunch break were arranged for shifts longer than three hours. There are 3712 possible shifts in total with various starting times, shift lengths, and break arrangement combinations that all abide by working time regulations.

Given all the data described above, plausible rosters are generated biasing to the customer demand and employee satisfaction according to the proposed approach. Offered load r for each 15-minute planning period was calculated by changing the parameter ε, having utilization that ranged widely from 38% to 96% in steps of 2%, as we believe it is important to show management as complete a frontier as might interest them. Using different values of utilizations, 30 demand profiles in total are calculated; 35 plausible rosters are generated for each of them, which produce 1050 plausible rosters. When biasing the rosters to employee satisfaction, the objectives are to maximize the average preference score given for the work starting time and to satisfy employees' desired shift lengths. All randomly generated rosters are tested using a simulation model that incorporates all the data provided by the call centre with 100 replications to obtain the service quality indicators for each of them. The same CI objective in the scheduling problem is also applied here. We would like to have the average of most roster's quality indicators within +/-10% CI. The details are provided in the following table.

Indicators	Number of Schedules NOT within Target CI	Largest CI Encounters among All Rosters
Average waiting time	5	11.49%
Maximum waiting time	0	8.17%
Average queue length	7	11.86%
Maximum queue length	0	7.20%
Average Reneging rate	96	13.65%
Average Blocking rate	0	0%
Average service level	0	5.65%
Average personnel utilization	0	0.45%

Table 7-12. CI information of all the quality indicators evaluated for plausible rosters

Again, even though there is more variation in the Reneging rate than desired, the actual CIs are short in absolute terms: the Reneging rate with longest CI is 0.08% +/- 0.01092%.

Since based on comparisons in Section 7.3 the dynamic match heuristic gives the best results among the three proposed methods, the plausible rosters are generated using only this method. Again, the SIPP results are used to compare with the ones given by the proposed approach in service quality performance. To explore the frontier, SIPP results are calculated using various service levels. However, the set covering model does not suit this case since it cannot consider two objectives at the same time. It is possible to combine the two objectives by designing a survey asking for preference for shifts with both characteristics inclusive; nevertheless, it will result in a list of several hundreds of shifts, which is unrealistic to ask from people in real life. Hence, the employee satisfaction results will only be compared to a roster provided by the company with each value of the indicator manually calculated.

Among 1050 randomly generated rosters, 1021 efficient ones are obtained using FDH. Blocking rate and the percentage of employees receiving the worst shifts are left out in the FDH analysis as they are zero for all the plausible rosters generated. The following graphs show the tradeoffs between labour costs and selected service quality and employee satisfaction indicators comparing to the SIPP results (generated using various service levels from 34% to 98%) and/or the company's current rostering position.

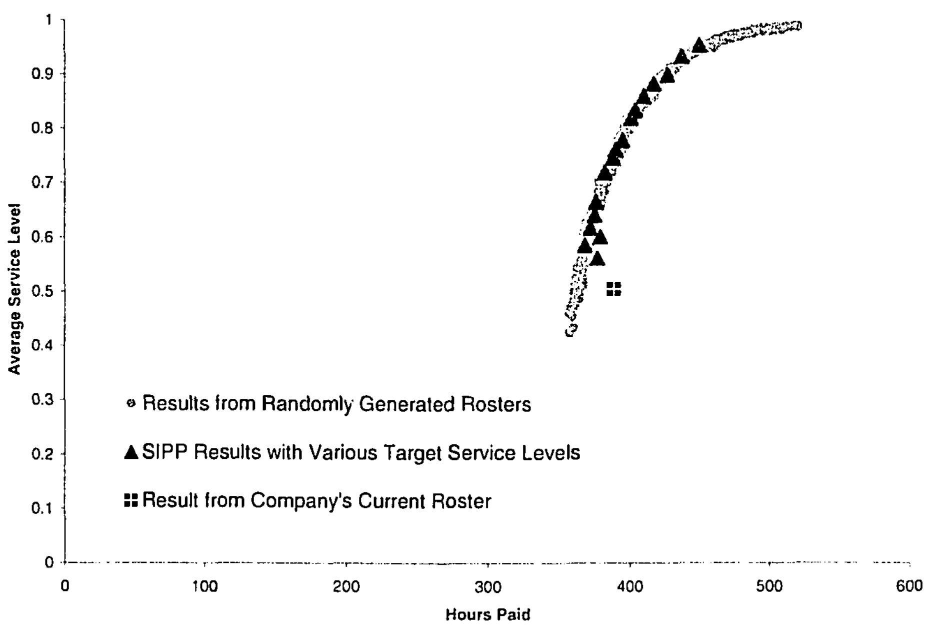


Figure 7-15. Average Service Level vs. Cost

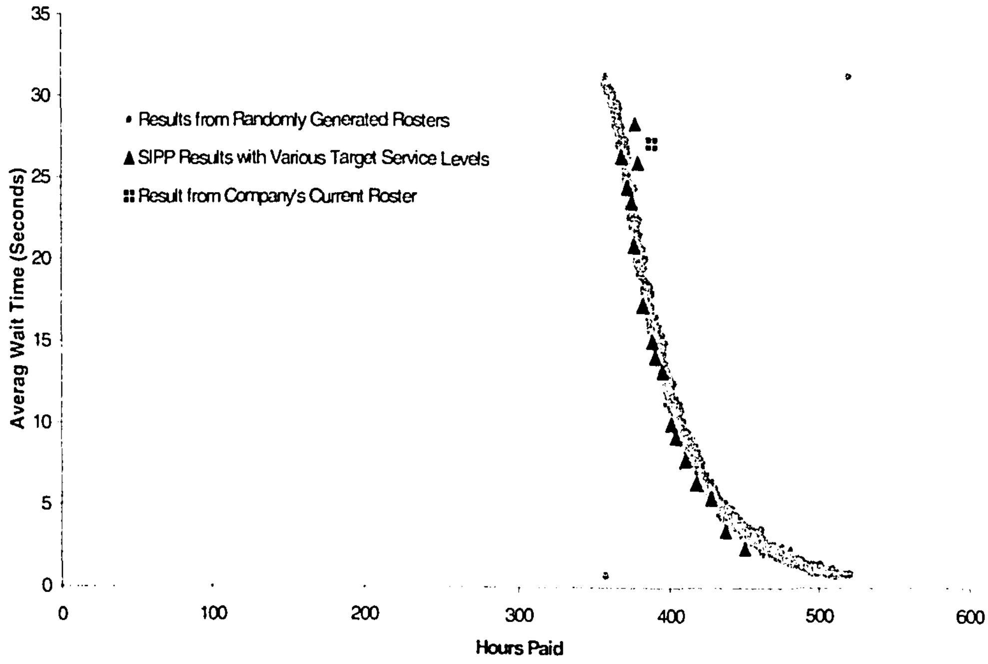


Figure 7-16. Average Wait Time vs. Cost

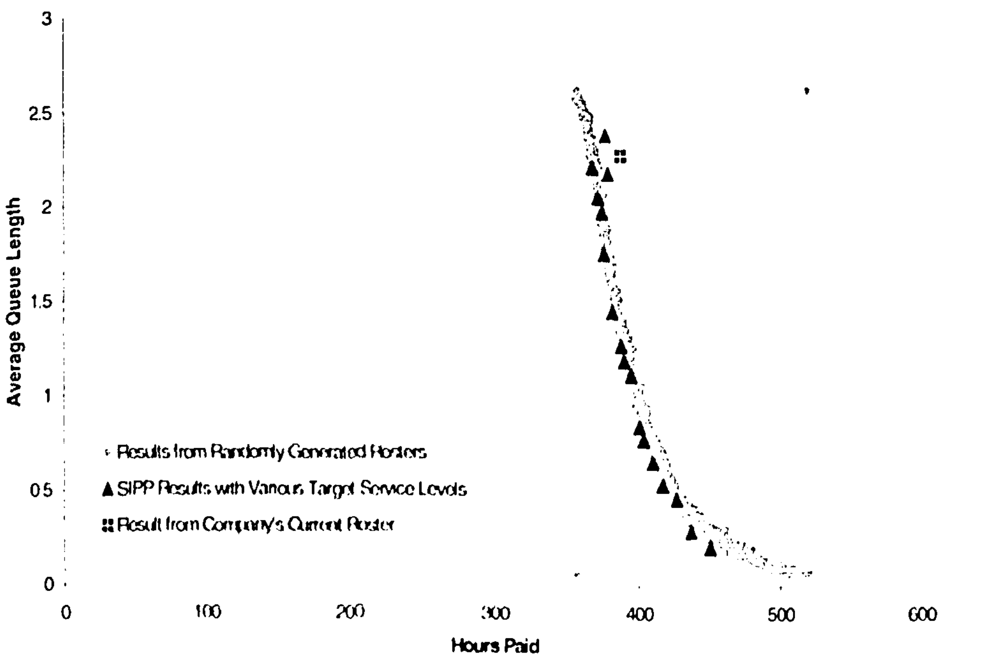


Figure 7-17. Average Queue Length vs. Cost

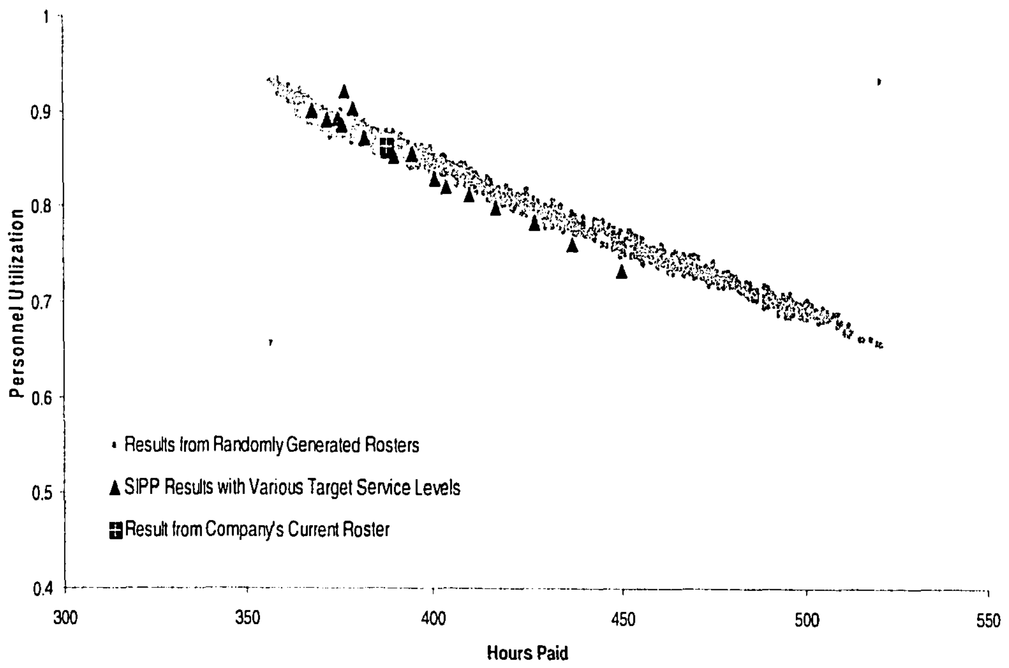


Figure 7-18. Average Utilization vs. Cost

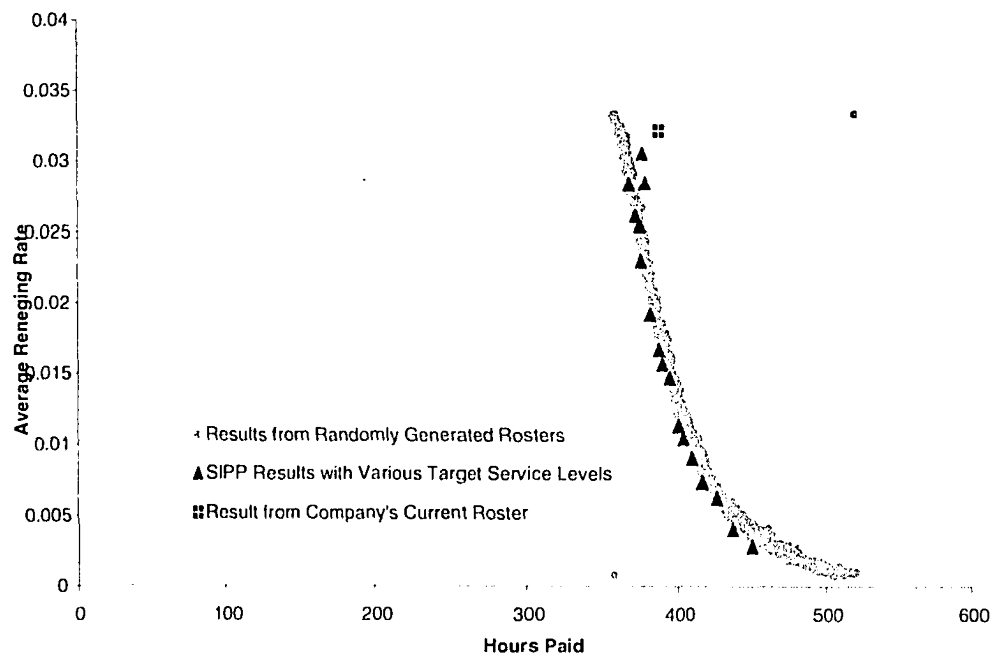


Figure 7-19. Average Reneging Rate vs. Cost

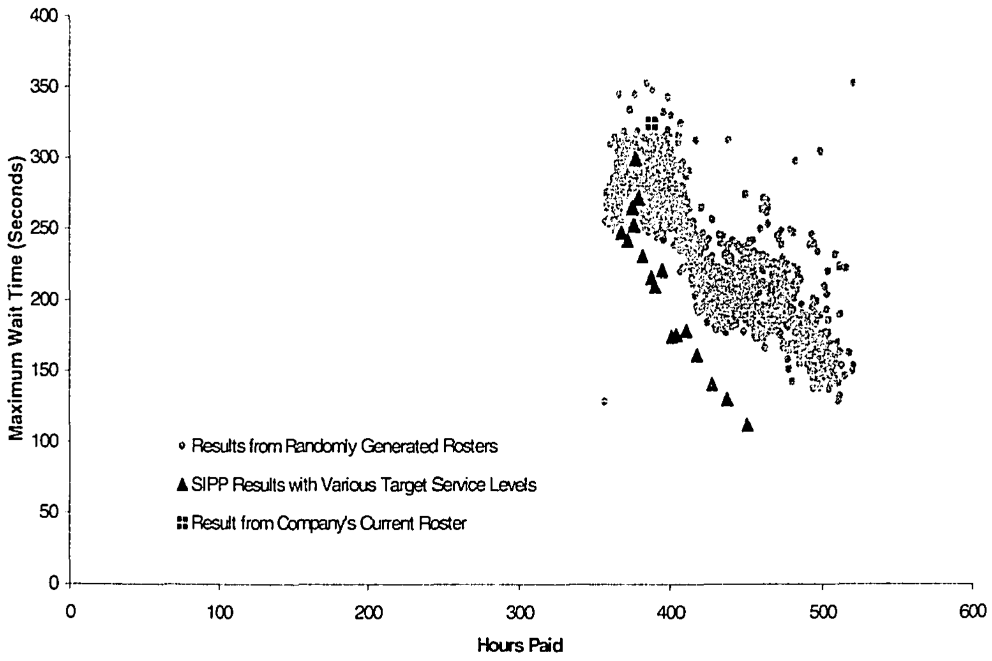


Figure 7-20. Maximum Wait Time vs. Cost

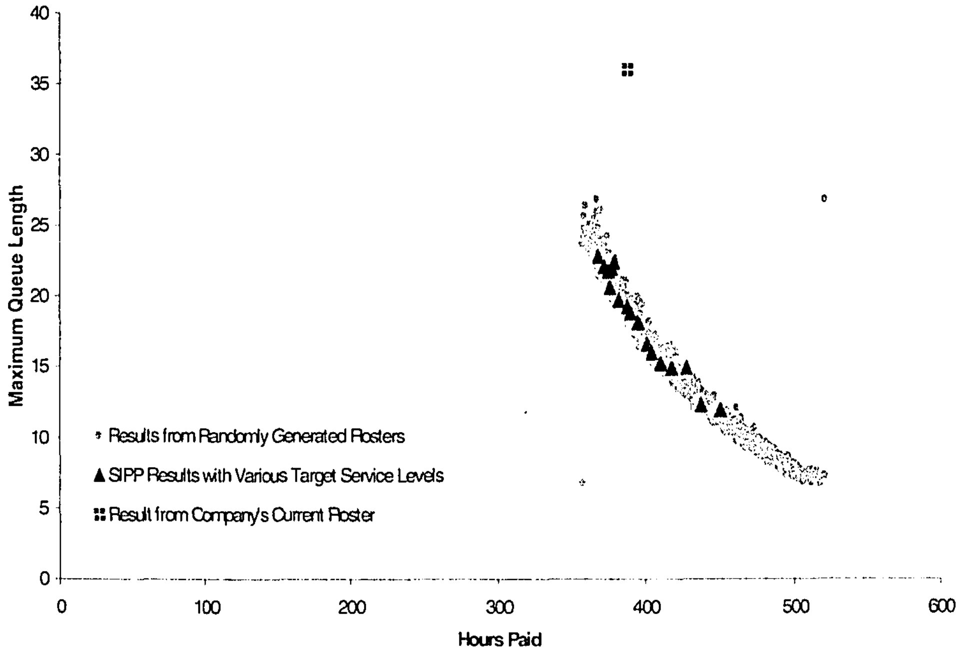


Figure 7-21. Maximum Queue Length vs. Cost

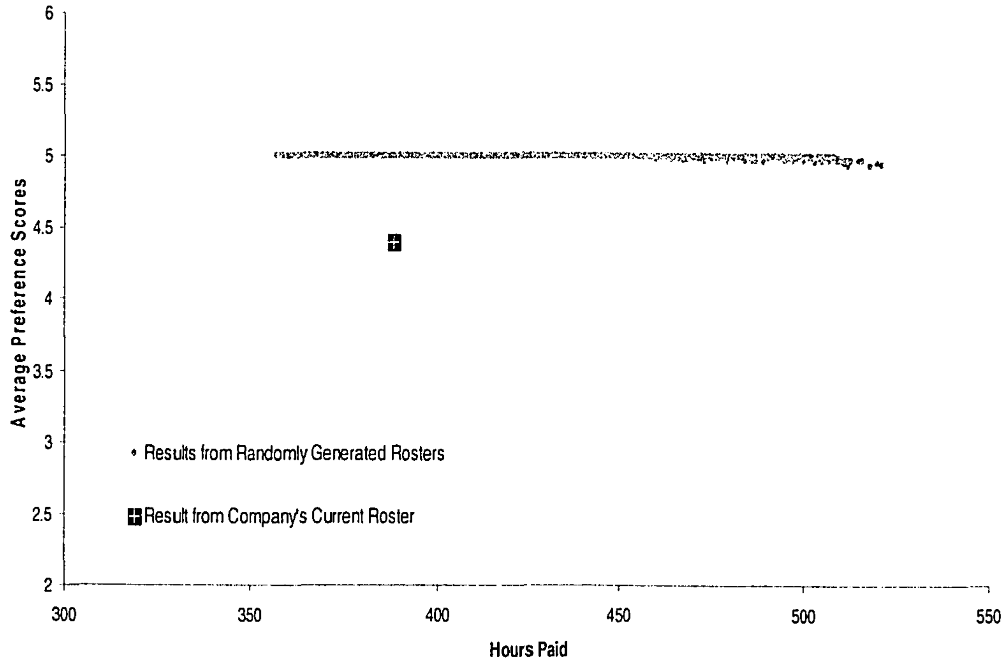


Figure 7-22. Average Preference Score vs. Cost

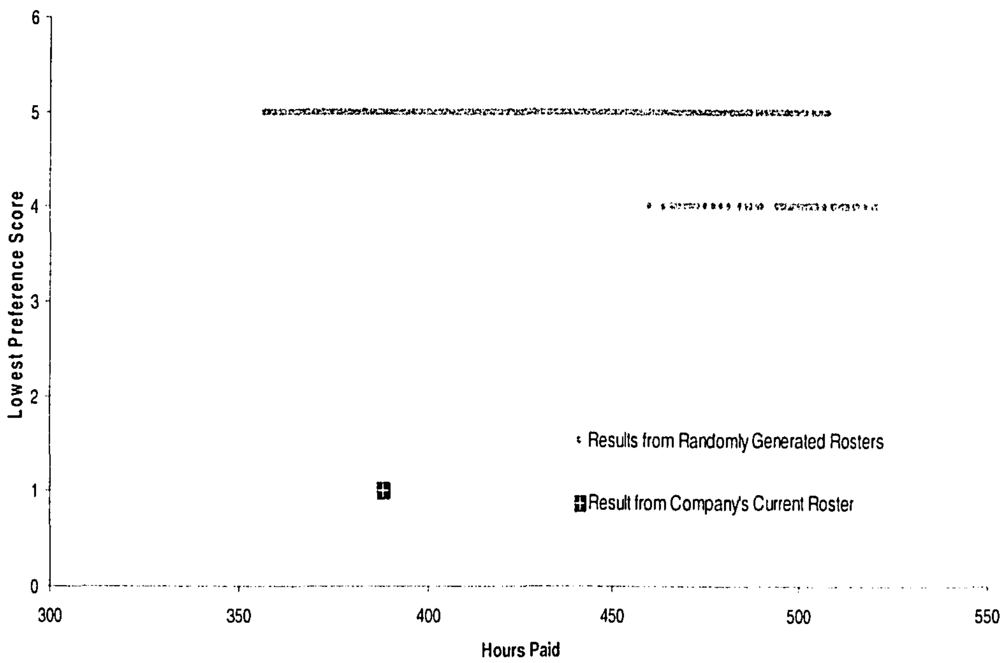


Figure 7-23. Lowest Preference Score vs. Cost

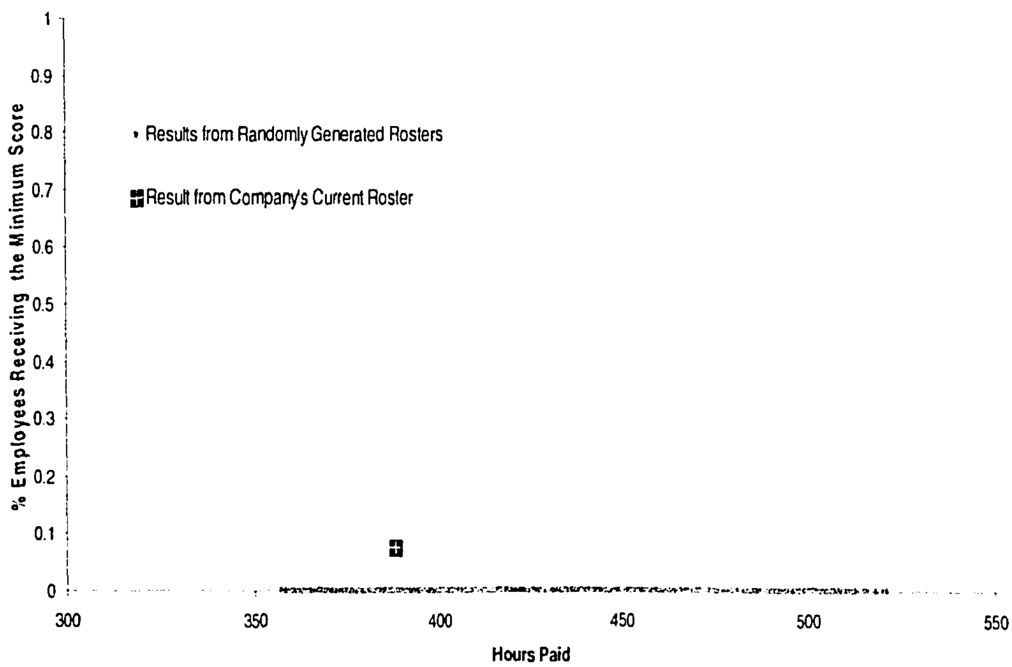


Figure 7-24. % of Minimum Score vs. Cost

The results from the SIPP method and company's current roster are generated by the same simulation model and used as benchmarks for the proposed approach. As an NP complete problem, the real frontier is unknown. To ensure that our results are at least not far off from the frontier, we used the optimized SIPP solution for comparison. From Figures 7-15 to 7-24 shown above, we can see that randomly generated rosters perform very well in terms of all service quality indicators except maximum wait time. The frontier is well above the SIPP optimized results especially when staffing cost is low; when staffing cost is high, there is limited room to improve. However, for maximum wait time, the SIPP results are better than the proposed approach. We believe the reason is that one or two planning periods are understaffed which results in a very long waiting time for a few customers. The performance results for these few customers dragged the maximum wait time down.

From Figure 7-23, it is noticeable that the lowest preference score of a shift an employee is ever assigned is 4 or 5, which means that plausible rosters satisfy almost all the employees. Data from the company shows that there were five employees being assigned shifts with preference score 1 – the least desired. Therefore, the proposed approach provides much better employee satisfaction solution.

The significance of our approach compare to the SIPP method is the speed of generating plausible rosters that explore the frontie and the ability to provide excellent results for various employee satisfaction criteria. Unlike integer linear programming, of which the run time of finding an optimized roster increases exponentially with the size of the problem (in this case number of employees); our approach is quicker to generate plausible rosters and the run time increases proportionally with the size. It took hours to generate a roster using the SIPP method yet minutes using the proposed approach for rostering approximately 100 employees.

Using the proposed approach, management can easily position the service quality and employee satisfaction they would like to achieve. When the desired value is provided for each indicator from the management, the ideal performance can be projected to the frontier of the plausible rosters and the roster that fits best the management's requirement can be found. Three sample sets of target quality and employee satisfaction indicators for rosters are projected to the measurements of plausible rosters and the corresponding closest efficient rosters are located. The results are given in Table 7-13 in next page.

Rosters	Avg. Renege %	Avg. Wait	Max Wait	Avg. Q. Length	Max Q. Length	Service Level %	Utilization	Hours Paid	Avg. Pref. Score	Min. Pref. Score
Target #1	0.10%	1	230	0.1	7	98.00%	0.675	510	4.99	4
Roster #1049	0.08%	0.766	140.62	0.065	6.87	98.79%	0.682	506	5	5
Target #2	15.00	1.1	190	0.09	7.4	98.00%	0.65	500	4.98	4
Roster #982	0.12%	1.015	189.36	0.087	7.25	98.58%	0.691	499	4.988	4
Target #3	1.50%	15	260	1.2	20	75.00%	0.85	395	4	5
Roster #338	1.49%	13.037	254.71	1.1	17.1	78.91%	0.855	395	4	5

Table 7-13. Targets and corresponding rosters

8. Conclusions, Discussion, and Future Research

This research proposes a novel scheduling and rostering framework. It is important to note that the approach does not seek to improve the traditional SIPP method. Though the SIPP method is used throughout the study to benchmark the performance of the proposed approach, it has little in common with the proposed method. In fact, one of the motivations of proposing the new scheduling method is to avoid the simple objective function of the SIPP method and its many variations – minimizing costs while meeting a specified service level for each planning period. Instead, a framework is created to consider multiple, complimentary objectives simultaneously.

Furthermore, it is important to recognize that the simulation itself is not part of the new framework. Simulation modeling in this study is applied only as a tool to calculate service quality indicators. Indeed, the simulation model can be easily replaced by any analytical model deemed appropriate to a specific situation. The framework might benefit from the improved accuracy and speed of an analytical model – if one that accommodates general service systems can be developed in the future.

Finally, the proposed approach neither employs optimization techniques nor specifies an objective function. Indeed, it is questionable whether an objective function that accounts for all the criteria could even be constructed, and we do not believe that management would be able to set meaningful weights (i.e. prices) on the various criteria, especially since many of them are interrelated. Instead, an ideal schedule or roster is defined by the management in terms of performance indicator (criteria) values. This ideal schedule or roster contains management's preference information and is therefore tailored to the specific business strategy of a particular company. Moreover, the use of FDH to project the ideal schedule or roster in consideration of management's preferences does not require cost estimation. The idea of inducing randomness in generating plausible schedules or rosters while biasing towards the demand profile is proposed due to the computational infeasibility of expressing the multidimensional efficient frontier in explicit form. The optimized SIPP results are used as a benchmark to ensure that plausible schedules or rosters form a realistic approximation of the actual efficient frontier. Furthermore, adding more dimensions to a problem is easy to implement in the proposed approach whereas it would complicate optimization models tremendously.

The points discussed above highlight what we would like to accomplish using the proposed methods: evaluate the service quality in a comprehensive and flexible manner (beyond the scope of the SIPP method); avoid strong limiting assumptions (the main reason a simulation model is adopted); and allow managers to learn about the interactions of cost and various criteria (without pricing them).

Furthermore, the multi-objective framework can be extended to accommodate more sophisticated systems. One important potential extension is to apply it in a multi-tasking time-varying service system, where each service has its own unique time-varying arrivals. For such a system it is common to cross-train employees to optimize utilization. Rostering multi-tasking employees is clearly more complicated than what the current proposed rostering method can provide. Nonetheless, if guidelines are obtained, we can modify the method to fit this new problem. Another critical extension is to generate rosters for a multiple day (i.e. weekly) period.

8.1 Extension to Multi-Tasking

Chapter 6 investigated whether cross-training is a good strategy for various types of time-varying $M(t)/G/s(t)$ queueing systems with two types of demand and how to optimally schedule flexible employees in a time-varying stochastic system, especially in the case where not all employees are cross-trained. Two particular variations are considered: different combinations of arrival shapes and the levels of demand. It was discovered that the former had little impact on cross-training decision whereas the latter did. While cross-training employees works well with various combinations of arrival shapes, it is not desirable in a high demand volume environment.

We then looked further into the guidelines for scheduling cross-trained employees. Two principles were discovered: in a system with different sizes of employee pools, cross training all employees in the smaller pool (and maybe some in the larger) gives the most benefits; and it is always better to schedule cross-trained employees in time periods where the volumes of the two demands are similar.

Since shifts can be generated while biasing not only to demand and employee satisfaction, but also to employee skills and suitability for cross-training, the proposed rostering heuristic can be modified to implement the above-mentioned principles and generate rosters with various numbers of total employees and cross-trained employees. After a shift is generated using the heuristic, it will be assigned simultaneously to an employee not only according to his/her preference but also to the skills. According to the suggested range of number of Cross-trained employees, plausible rosters can be generated by assigning shifts with appropriate timing to employees from appropriate pool. Evaluating the performance of the generated rosters, we can determine the optimized combination of the number of total employees on duty and the percentage who should be cross-trained.

8.2 Extension to Multi-period

While the proposed workforce rostering approach is suitable for generating daily rosters, it is rare in practice for rosters to be produced on a daily basis. Usually, a weekly or even monthly roster is generated, and minor adjustments can be applied day by day if necessary. More constraints need to be considered when generating rosters for a longer period of time, such as weekly days off, total weekly working hours, vacation arrangements, etc (Jarrah, Bard and deSilva 1994). With additional components and modifications the proposed rostering approach can be extended to solve a multi-day rostering problem. The new method would not only maintain its current strengths – enabling management to find a roster that meets the organization's needs in terms of costs and various aspects of service quality and employee satisfaction – but also have the ability to incorporate all additional constraints unique to the multi-day problem.

There are three aspects to be considered in a multi-day rostering problem – arranging days off, scheduling and rostering shifts for each day, and considering interactions of shift restrictions between days. The first aspect can be resolved using an integer programming formulation to decide who is available on each day based on the daily demand and employees' preference on days off. Once this information is obtained, the multi-day problem is decomposed into several daily problems with different sets of employees available for each day. Naturally, the proposed rostering method is able to generate plausible rosters biased towards demand and employee satisfaction for each day. However, other constraints that depend on the previous day's schedule might emerge. For example, a person who is assigned a night shift cannot receive an early morning shift the next day. Depending on the pooling of the desired available shifts, it is also very possible that someone will have to work shifts that no one wants. Thus, a rotation mechanism should be adopted in this case to ensure fair treatment of all employees.

When daily shifts in a weekly schedule are generated independently, none of the above mentioned constraints will be considered. One remedy might be to update employee shift preference data and employee priority ranking dynamically according to his or her previous shift. At this point, all constraints are fully reflected in the preference data and priority ranking data. The proposed rostering method can then be easily modified to assign shifts to those having higher priority, and thus be applied to the multi-period setting. For example, in the case of using constant sum scaling, employees who received undesirable shifts can be given an arbitrarily large score to their preferred shifts. If direct rating is employed, they will have the privilege of receiving preferred shifts before other employees.

8.3 Conclusion

In this study, we have proposed an original workforce scheduling and rostering framework and investigated an employee cross-training strategy. It has been shown that the proposed rostering method provides us with a much broader view of service quality and employee satisfaction. It facilitates greater understanding by the management of the interactions among labour costs, service quality, and employee satisfaction, and offers rosters that are tailor-made to fit a company's particular business strategy. The approach also has the versatility of fitting into various, more sophisticated systems with certain modifications. We have developed guidelines for scheduling cross-trained employees that assist in modifying the framework to suit multi-function time-varying service systems. Most importantly, very promising results have been obtained from the application of both the proposed scheduling and rostering approaches to real-world, empirical data from a call center. These approaches are moreover applicable in all service industries with time-varying demand.

References

Agnihothri, S.R., Taylor, P.F., 1991. Staffing a Centralized Appointment Scheduling Department in Lourdes Hospital. Interfaces 21 (5), 1-11.

Agnihothri, S.R., Mishra, A.K., Simmons, D.E., 2003. Workforce Cross-Training Decisions in Field Service Systems with Two Job Types. Journal of the Operational Research Society 54 (4), 410-418.

Agnihothri, S. R., Mishra, A.K., 2004. Cross-Training Decision in Field Services with Three Job Types and Server-Job Mismatch. Decision Sciences 35 (2), 239-250.

Ali, A.I., Seiford, L.M., 1990. Translation Invariance in Data Envelopment Analysis. Operations Research Letters 9, 403-405.

Andrews, B., Parsons, H., 1993. Establishing Telephone-Agent Staffing Levels through Economic Optimization. Interfaces 23 (2), 14-20.

Andersen, P., Petersen, N.C., 1993. A Procedure for Ranking Efficient Units in Data Envelopment Analysis. Management Science 39, 1261-1264.

Bailey, J.E., 1985. Integrated Days off and Shift Personnel Scheduling. Computers and Industrial Engineering 9 (4), 395-404.

Baker, K., 1976. Workforce Allocation in Cyclical Scheduling Problems: A Survey. Operational Research Quarterly 27 (1), 155-167.

Bartholdi, J.J., 1981. A Guaranteed-Accuracy Round-Off Algorithm for Cyclic Scheduling and Set Covering. Operations Research 29(3), 501-510.

Bateman, T., Organ, D., 1983. Job Satisfaction and the Good Soldier: The Relationship between Affect and Employee "Citizenship". Academy of Management Journal 26, 587-595.

Bearden, W.O., Netemeyer, R.G., Mobley, M.F., 1993. Handbook of marketing scales: multi-item measures for marketing and consumer behavior research. Newbury Park, CA: Sage.

Bechtold, S.E., Jacobs, L.W., 1990. Implicit Modeling of flexible Break Assignments in Optimal Shift Scheduling. Management Science 36 (11), 1339-1351.

Berman, O., Larson, R.C., Pinker, E., 1997. Scheduling Workforce and Workflow in a High Volume Factory. Management Science 43 (2), 158-172.

Bottomley, P.A., Doyle, J.R., Green, R.H., 2000. Testing the Reliability of Weight Elicitation Methods: Direct Rating Versus Point Allocation. Journal of Marketing Research 37(4), 508-513.

Brusco, M.J., Jacobs, L.W., Bongiorno, R.J., Lyons, D.V., Tang, B., 1995. Improving Personnel Scheduling at Airline Stations. Operations Research 43 (5), 741-751.

Brusco, M.J., Jacobs, L.W., 1998. Personnel Tour Scheduling When Starting-Time Restrictions Are Present. Management Science 44(4), 534-547.

Burns, R.N., Carter, M.W., 1985. Workforce Size and Single Shift Schedules with Variable Demands. Management Science 31 (5), 599-607.

Byrne, J., 1988. A Preferential Bidding System for Technical Aircrew. AGIFORS Symposium Proceeding 28, 87-99

Callahan, B.B., Khan, M.R., 1993. Planning Laboratory Staffing with a Queueing Model. European Journal of Operational Research 67 (3), 321-331.

Campbell, G. M. 1999. Cross-Utilization of Workers Whose Capabilities Differ. Management Science, 45 (5) 722-732.

Chakravarthy S.R., Agnihothri, S.R., 2005. Optimal Workforce Mix in Service Systems with Two Types of Customers. Production and Operations Management, forthcoming.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring Efficiency of Decision Making Units. European Journal of Operational Research 2 (6), 429-444.

Chung, C.A., 2004. Simulation Modeling Handbook: A Practical Approach. CRC Press LLC, Boca Raton, FL, USA.

Cooper, W.W., Seiford, L.M., Tone, K., 2000. Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software. Kluwer Academic Publishers, Norwell.

Dantzig, G.B., 1954. A Comment on Edie's Traffic Delays at Toll Booths. Operations Research 2 (3), 339-341.

DeVellis, R.F., 1991. Scales Development: Theory and Applications. Newbury Park, CA: Sage.

Doyle, J.R., 1999. Elicitation and Context Effects in Judgments: Fixed Sum versus Fixed Scale Frames. Management Science 45 (7), 972-979.

Doyle, J.R., Green, R.H., Bottomley, P.A., 1997. Judging Relative Importance: Direct Rating and Point Allocation Are Not Equivalent. Organizational Behavior and Human Decision Processes 70 (1), 65-72.

Easton F.F., Rossin, D.F., 1991. Equivalent Alternate Solutions for the Tour Scheduling Problem. Decision Sciences 22 (5), 985-1007.

Easton, F.F., Rossin, D.F., 1996. A Stochastic Goal Program for Employee Scheduling. Decision Sciences 27 (3), 541-568.

Eavens, B. C., Rothblum U. G. 1988. A Discrete Time Average Cost Flexible Manufacturing and Operator Scheduling Model Solved by Deconvexfication over Time. Operations Research 36 (2) 242-257.

Edie, L.C., 1954. Traffic Delays at Toll Booths. Operations Research 2 (2), 107-138.

Elliott, T., 1989. Cost Analysis of Alternative Scheduling. Nursing Management 20 (4), 42-47.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D., 2004. Staff Scheduling and Rostering: A Review of Applications, Methods and Models. European Journal of Operational Research 153 (1), 3-27.

Felan, J. T., Fry, T.D., 2001. Multi-Level Heterogeneous Worker Flexibility in a Dual-Resource Constrained (DRC) Job-Shop. International Journal of Production Research 39 (14), 3041-3059.

Figueira, J. Greco, S., Ehrgott, M., 2005. Multiple Criteria Decision Analysis. State of the Art Surveys, Springer Science + Business Media, Inc. New York.

Fryer, J.S., 1974. Labor Flexibility in Multiechelon Dual-Constrained Job Shops. Management Science 19 (9), 1001-1012.

Foegen, J. H., 1993. The Case for Employee Versatility. Business and Economic Review 39 (3) 30-31.

Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., Nourbakhsh, I., 2002. Staff Scheduling for Inbound Call Centers and Customer Contact Centers. AI Magazine 23(4), 30-40.

Gaballa, A., Pearce W., 1979. Telephone Sales Manpower Planning at Qantas. Interfaces 9 (3), 1-9.

Gamache, M., Soumis, F., Villeneuve, D., Desrosiers, J., Gelinas, E., 1998. The Preferential Bidding System at Air Canada. Transportation Science 32(3) 246-255.

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

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

Glover, F., McMillan, C., 1986. The General Employee Scheduling Problem: An Integration of MS and AI. Computers and Operations Research 13 (5), 563-573.

Glover, F., McMillan, C., Glover, R., 1984. A Heuristic Programming Approach to the Employee Scheduling Problem and Some Thoughts on "Managerial Robots". Journal of Operations Management 4 (2), 113-128.

Gopalakrishnan, M., Gopalakrishnan, S., Miller, D.M., 1993. A Decision Support System for Scheduling Personnel in a Newspaper Publishing Environment. Interfaces 23 (4), 104-115.

Grassmann, W.K., 1988. Finding the Right Number of Servers in Real-World Queueing Systems. Interfaces 18 (2), 94-104.

Green, L., 1985. A Queueing System with General-Use and Limited-Use Servers. Operations Research 33 (1) 168-182.

Green, L.V., Kolesar, P.J., Soares, J., 2001. Improving the SIPP Approach from Staffing Service Systems that Have Cyclic Demands. Operations Research 49 (4), 549-564.

Harris, C.M., Hoffman, K.L., Saunders, P.B., 1987. Modeling the IRS Telephone Taxpayer Information System. Operations Research 35 (4), 504-523.

Haynes, S., Nelson, N.K., Blaine, D., 1999. Psychometric Issues in Assessment Research. In P.C. Kendall, J.N. Butcher, & G. Holmbeck (Eds.), Handbook of Research Methods in Clinical Psychology (pp. 125-154). New York: John Wiley & Sons.

Holloran, T.J., Byrn, J.E., 1986. United Airlines Station Manpower Planning System. Interfaces 16 (1) 39-50.

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

Hopp, W.J., Tekin, E., Van Oyen, M.P., 2004. Benefits of Skill Chaining in Serial Production Line with Cross-Trained Workers. Management Science 50 (1), 83-98.

Hottenstein, M.P., Bowman, S.A., 1998. Cross-Training and Worker Flexibility: A Review of DRC System Research. The Journal of High Technology Management Research 9 (2), 157-174.

Hung R., 1992. Improving Productivity and Quality through Workforce Scheduling. Industrial Management 34 (6), 4-6.

Ingolfsson, A., Akhmetshina E., Budge, S., Li, Y., Wu, X., 2005. A Survey and Experimental Comparison of Service Level Approximation Methods for Non-Stationary $M(t)/M/s(t)$ Queueing Systems. INFORMS Journal on Computing. Forthcoming.

Ingolfsson, A., Cabral, E., Wu, X., 2003. Combining Integer Programming and the Randomization Method to Schedule Employees. University of Alberta Research Report No. 02-1.

Jackson, K., 2002. Thinking Beyond the Old 80/20 Rule. Call Centre Magazine 15 (1), 54-56.

Jacobs, L.W., Brusco, M.J., 1996. Overlapping Start-time Bankds in Implicit Tour Scheduling. Management Science 42 (9), 1247-1259.

Jarrah, A.I., Bard, J.F., deSilva, A., 1994. Solving Large-scale Tour Scheduling Problems. Management Science 40 (9), 1124-1144.

Jenkins, J.D., Taber, T.D., 1977. A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability. Journal of Applied Psychology 62, 392-398.

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

Joro, T., Korhonen P., Wallenius, J., 1998. Structural Comparison of Data Envelopment Analysis and Multiple Objective Linear Programming. Management Science 44 (7), 962-970.

Katz, K.L., Larson, B.M., Larson, R.C., 1991. Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage. Sloan Management Review 32 (2), 44-53.

Kelton, W.D., Sadowski, R.P., Sadowski, D.A., 2001. Simulation with Arena – 2nd Bk&Cdr Edition. The McGraw-Hill Companies, Inc. USA.

Koelling, C., Bailey, J.A., 1984. A Multiple Criteria Decision Aid for Personnel Scheduling. IIE Transactions 16 (4), 299-307.

Lauer, J., Jacobs, L.W., Brusco, M.J., Bechtold, S.E., 1994. An Interactive, Optimization-base Decision Support System for Scheduling Part-time Computer Lab Attendants. OMEGA International Journal of Management Science 22 (6), 613-626.

Law, A.M., Kelton, W.D., 2000. Simulation Modeling and Analysis – 3rd Edition. The McGraw-Hill Companies, Inc. USA.

Lissitz, R.W., Green, S.B., 1975. Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology 60, 10-13.

Loucks, J.S., Jacobs, F.R., 1991. Tour Scheduling and Task Assignment of a Heterogeneous Work Force: a Heuristic Approach. Decision Sciences 22 (4), 719-738.

Love, R.R. Jr., Hoey, J.M., 1990. Management Science Improves Fast-food Operations. Interfaces 20 (2), 21-29.

Lyons, R.F., 1992. Cross-Training: a Richer Staff for Leaner Budgets. Nursing Management 23 (1) 43-44.

Mabert, V.A., 1979. A Case Study of Encoder Shift Scheduling under Uncertainty. Management Science 25 (7), 623-631.

Mahoney, T.A., 1978. The Re-Arranged Work Week: Evaluation of Different Work Schedules. California Management Review 20 (4), 31-39.

Mehrotra, V., 1997. Ringing up Big Business. OR/MS Today 24 (4), 18-24.

Malhotra, N.K., 2004. Marketing Research: an Applied Orientation (4th ed.). New Jersey: Pearson Education.

Mandelbaum, A., Massey, W.A., Reiman, M.I., Stolyar, A., Rider, B., 2002. Queue Lengths and Waiting Times for Multi-server Queues with Abandonment and Retrials. Telecommunication Systems 21 (2-4) 149-171.

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

McCune, J.C., 1994. On the Train Gang. Management Review 83 (10) 57-60.

McKelvie, S.J., 1978. Graphic Rating Scales – How Many Categories? British Journal of Psychology 69, 185-202.

Moondra, S.L., 1976. An L.P. Model for Work Force Scheduling in Banks. Journal of Bank Research 7 (4) 299-301.

Moore, R., Evans, J., Ngo, H., 1978. Computerized Tailored Blocking. AGIFORMS Symposium Proceeding 18, 343-361.

Netemeyer, R.G., Bearden, W.O., Sharma, S., 2003. Scaling Procedures: Issues and Applications. Thousand Oaks, London, New Delhi, Sage Publications.

Nunnally, J., Bernstein, I.H., 1994. Psychometric Theory (3rd ed.). New York: McGraw-Hill.

Ostroff, C., 1992. The Relationship between Satisfaction, Attitudes, and Performance: An Organizational Level Analysis. Journal of Applied Psychology 77, 963-974.

Ozden, M., Ho, Y., 2003. A Probabilistic Solution-Generator for Simulation. European Journal of Operational Research 146 (1), 35-51.

Park, P.S., 1991. The Examination of Worker Cross-training in a Dual Resource Constrained Job Shop. European Journal of Operational Research 52 (3), 291-299.

Peter, J.P., 1979. Reliability: A Review of Psychometric Basics and Recent Marketing Practices. Journal of Marketing Research 16 (2), 6-17.

Pfeffer J., 1995. Producing Sustainable Competitive Advantage through the Effective Management of People. The Academy of Management Executive 9 (1), 55-72.

Pinker, E. J., Shumsky R.A., 2000. The Efficiency-Quality Trade-off of Cross-Trained Workers. Manufacturing & Service Operations Management 2 (1), 32-48.

Preston, C.C., Colman, A.M., 2000. Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. Acta Psychologica 104, 1-15.

Pristker, A.A.B., 1986. Introduction to Simulation and SLAM-II - 3rd Edition. Wiley, New York, NY, USA.

Rockwell Software Inc, 2000, Arena 5.0.

Schlesinger, L.A., Heskett, J.L., 1991. Customer Satisfaction Is Rooted in Employee Satisfaction. Harvard Business Review 69, 148-149.

Seila, A., Ceric, V., Tadikamalla, P., 2003. Applied Simulation Modeling Appendix A.3.1, Brooks/Cole – Thomson Learning, Belmont, CA, USA.

Shaw, M.E., Wright, J.M., 1967. Scales for the Measurement of Attitudes. New York: McGraw-Hill.

Silvestro, R., Silvestro C., 2000. An Evaluation of Nurse Rostering Practices in the National Health Service. Journal of Advanced Nursing 32 (3), 525-535.

Stuller, J., 1999. Making Call-Center Voices Smile: a Business Case for Better Training. Training 36 (4), 26-32.

Sze, D.Y., 1984. A Queueing Model for Telephone Operator Staffing. Operations Research 32 (2), 229-249.

Taha, H.A., 1981. Queueing Theory in Practice. Interfaces 11 (1), 43-49.

Taylor, P.E., Huxley, S.J., 1989. A Break from Tradition for the San Francisco Police: Patrol Officer Scheduling Using an Optimization-Based Decision Support System. Interfaces 19 (1), 4-24.

Thompson, G.M., 1990. Shift Scheduling in Services When Employees Have Limited Availability: An L. P. Approach. Journal of Operations Management 9 (3), 352-370.

Thompson, G.M., 1995a. Labor Scheduling Using NPV Estimates of the Marginal Benefit of Additional Labor Capacity. Journal of Operations Management 13 (1), 67-86.

Thompson, G.M., 1995b. Improved Implicit Optimal modeling of the Shift Scheduling Problem. Management Science 41 (4) 595-607.

Thompson, G.M., 1996. Optimal Scheduling of Shifts and Breaks Using Employees Having Limited Time-availability. International Journal of Service Industry Management 7 (1) 56-73.

Thompson, G.M., 1999. Labor Scheduling, Part 3: Developing a Workforce Schedule. Cornell Hotel and Restaurant Administration Quarterly 40 (1), 86-96.

Tien, J.M., Kamiyama, A., 1982. On Manpower Scheduling Algorithms. SIAM Review 24 (3), 275-287.

Tulkens, H., 1993. On FDH Analysis: Some Methodological Issues and Applications to Retail Banking, Courts and Urban Transit. Journal of Productivity Analysis 4(3), 183 - 210.

Turney, J.R., Cohen, S.L., 1983. Alternative Work Schedules Increase Employee Satisfaction. Personnel Journal 62(3), 202-207.

US Department of Labour. http://www.bls.gov/oco/cg/cgs035.htm.

Vakharia, A.J., Selim, H.S., Husted, R.R., 1992. Efficient Scheduling of Part-time Employees. OMEGA International Journal of Management Science 20 (2), 201-213.

Wallace, R.B., Whitt, W., 2004. A Staffing Algorithm for Call Centers with Skill-Based Routing. Working paper. http://www.columbia.edu/~ww2040/recent.html.

Whitt, W., 2004. A Diffusion Approximation for the G/GI/n/m Queue. Operations Research 52(6), 922-941.

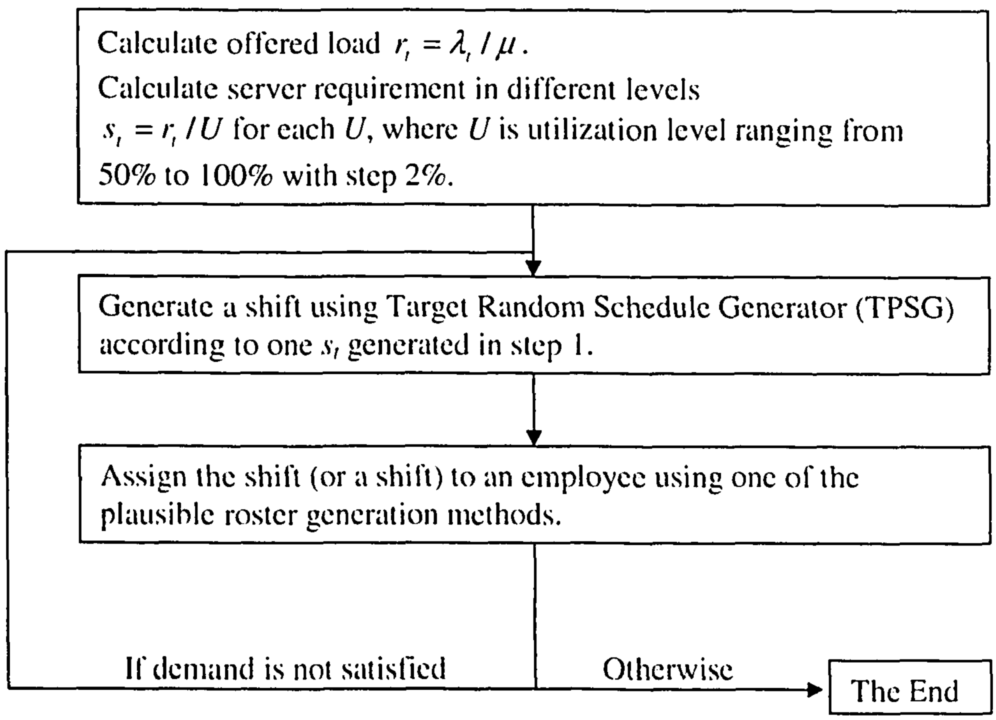
Yeung, A.K., Berman, B., 1997. Adding Value through Human Resources: Reorienting Human Resource Measurement to Drive Business Performance. Human Resource Management 36(3), 321-335.

Zhang, Z.G., Tian, N., 2004. An Analysis of Queueing Systems with Multi-task Servers. European Journal of Operational Research 156(2), 375-389.

Appendices

Appendix 1: Plausible Roster Generation

There are several steps involved in generating plausible rosters that bias to demand and employee preferences. The following flow chart roughly shows how a plausible roster is produced given average arrival rate in each planning period (λ_t) and average service rate (μ). The only exception is that when feed and fill is applied, the step 2 in the plausible roster generation method is not needed until the updated server requirement s_t in at least one planning period reaches zero. For detail please refer to Section 3.4.2.



TPSG has one method which was introduced in Section 3.2. It generates plausible shifts in a way that the aggregated number of servers resembles the pattern of s_t. Each generated shift will be assigned to an employee using one of the plausible generation methods before the next shift is to be generated. The three plausible roster generation methods and their pseudocodes are presented in the following sections.

Naïve Match Method

The naïve match method will assign the generated shift SHIFT(n) to an available employee who allocates the most points to the shift. If there is a tie, a random one among them will be chosen. Step 0 is not part of the method.

List of notations:

N is the total number of employees need to be scheduled;

P(m, n) is the score employee m assigns to shift n;

MaxP is the maximum preference score an employee is assign to a shift;

$\Lambda(m)$ is a binary array that denotes the availability of employee m. 1: unavailable, 0: available;

X(l) is an array to record all the employees who have a tie in the maximum preference score;

RAN(x, y) is a random number generated between x and y;

R(m) records the shift that is assigned to employee m.

Pseudocode of naïve match method

```
Step 0: TPSG generates SHIFT(n)
Step 1: find the available employees allocates the most
preference score to SHIFT(n)
LET MaxP = 0
LET l = 0
ERASE X()
FOR i := 1 TO N DO
      IF P(i, n) > MaxP && Λ(i) == 0
            THEN MaxP = P(i, n)
FOR i := 1 to N Do
      IF P(i, n) = MaxP && Λ(i) == 0
            THEN l = l + 1
                 X(l) = i
```

```
Step 2: randomly make one of the chosen employees not available
IF l = 1
      THEN Λ(X(1)) = 1
      ELSE  i = RAN(1, l)
            Λ(X(i)) = 1
Step 3: record the shift that is assigned to the employee
R(X(i)) = n
Back to step 0 until the demand is met.
```

Feed and Fill Method

As the name implies, this method has two stages. First, random employees are asked to give the best shift they like to work without considering the demand part. When the demand in any planning period diminishes to 0, employees are selected to fill in the demand. The second stage is exactly the same as the naïve match method.

Additional list of notations:

T is the total number of shifts that will be assigned to employees;

Pseudocode of feed and fill method

```
Do until demand in at least one planning period is 0
      Step 1: randomly pick employees and assign them the best
      shift
      LET k = RAN(1, N)
      IF Λ(k) == 0
      THEN
            LET Λ(k) == 1
            LET MaxP = 0
            FOR i := 1 TO T
                  IF P(k, i) > MaxP
                  THEN MaxP = P(k, i)
                        R(k) = i
      Step 2: Update the demand profile
Step 3: Repeat the naïve match method until the demand is met.
```

Dynamic Match Method

When the first random shift is generated by TPSG all the employees not considering the shift as the worst one are possible candidates for that shift in the dynamic match method. Each possible candidate will lead to a thread of roster by assigning the later generated shifts one by one to the available employees with the highest preference score. The roster with the highest average preference score is the one that will be chosen. There are many vectors involved to record the many possible threads of solutions and related ties when more than one employee assigns same preference score to the same shift. The programming is therefore more complicated.

Additional list of notations:

C counts the number of employees who were assigned a shift;

Θ is the number of threads of rosters being generated;

DP(i, m) records the preference score of the shift employee m is assigned in thread i;

DA(i, m) records if a shift is assigned to employee m in thread i. 0: not assigned, + (shift #): assigned, - (shift #): tied;

DM(i) records the cumulative preference score in trace i;

DX(i, j) records the number of ties for thread i and shift j;

ι counts the number of ties;

e records the employee # that has the highest preference score to a given shift

Pseudocode of dynamic match method

```
Step 0: TPSG generates SHIFT(n)
Step 1: Assign SHIFT(n) to an available employee
IF C = 0
Sub-step 1-1: In the first round of iteration, find out how many
threads there are and record them
THEN Θ = 0
      FOR i := 1 TO N
            IF P(i, n) > Least Preference Score
```

```
            THEN Θ = Θ + 1
                  DP(Θ, i) = P(i, n)
                  DΛ(Θ, i) = n
                  DM(Θ, i) = DP(Θ, i)
ELSE
```

Sub-step 1-2: For each thread, assign SHIFT(n) to the employee allocates the highest preference score to it. If there is a tie, record the tie.

```
      FOR i := 1 TO Θ
            LET MaxP = Max_j{P(j)}, ∀ DΛ(j) <= 0
            LET DM(i) = DM(i) + MaxP
            Sub-step 1-2-1: Find out if there is a tie
            LET ι = 0
            FOR j := 1 TO N
                  IF P(j) == MaxP && DΛ(i, j) < 0
                        ι = ι + 1
                        e = j
            Sub-step 1-2-2: If there is no tie, assign the shift
            to the employee e. If the employee is in a tie, clear
            up the tie.
            IF ι = 1
            THEN
                  IF DΛ(i, e) < 0
                  THEN DX(i, -DΛ(i, e)) = DX(i, -DΛ(i, e)) - 1
                        DP(i, e) = MaxP
                        Sub-step 1-2-2-1: IF the tie employee e
                        was in a tie that has only two elements,
                        the other employee should be found and
                        work for the tied shift.
                        IF DX(i, -DΛ(i, e)) = 1
                        THEN DX(i, -DΛ(i, e)) = 0
                              FOR j := 1 to N
                                 IF DΛ(i, j) = DΛ(i, e) && j ≠ e
                                    DP(i, j) = P(j, -DΛ(i, j))
                                    DΛ(i, j) = - DΛ(i, j)
                  DΛ(i, e) = n
```

```
ELSE
Sub-step 1-2-3: If there is a tie, find out if the
employees involved are in other ties, if so remove
them from the previous ties and form the new tie.
FOR j := 1 TO N
      IF P(j, n) = MaxP
      THEN
            IF DΛ(i, j) < 0
            THEN
              DX(i, -DΛ(i, j)) = DX(i, -DΛ(i, j))-1
              Sub-step 1-2-3-1: IF the employee j was
              in a previous tie that has only two
              elements, the other employee should be
              found and work for the previous tied
              shift.
              IF DX(i, -DΛ(i, j)) = 1
              THEN DX(i, -DΛ(i, j)) = 0
                    FOR k := 1 to N
                     IF DΛ(i, k) = DΛ(i, j) && k ≠ j
                        DP(i, k) = P(k, -DΛ(i, k))
                        DΛ(i, k) = - DΛ(i, k)
              DX(i, n) = DX(i, n) + 1
              DΛ(i, j) = -n
            ELSE DX(i, n) = DX(i, n) + 1
              DΛ(i, j) = -n
Sub-step 1-2-4: At the end, check the number of
elements in the new tie, in case that two same
employee tie in two same shifts.
IF DX(i, n) = 1
THEN DX(i, n) = 0
      DΛ(i, n) = n
      DP(i, n) = MaxP
```

Back to step 0 until the demand is met.

Appendix 2: Employee Preference Survey

Employee Information Form

Code: ______________________ Date: ____________

Preferences:

For those who work part time, please answer questions 1 and 2.

1. **Maximum** number of weekly hours available: __________ hours

2. Please indicate the minimum, maximum and most preferred number of working hours on each day:

	Minimum (Hours)	**Preferred** (Hours)	**Maximum** (Hours)
Mon.			
Tue.			
Wed.			
Thru.			
Fri.			
Sat.			

If you have no preferences on the starting time of shifts, please check □,
otherwise do not check and please fill in all the blanks in the following table:

Please indicate your preferences of the **starting** time with scale 5 to 1, where 5 the most preferred and 1 the least satisfactory under each time slot. (note: 7:00 am means anytime between 7:00 and 7:45 am.)

Mondays:

7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

Tuesdays:

7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

Wednesdays:

7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

Thursdays:

7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

Fridays:

7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

Saturdays:

8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM

- □ **All/Most Saturdays**
- □ **In rotation on Saturday, designated day off ________**
- □ **Prefer not to work on Saturdays**

If you work part time, please indicate your preferences of the starting times for the rest of the day.

Mondays:

2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM

Tuesdays:

2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM

Wednesdays:

2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM

Thursdays:

2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM

Fridays:

2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM